Ecole Nationale d'Ingénieurs de Tunis

UNIVERSITÉ PARIS DESCARTES

THÈSE

pour l'obtention de grade de

DOCTEUR EN TÉLÉCOMMUNICATIONS

de l' École Nationale d'Ingénieurs de Tunis

et de l'Université Paris Descartes

ÉCOLE DOCTORALE SCIENCES ET TECHNIQUE DE L'INGENIEUR

ÉCOLE DOCTORALE INFORMATIQUE, TÉLÉCOMMUNICATION ET ÉLECTRONIQUE

Présentée par

Imen SAMAALI GHARBI

Tatouage pour le renforcement de la qualité audio des systèmes de communication bas débit

Directeurs de thèse :
Mme. Monia Turki : Profeseur, ENIT
Mme. Madeleine Bonnet : Profeseur, Université Paris Descartes

Soutenue le 16 Janvier 2013 devant le jury composé de :

Président :	Hichem Besbes	-	COSIM, Sup'Com
Rapporteurs :	Sofia Ben Jebara	-	TECHTRA, Sup'Com
	Laurent Girin	-	GIPSA-LAB, INPG
Directeur :	Monia Turki	-	U2S, ENIT
Co-encadreur :	Gaël Mahé	-	LIPADE, Université Paris Descartes
Examinateur :	Abdallah Adib	-	LIM, FST Mohammedia
Invité :	Imed Jedidi	-	Ingénieur Telnet.

Résumé

L'objectif de cette thèse est d'étudier l'idée du tatouage dans le traitement du son. Les recherches en tatouage audio se sont principalement tournées vers des applications sécuritaires ou de transmission de données auxiliaires. Une des applications visées par ce concept consiste à améliorer la qualité du signal hôte ayant subi des transformations et ceci en exploitant l'information qu'il véhicule. Le tatouage audio est donc considéré comme mémoire porteuse d'informations sur le signal originel.

La compression à bas débit des signaux audio est une des applications visée par ce concept. Dans ce cadre, deux objectifs sont proposés :

- la réduction du pré-écho et de l'amollissement d'attaque, deux phénomènes introduits par les codeurs audio perceptifs, en particulier les codeurs AAC et MP3 ;
- la préservation de l'harmonicité des signaux audio dégradée par les codeurs perceptifs à extension de bande, en particulier le codeur HE-AAC.

La **première partie** de ce manuscrit présente les principes de base des systèmes de codage bas débit et étudie les différentes distorsions introduites par ces derniers. Fondées sur cette étude, deux solutions sont proposées. La première, visant principalement la réduction du pré-écho, consiste à corriger l'enveloppe temporelle du signal après réception en exploitant la connaissance a priori de l'enveloppe temporelle du signal original, supposée transmise par un canal auxiliaire à faible débit ($<$ 500 bits/s). La seconde solution vise à corriger les ruptures d'harmonicité générées par les codeurs à extension de bande. Ce phénomène touche essentiellement les signaux fortement harmoniques (exemple : violon) et est perçu comme une dissonance. Une préservation de l'harmonicité des signaux audio par des opérations de translation spectrale est alors proposée, les paramètres étant là encore transmis par un canal auxiliaire à faible débit.

La **seconde partie** de ce document est consacrée à l'intégration du tatouage audio dans les techniques de renforcement de la qualité des signaux audio précitées. Dans ce contexte, le tatouage audio remplace le canal auxiliaire précédent et œuvre comme une mémoire du signal originel, porteuse d'informations nécessaires pour la correction d'harmonicité et la réduction de pré-écho. Cette seconde partie a été précédée par une étape approfondie de l'évaluation des performances de la technique de tatouage adoptée en terme de robustesse à la compression MPEG(MP3, AAC et aacPlus).

Abstract

The goal of this thesis is to explore the idea of watermark for sound enhancement. Classically, watermark schemes are oriented towards security applications or maximization of the transmitted bit rates. Our approach is completely different. Our goal is to study how an audio watermarking can improve the quality of the host audio signal by exploiting the informations it conveys. The audio watermarking is considered as a memory that carries information about the original signal.

The low bitrate compression of audio signals is one of the applications covered by this concept. In this context, two objectives are proposed :

- reducing the pre-echo and the attack softening, two phenomena introduced by the perceptual audio coders, particularly AAC and MP3 encoders ;
- preserving the harmonicity of audio signals, distorted by coders with bandwidth extension, especially HE-AAC encoder. These coders are limited in the reconstruction of the high-frequency spectrum mainly because of the potential unpredictability of the fine structure of the latter, as well as imperfect indicators of tonal to noise.

The **first part of this manuscript** presents the basic principles of low rate coding systems and studies the various distortions introduced by the latter. Based on this study, two solutions are proposed. The first one, principally aimed at reducing the pre-echo, consist in correcting the time envelope of the signal after reception by exploiting the prior knowledge of the temporal envelope of the original signal, which is assumed transmitted by an auxiliary channel at low bitrate ($<$500 bps). The second solution is to correct the harmonicity generated by coders with bandwidth extension. This primarily affects strongly harmonic signals (e.g. violin) and is perceived as a dissonance. We propose then to preserve the harmonicity of audio signals by spectral translations. The parameters being passed again by an auxiliary channel at low bitrate.

The **second part of this document** is dedicated to the integration of audio watermarking techniques in the solution presented in the first part. In this context, the audio watermarking replaces the previous auxiliary channel and is regarded as a memory of the original signal, carrying information necessary for the correction of harmonicity and the pre-echo reduction.

Préambule

Les travaux de thèse présentés dans ce document sont élaborés dans le cadre d'une cotutelle de thèse entre l'Ecole Nationale d'Ingénieurs de Tunis et l'Université Paris Descartes et ont pour objectif l'amélioration de la qualité du signal audio obtenu à la sortie des nouveaux systèmes de communications numériques pour différentes applications (Voix sur IP, Transmission radio-mobile 3^{ieme} génération, radio mondiale, etc,...). Ces travaux sont soutenus par le projet WARRIS (Watermarking Réflexif pour le renforcement des Images et du Son) du programme Jeunes Chercheurs de l'ANR (français) de la Recherche et le projet de coopération Tuniso-Français, "Evaluation et renforcement de la qualité en communication audio : débruitage et codage à débit variable", financé par le Comité Mixte de Coopération Universiatire (CMCU).

Glossaire

Abréviations scientifiques

AAC	Advanced Audio Coding
AD	Algebraic Detector
AR	AutoRegressive
ARMA	AutoRegressive Moving Average
CM-BWE	Continuously Modulation Bandwidth Extension
DCT	Discrete Cosine Transform
DSP	Densité Spectrale de Puissance
HBE	Harmonic Bandwidth Extension
HE-AAC	High-Efficiency Advanced Audio Coding
IDCT	Inverse Discrete Cosine Transform
IS	Indice de Stationnarité
LSF	Line Spectral Frequency
MA	Moving Average
MPEG	Moving Picture Experts Group
MP3	MPEG 1 Audio Layer 3
SBR	Spectral Bandwidth Replication
SNR	Signal To Noise Ratio
SDG	Subjective Difference Grade
TCD	Transformée en Cosinus Discrète
TCDM	Transformée en Cosinus Discrète Modifiée
TFD	Transformée de Fourier Discrète
TFR	Time Frequency Representation
TNS	Temporal Noise Shaping
TM	Temporal Masking
ODG	Objective Difference Grade
VQ	Vector quantization
WM	Watermarking

Abréviations institutionnelles

ANR	Agence Nationale de la Recherche
CMCU	Comité Mixte de Coopération Franco-Tunisienne
ENIT	Ecole Nationale d'Ingénieurs de Tunis
INRIA	Institut National de Recherche en Informatique et Automatique
UFR	Unité de Formation et de Recherche
WARRIS	Watermarking Réflexif pour le Renforcement des Images et du Son

Opérations

X^T transposée de X

$a_n * b_n$ produit de convolution de a_n et b_n

$X \odot Y$ produit d'Hadamard

$E[x]$ espérance de x

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le tatouage audio (ou audio watermarking) consiste à cacher des informations à l'intérieur des sons, de manière inaudible et ceci en exploitant les modèles psycho-acoustiques du système auditif humain, en particulier la propriété de masquage. Les applications conventionnelles du tatouage audio peuvent être classées en deux catégories :

- le tatouage sécuritaire, ayant principalement pour objectif la protection des documents. Dans ce sens, les efforts se sont orientés vers une maximisation de la robustesse du tatouage aux attaques.
- le tatouage support de transmission virtuel, véhiculant des informations supplémentaires servant différents intérêts, soit pour l'auditeur (comme par exemple l'indexation et l'annotation des documents), soit pour des applications cibles (à titre d'exemple l'envoi de spots publicitaires). Dans ce cas, la contrainte de robustesse est moins forte mais un débit d'insertion élevé peut être requis.

Le rôle que nous assignons au tatouage est différent. Il s'agit d'utiliser le tatouage audio pour améliorer la qualité du signal hôte ayant subi des transformations et ceci en exploitant l'information qu'il véhicule. Ce contexte applicatif a été déjà introduit dans les travaux de [Gilloire 1998, Larbi 2005b, Baras 2005]. Dans ce cadre, le tatouage audio est considéré comme mémoire porteuse d'informations sur le signal originel. Il doit satisfaire un ensemble de contraintes dont essentiellement celles liées à la capacité d'insertion et à la robustesse aux distorions introduites par la compression MPEG et les opérations de filtrage. Les techniques de tatouage développées sont limitées en terme de capacité d'insertion ($<$ à 500 bps pour la robustesse souhaitée). Ceci implique la recherche de la représentation la plus compacte des informations les plus pertinentes pour l'application considérée.

La compression bas débit des signaux audio est une des applications visées par ce concept. Celle-ci s'appuie sur les modèles psycho-acoustiques et sur les techniques d'extension de bande. Ces techniques de compression permettent une réduction notable de débit de compression et donc de la bande passante nécessaire à la transmission d'un signal audio numérique, tout en bénificiant d'une qualité d'écoute adaptée à l'application désirée. Cependant, plusieurs études ont souligné des dégradations subies par le signal à la sortie des codeurs audio fonctionnant à bas débits [Reiss 2004, Erne 2002, Vercellesi 2007, Vitali 2011, Herre 1999].

Notre objectif est l'amélioration de la qualité audio à la sortie des systèmes de communication bas débit en utilisant le tatouage audio comme canal de transmission

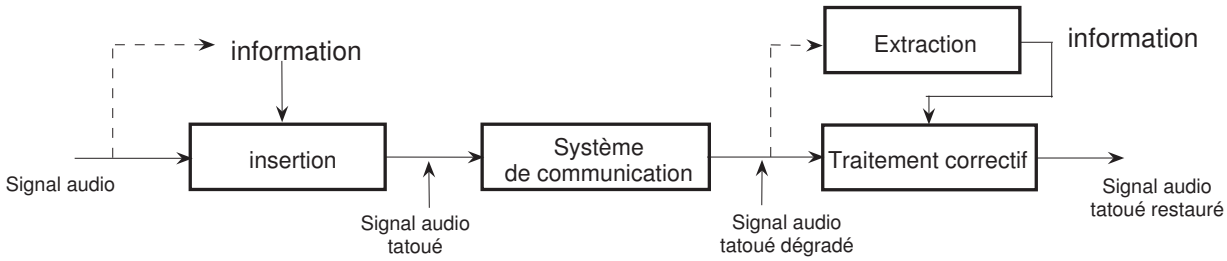


FIGURE 1 – Schéma général de l'approche proposée

virtuel. La figure 1 schématise la démarche proposée pour corriger certaines distorsions introduites par les codeurs audio bas débit. Les principaux modules de traitement développés sont représentés en gras. Les informations utiles pour la correction du signal audio à la sortie du système de communication sont insérées par tatouage dans le signal lui même. Les informations insérées servent à restaurer le signal audio par un traitement correctif spécifique.

Les codeurs HE-AAC (High Efficiency AAC, intégré dans MPEG4), AAC et MP3 sont des codeurs visés par les travaux de cette thèse. Ces systèmes sont utilisés dans diverses applications de transmission telle que la diffusion de la musique sur les réseaux mobiles 3G, la TV numérique, etc. Deux corrections seront proposées et étudiées :

- la réduction du pré-écho et de l'amollissement d'attaque, deux phénomènes introduits par les codeurs audio perceptifs, en particulier les codeurs AAC et MP3 ;
- la préservation de l'harmonicité des signaux audio dégradée par les codeurs perceptifs à extension de bande, en particulier le codeur HE-AAC.

Le présent manuscrit est structuré en deux parties, comme suit :

La **première partie** est consacrée à la présentation des principaux traitements correctifs proposés pour réduire les distortions introduites par les systèmes de compression bas débit. Dans le chapitre 1, nous présentons les principes de base des systèmes de codage bas débit et l'étude des différentes distorsions introduites par ces derniers, à savoir le pré-écho et la rupture d'harmonicité. Une première solution, détaillée au chapitre 2, vise principalement à la réduction du pré-écho et la restauration des attaques. Elle consiste à corriger l'enveloppe temporelle du signal après réception en exploitant la connaissance *a priori* de l'enveloppe temporelle du signal originel, supposée transmise par un canal auxiliaire à faible débit (< 500 bps). Une évaluation du système proposé par des critères objectifs dédiés à la mesure de la qualité audio perçue sera évoquée dans ce chapitre. La seconde solution, objet du chapitre 3, vise à corriger les ruptures d'harmonicité générées par les codeurs à extension de bande [Nagel 2009, Nagel 2010]. Ce phénomène touche essentiellement les signaux fortement harmoniques (exemple :

violon) et est perçu comme une dissonance. Un autre problème affecte essentiellement les signaux faiblement harmonique (exemple : glockenspiel). Il se traduit par une synthèse de tonales isolées en hautes fréquences complètement différentes de celle du signal original [Wolters 2003]. Une correction de tonalité des signaux audio par des opérations de translation spectrale est alors proposée, les paramètres étant là encore transmis par un canal auxiliaire à faible débit.

La **seconde partie** de ce document est consacrée à l'intégration du tatouage audio dans les techniques de renforcement des signaux audio précitées. Dans ce contexte, le tatouage audio remplace le canal auxiliaire précédent et œuvre comme une mémoire du signal originel, porteuse d'informations nécessaires pour la correction d'harmonicité et la réduction de pré-écho. Pour cela, le chapitre 4 décrit le système de tatouage additif considéré et présente ses performances en terme de robustesse à la compression MPEG dans un contexte de transmission d'informations à des débits inférieurs à 500 bps. Un premier système de réduction de pré-écho assisté par tatouage audio fait l'objet du chapitre 5. Le chapitre 6 présente un système de préservation d'harmonicité assisté également par tatouage audio.

Première partie

Techniques de renforcement de la qualité audio des systèmes de communication bas débit

Cette première partie est consacré à la présentation des principes de base des systèmes de codage bas débit, en particulier les codeurs MP3, AAC et HE-AAC et à la mise en évidence des principaux défauts de ces codeurs. Ces défauts se manifestent aussi bien sur le contenu temporel du signal (pré-écho et amolissement d'attaque) que sur le contenu fréquentiel (non préservation de tonalité et rupture d'harmonicité). Face à ces défauts, deux solutions sont proposées : réduction de pré-écho par correction d'enveloppe temporelle et préservation de tonalité par translation spectrale.

Chapitre 1

Systèmes de communications audio : problématiques liées à la compression bas débit

Sommaire

1.1 Introduction

La compression bas débit des signaux audio est une des applications visées par le concept proposé. Dans les codeurs audio développés depuis les années 90, la compression s'appuie sur les modèles psycho-acoustiques qui permettent de ne pas coder les informations non audibles et de quantifier optimalement les informations audibles. Ces techniques

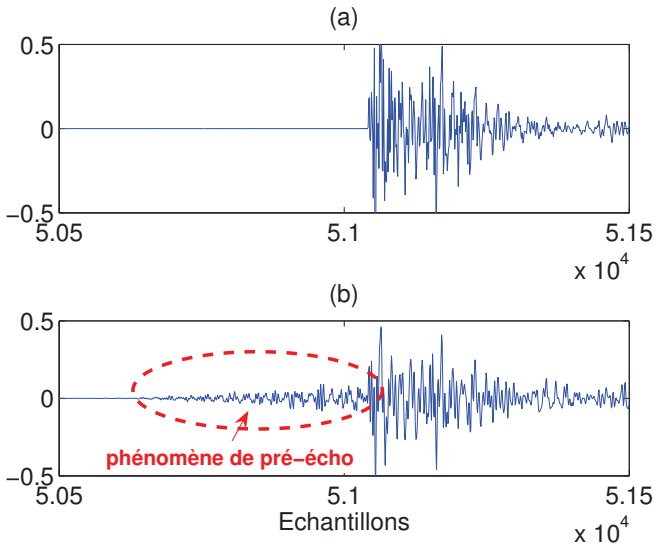


FIGURE 1.1 – Exemple de pré-écho sur un extrait de castagnettes : (a) signal original, (b) signal synthétisé par le codeur Lame à 48 kbps.

de compression reduisent efficacement le débit nécessaire à la transmission d'un signal audio tout en bénéficiant d'une qualité d'écoute adaptée à l'application désirée. Cependant, plusieurs études ont souligné des dégradations perceptivement gênantes subies par le signal après décodage.

Nous nous intéressons dans ce chapitre aux codeurs MPEG audio : codage perceptif et codage à extension de bande. En particulier, nous mettrons l'accent sur les défaillances majeures introduites par ces types de codeurs et les principales solutions apportées dans la littérature.

1.2 Codage par transformée et phénomène de pré-cho

1.2.1 Pré-écho : définition et origine

Un pré-écho est un artefact qui découle du processus de codage par transformée [Erne 2002, Reiss 2004]. Il se traduit par l'apparition d'un bruit antérieurement à une augmentation rapide de puissance. Le niveau d'énergie du pré-écho est inférieur à celui du signal pour les échantillons de forte amplitude suivant immédiatement la transition mais il est supérieur à celui du signal pour les échantillons d'énergie plus faible, notamment sur la partie précédant la transition. Dans cette dernière, le rapport signal à bruit est faible, il en découle une dégradation pouvant être très gênante à l'écoute. Le bruit qui précède la transition est appelé pré-écho alors que celui postérieur à la transition est appelé post-écho. Ce dernier est moins gênant car il y a en général moins de fins brusques

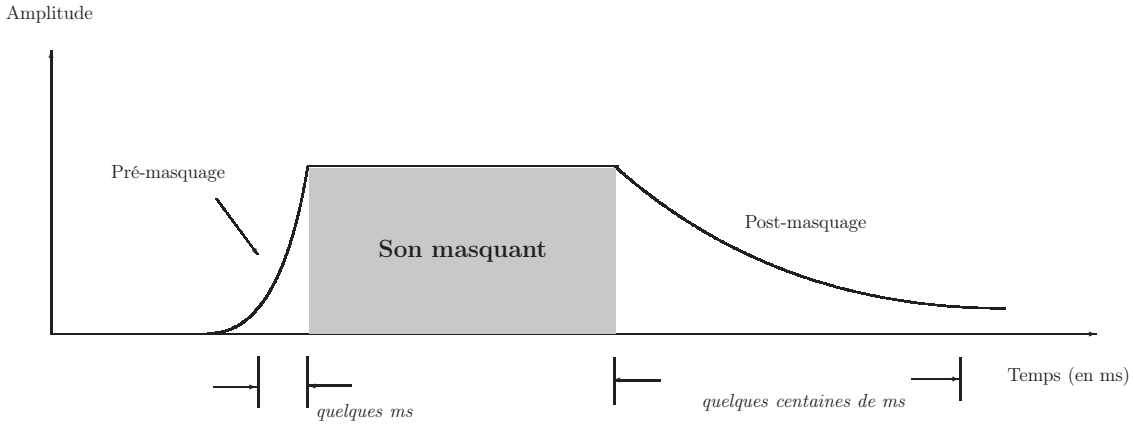


FIGURE 1.2 – Courbe de pré-masquage et de post-masquage temporel.

que d'attaques brusques dans les signaux audio.

Le phénomène de pré-écho est illustré par la figure 1.1 où nous montrons l'exemple d'une séquence de castagnettes et sa version reconstruite par le codeur/décodeur MP3 (Lame [1]) à 48 kbits/s. En comparant les figures 1.1 (a) et (b), on observe bien la présence d'un bruit qui affecte la partie précèdant la transition.

Masquage temporel et pré-écho

On parle de masquage temporel lorsqu'un son de faible puissance est masqué par l'apparition à un instant différent d'un autre son de forte puissance, appelé masquant. Si le son masquant est après le son masqué, on parle de pré-masquage ou masquage antérieur. Dans le cas contraire, on parle du post-masquage, ou masquage postérieur.

Selon [Pickett 1959], deux zones de pré-masquage et de post-masquage de durées différentes sont définies :

- la durée effective du pré-masquage est assez limitée, de l'ordre de 5 ms. Selon leur puissance, les sons survenant dans les 5 ms précédant le masquage peuvent ne pas être perçus.
- le post-masquage est beaucoup plus important, il est de l'ordre de 200 ms et dépend des caractéristiques du son masquant. Après un son fort, l'oreille pourra ne percevoir les sons les plus faibles qu'au terme de ce laps de temps.

La distinction des zones de pré et de post masquage est illustrée par la courbe de masquage temporel de la figure 1.2. Elles sont définies respectivement lors du passage des séquences de faibles énergies à des séquences de fortes énergies et inversement.

Le phénomène de masquage temporel contribue ainsi à la réduction de l'effet gênant du pré-écho introduit avant la transition. En effet, pour une durée de pré-écho inférieure

1. LAME version 3.98 (http ://www.mp3dev.org/).

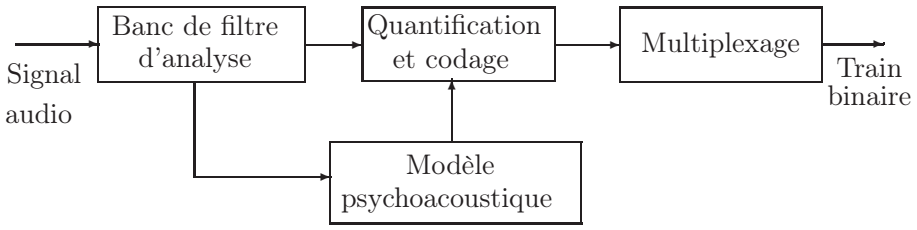


FIGURE 1.3 – Principe d'un codeur perceptif.

à la durée du pré-masquage, le bruit peut être complétement masqué. Dans le cas contraire, le pré-écho devient audible et peut en conséquence dégrader la qualité des signaux.

Afin de déterminer les origines du pré-écho, il est important d'examiner l'algorithme général d'un codeur perceptif. Nous présentons dans ce qui suit le principe de ce type de codage et nous examinerons également les trois facteurs qui contribuent à la formation du pré-écho, à savoir : le bruit de quantification, la taille de la fenêtre d'analyse et la non stationnarité du signal audio.

1.2.2 Codage perceptif par transformée

Un codeur perceptif par transformée fonctionne en analysant, à court terme, le signal dans le domaine fréquentiel. Le principe de ce codeur est schématisé par la figure 1.3. Il inclut un outil de transformation qui consiste en un banc de filtres ou une transformée. Les codeurs audio perceptifs utilisent généralement la *Transformée en Cosinus Discrète Modifiée (TCDM).*

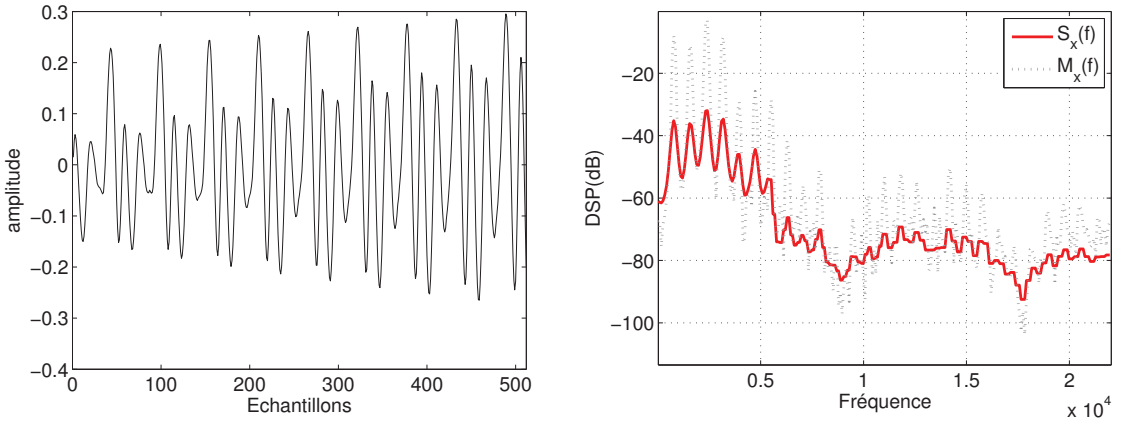
La TCDM est une transformation à valeurs réelles. Elle est équivalente à un banc de filtres modulés avec $N/M = 2$, où N et M représentent respectivement le nombre d'échantillons traités par bloc et le nombre de sous-bandes du banc de filtres.

Soient $\{x_i(n)|n = 0..N-1\}$ les N échantillons temporels du bloc d'analyse d'indice i, le spectre $X_i(k)$ obtenu par TCDM correspondant au bloc i est donné par :

$$X_i(k) = \sum_{n=0}^{N-1} h(n)x_i(n)\cos\left(\ \tfrac{2\pi}{N}(n+n_0)(k+1/2)\ \right), \quad \text{avec } n_0 = \frac{N}{4}+\frac{1}{2}, \tag{1.1}$$

où $h(n)$ représente la fenêtre d'analyse.

Les coefficients spectraux subissent ensuite différents traitements, dont principalement l'opération de quantification. Cette quantification est contrôlée par le modèle psychoacoustique qui représente la caractéristique du codage perceptif. La principale propriété psychoacoustique utilisée par le codeur perceptif est le seuil de masquage. En effet, certaines parties du signal sonore ne sont pas perçues par l'oreille humaine en présence d'autres signaux, elles sont considérées par le codeur comme insignifiantes.



(a) (b)

FIGURE 1.4 – (a) extrait d'un signal de violon (512 éch.), (b) DSP du signal audio $S_x(f)$, seuil de masquage correspondant $M_x(f)$.

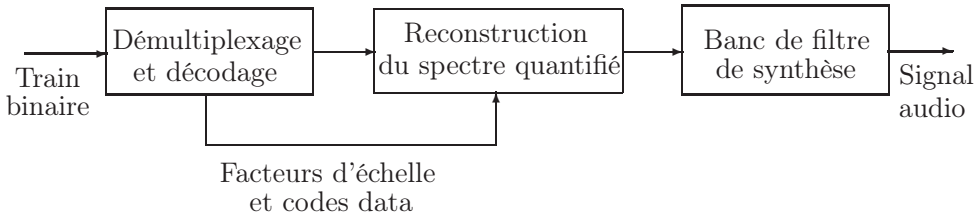


FIGURE 1.5 – Principe d'un décodeur perceptif

Le modèle psychoacoustique modélise le système auditif humain par un banc de filtres. Ce modèle analyse le signal d'entrée sur plusieurs blocs temporels consécutifs d'échantillons en transformant chaque bloc dans le domaine fréquentiel. Il modélise ensuite les propriétés de masquage de l'oreille et estime pour chaque bloc un niveau de bruit juste audible, appelé *seuil de masquage*. Ce seuil représente la limite supérieure du bruit pouvant être ajouté au signal original par les opérations de quantification, sans que ce bruit devienne audible. Sur la figure 1.4, on visualise la Densité Spectrale de Puissance (DSP) d'une trame de 512 échantillons d'un signal de violon échantillonné à 44100 Hz et le seuil de masquage correspondant calculé selon le modèle de Léandro [2][Moreau 1995].

Dans la partie quantification et codage, le codeur essaie d'allouer le nombre de bits disponibles sous deux contraintes : le débit et le seuil de masquage. Comme illsutré sur la figure 1.3, le modèle d'audition contrôle l'opération de quantification.

Le décodeur est beaucoup moins complexe, car il ne contient pas de modèle psychoacoustique ni la procédure d'allocation de bits, comme le montre la figure 1.5. Sa tâche est de reconstruire un signal audio à partir des informations contenues dans le train binaire.

2. Le modèle de Léandro est une version simplifiée du modèle n°1 de MPEG-1 layer 1.

1.2.3 Facteurs de pré-écho

La création de pré-écho résulte des conditions suivantes : le signal audio doit être suffisamment non-stationnaire dans la durée de la fenêtre d'analyse et doit aussi être suffisamment énergique pour créer un grand niveau de bruit de quantification.

1.2.3.1 Bruit de quantification

Comme mentionné précédemment, l'algorithme de quantification tire profit des modèles psycho-acoustiques pour cacher le bruit dans le signal existant. Ces modèles définissent dans chaque sous-bande un niveau de bruit acceptable en fonction du seuil de masquage. Le codeur attribue à chaque coefficient spectral (ou coefficient de sous-bande) un nombre de bits fonction du rapport signal à bruit souhaité dans la sous-bande concernée. Ce bruit, proportionnel au niveau moyen du signal, est uniformément réparti sur la totalité du support temporel.

Certains sons sont caractérisés par des attaques brusques qui se traduisent par des transitions très rapides et de grande amplitude dans le domaine temporel. C'est le cas, en particulier, pour certains sons musicaux, tels que les percussions, et pour certains sons de parole, notamment les consonnes plosives. Cette variation est accompagnée d'un changement tout aussi brusque du signal dans le domaine fréquentiel.

Le codage étant réalisé par blocs successifs d'échantillons, les transitions apparaissent donc en un point quelconque du bloc. Or en codage par transformée, le bruit de quantification est réparti temporellement de façon uniforme sur toute la durée du bloc d'échantillons. Ceci se traduit par l'apparition de pré-écho antérieurement à la transition et de post-écho postérieurement à la transition. Ainsi, un choix de taille variable de trame d'analyse s'est imposé dans le souci de réduire le phénomène de pré-écho.

1.2.3.2 Taille de fenêtre d'analyse

Le choix de la longueur de la fenêtre d'analyse implique un compromis entre la résolution temporelle et la résolution fréquentielle. Selon les techniques de codage, deux types de longueur de fenêtre sont définies : fenêtre longue pour $N = 2048$ et fenêtre courte pour $N = 256$.

Résolution fréquentielle

La résolution fréquentielle dépend de la longueur N de la trame d'analyse. A titre d'exemple, pour une fréquence d'échantillonnage de 48 kHz, la résolution fréquentielle est de l'ordre de 23 Hz (fe/N) en mode long (N=2048) et de l'ordre de 187 Hz en mode court (N=256).

Afin de garantir l'inaudibilité du bruit de quantification, une analyse par des fenêtres de type long est favorisée. En effet, ces fenêtres offrent une bonne résolution fréquentielle qui modèlise correctement l'effet de masquage des composantes tonales pour des sons stationnaires de type tonal. Les fenêtres d'analyse type court présentent une mauvaise résolution fréquentielle, ce qui rend difficile de modéliser avec précision les effets de masquage des composantes tonales. En effet, vu que la largeur des bandes critiques pour les basses fréquences est de 100 Hz, la faible résolution fréquentielle (187 Hz) se traduit par un nombre réduit de coefficients (voire nul) dans ces bandes et par conséquent, l'intérêt des modèles psycho-acoustiques est perdu et on aboutit à une quantification très grossière. En outre, le codage des courtes fenêtres pour des trames relativement stationnaires est très coûteux en terme de débit (nombre d'information à transmettre important). Pour ces raisons, il est souvent préférable d'utiliser une fenêtre de type long pour un codage par transformée des segments à caractère tonal et stationnaire. Pour les trames non stationnaires, il est souhaitable d'utiliser des fenêtres courtes qui réduisent les effets de perception du bruit de quantification.

Résolution temporelle

La résolution temporelle est inversement proportionnelle à la résolution fréquentielle. En l'absence de quantification, le codeur permet une reconstruction parfaite du signal, quelle que soit la longueur de la fenêtre. Toutefois, en présence de l'opération de quantification, l'utilisation d'une fenêtre type courte peut être utile. En effet, afin de profiter du phénomène de masquage fréquentiel du bruit de quantification, le codeur suppose que le signal audio varie lentement par rapport à la longueur de la fenêtre. Plus la fenêtre est longue, moins cette hypothèse sera vérifiée. Or les fonctions de bases de la transformée s'étendent sur la longueur de la fenêtre de sorte que le bruit de quantification est de puissance constante sur toute la fenêtre, ce qui favorise le pré-écho dans le cas d'une attaque située dans la fenêtre. Le pré-écho a d'autant plus de chances d'être court, donc masqué (masquage antérieur), que la fenêtre est courte. Nous allons développer ces aspects dans la section suivante.

La variation de la longueur de la fenêtre d'analyse en fonction de la nature du signal (stationnaire/non stationnaire) semble être une solution en faveur de l'atténuation des effets de pré-écho par le phénomène de pré-masquage temporel. Cependant, un nombre important de bits est réservé pour la transmission d'informations annexes décrivant la taille de la fenêtre d'analyse à appliquer.

1.2.4 Réduction de pré-écho

Pour réduire l'effet gênant précité du pré-echo, et dans une moindre mesure des post-échos, différentes solutions ont été proposées :

1.2.4.1 Fenêtres d'analyse à taille variable

Une première solution proposée dans [Edler 1989] consiste à réduire les pré-echos par une contraction dynamique des fenêtres. Comme la résolution temporelle et fréquentielle des signaux dépend de la longueur de la fenêtre d'analyse, les codeurs fréquentiels commutent entre les fenêtres longues, typiquement N=2048 échantillons, pour les segments stationnaires, et des fenêtres courtes, N=256 échantillons, pour des signaux à grande variation dynamique ou transitoire [3]. Le taux de recouvrement des fenêtres est de 50%. La condition de reconstruction parfaite impose des contraintes sur les choix des fenêtres d'analyse et de synthèse notées respectivement $h(n)$ et $f(n)$. Ces fenêtres sont choisies identiques avec la contrainte suivante [Moreau 1995] :

$$h(n)^2 + h(n + N/2)^2 = 1, \tag{1.2}$$

à la quelle on ajoute la contrainte de symétrie temporelle de $h(n)$:

$$h(n) = h(N - 1 - n). \tag{1.3}$$

La figure 1.6 illustre les différents types de fenêtre appliquées au cours du temps en fonction de la nature du signal (stationnaire et transitoire). La condition de reconstruction parfaite, donnée par l'équation 1.2, est vérifiée pour le recouvrement entre les trames de même type. Cependant, lors du passage d'un bloc long à un bloc court et inversement, il est impossible de réaliser un recouvrement de 50 % pour les deux fenêtres impliquées et par conséquent, la propriété de reconstruction parfaite n'est plus conservée. D'où la nécessité d'utiliser des fenêtres dites de *transition* précédant et suivant une séquence de fenêtres courtes.

L'inconvénient majeur de cette solution est qu'elle induit un retard supplémentaire de l'ordre de $N/2$ échantillons car si une transition commence dans la fenêtre suivante, il faut être en mesure de préparer la transition et de commuter sur une fenêtre de transition permettant de conserver la reconstruction parfaite.

Cas du MP3

Le codeur MPEG-1 couche 3 (MP3) définit une solution similaire à celle proposée par [Edler 1989], appelé *Masquage Temporel (TM)* [Ambikairajah 1997, Noll 2000], qui

3. Ce sont des valeurs utilisées notamment dans le cas du codage AAC.

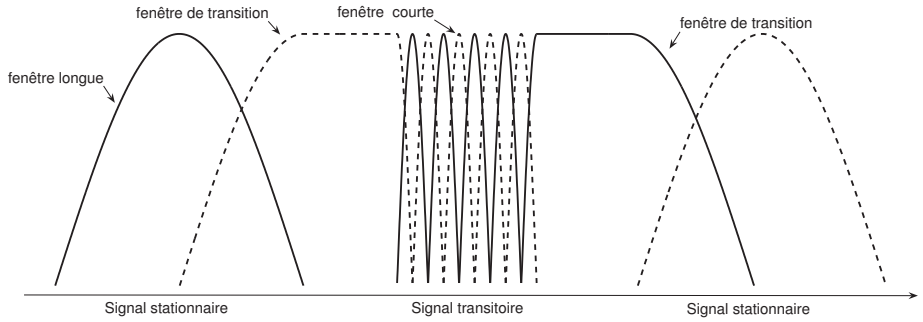


FIGURE 1.6 – Principe de l'agencement des fenêtres lors d'un changement de taille de fenêtre.

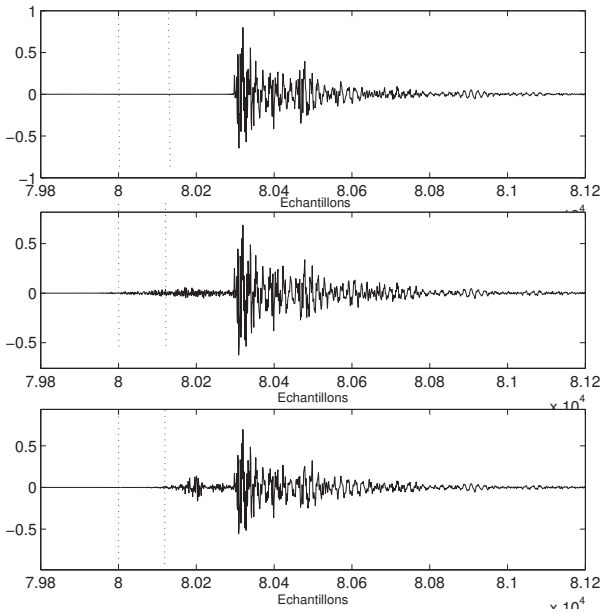


FIGURE 1.7 – Signal de castagnette (en haut), Signal reconstruit avec N=256 (au milieu), Signal reconstruit avec N=64 (en bas).

consiste à basculer entre deux fenêtres d'analyse de tailles différentes : 64 et 1024 échantillons. Les petits blocs ne sont utilisés que pour contrôler les pré-échos lors de la non stationnarité. Dans le cas contraire, le codeur reprend son traitement en utilisant des fenêtres de taille longue. Avec un tel traitement, l'erreur de quantification est rendue inaudible en faisant en sorte que la durée du bruit précédant l'attaque soit suffisament courte pour permettre un masquage antérieur. Ceci est illustré par la figure 1.7 où nous présentons une séquence de castagnettes échantillonée à 44.1 kHz et sa version codée/décodée avec lame à 64 kbits/s en utilisant deux types de fenêtres courtes : N=256 et N=64.

Lors de l'utilisation d'un système de fenêtrage adaptatif décrit précédement, le codeur doit se baser sur un critère pour décider de la longueur de fenêtre approprié. Face à ce

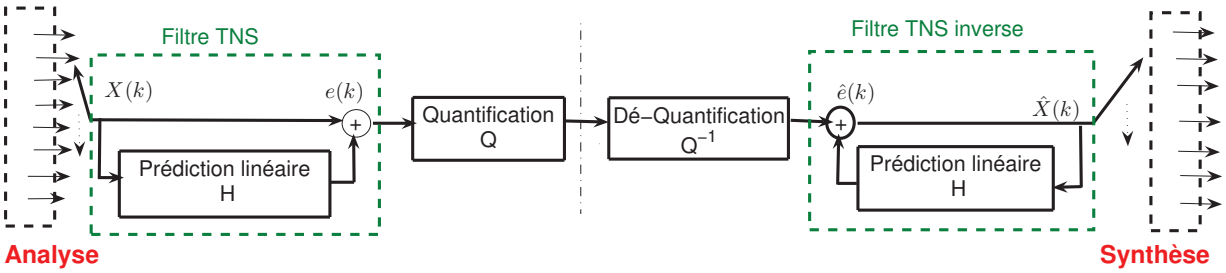


FIGURE 1.8 – Processus du codage/décodage avec la technique TNS

problème, un détecteur de transition a été conçu pour examiner les transitoires du signal audio qui pourraient conduire à des pré-échos. Lors de la détection d'une transition, le codeur bascule vers des fenêtres courtes excluant ainsi toute possiblité de génération de pré-écho.

Bien que le détecteur permette de déclencher le traitement pour la réduction de pré-écho, le détecteur d'attaque se trouve limité pour certains types de signaux transitoires comme les castagnettes, timbale, triangle ou certains types de signaux de parole. De ce fait, le pré-écho généré par le processus de codage reste nuisible.

1.2.4.2 Mise en forme temporelle du bruit

Pour réduire l'effet du pré-écho, le codeur AAC[4] définit l'option TNS[5] [Herre 1999] qui suit directement l'étape du banc de filtres d'analyse. Cette technique permet au codeur de bien contrôler la distribution temporelle du bruit de quantification.

Le principe de la technqiue TNS repose sur l'exploitation de la dualité temps/fréquence de l'analyse LPC standard. En effet, il est bien connu qu'une quantification scalaire prédictive en boucle ouverte d'un signal dans le domaine temporel se traduit par une erreur de quantification adaptée à la densité spectrale de puissance (DSP) du signal d'entrée [Jayant 1984]. En combinant cette observation avec la dualité temps/fréquence, on peut déduire que l'application d'un tel codage prédictif au contenu fréquentiel d'un signal permet de garantir une enveloppe temporelle du bruit de quantification adaptée à la distribution énergétique du signal d'entrée. En conséquence, cet effet peut être utilisé pour mettre en forme le bruit de quantification et de cette manière on réduit les problèmes liés au non-masquage du pré-écho, en particulier pour les signaux transitoires.

Le processus prédictif de codage/décodage avec la technique TNS est représenté par la

4. Advanced Audio Coding.
5. TNS : Temporal Noise Shaping.

figure 1.8. Un bloc additionnel, "Filtre TNS", est inséré après le banc de filtres d'analyse permettant d'effectuer une opération de filtrage des coefficients spectraux issus du banc de filtres. Cette opération consiste essentiellement au remplacement des valeurs spectrales (à laquelles doivent être appliquées le TNS) par l'erreur de prédiction.

Le processus de décodage TNS se fait par l'insertion d'un bloc supplémentaire, "Filtre TNS inverse", immédiatement avant le banc de filtres de synthèse (voir figure 1.8). Les valeurs spectrales résiduelles déquantifiées sont remplacées par les coefficients spectraux décodés au moyen d'un filtre prédicteur inverse tous pôles. Le processus de traitement TNS est signalé au décodeur par l'intermédiaire de données complémentaires représentées par l'information latérale comprenant une TNS on/off et des données du filtre correspondant aux coefficients de prédiction et au résidu spectral.

Ainsi, la forme temporelle du bruit de quantification est adaptée à la distribution énergétique du signal d'entrée. Alors que dans le cas du traitement standard (TNS off), le bruit de quantification est réparti presque uniformément sur toute la durée du signal. Ce contrôle de la distribution de l'énergie du bruit se trouve parfois limité, en particulier lorsque le débit est faible. Dans ce cas, la distribution du bruit de quantification n'épouse plus l'enveloppe temporelle du signal.

1.3 Codage par extension de bande : avantages et limites

La qualité audio des codages-décodages AAC et MP3 se dégrade lorsque le débit de codage diminue. En deçà de 96 kbits/s pour MP3 (mono) et 64 kbits/s pour AAC (mono), le bruit de quantification généré par le codeur dépasse le seuil de masquage et génère ainsi des artefacts audibles [Dietz 2002].

Pour maintenir une qualité audio transparente tout en réduisant le débit, plusieurs nouveaux schémas de codage ont été proposés, notamment des codages par extension de bande utilisant un codeur coeur MP3 ou AAC. Nous détaillons dans ce qui suit les principes de l'extension de bande.

1.3.1 Codage par extension de bande

La technique d'extension de bande, appelée aussi SBR[6] (Spectral Band Replication), repose sur l'exploitation de la forte corrélation entre les basses et les hautes fréquences d'un signal audio. Il est ainsi possible de reconstituer la bande haute fréquence du signal

6. La technique SBR est mise au point par Coding Technologies en 1999 dans le cadre de la normalisation du projet DRM (Digital Radio Mondiale).

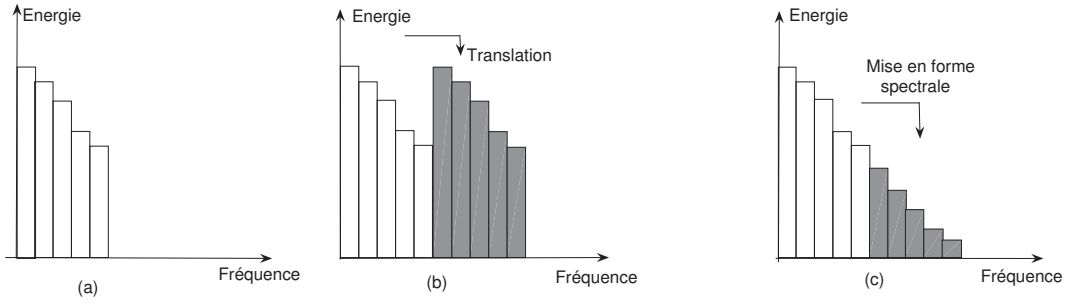


FIGURE 1.9 – (a)- Sélection de la bande basse fréquence, (b)- translation spectrale pour la régénération de la structure fine haute fréquence et (c)- ajustement de l'enveloppe spectrale du signal synthétisé.

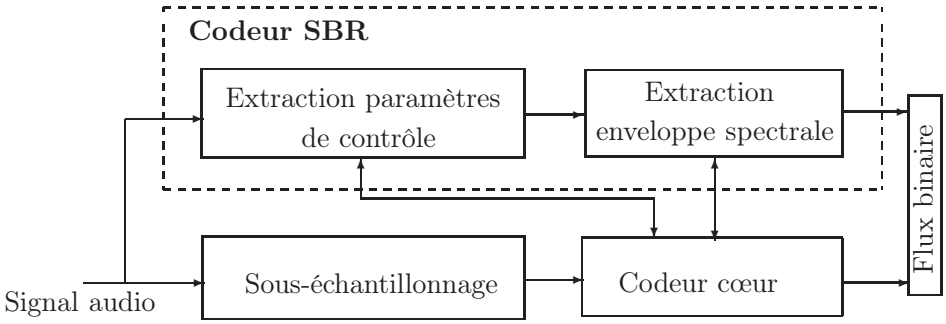


FIGURE 1.10 – Principe d'un codeur à extension de bande [Dietz 2002].

audio en s'appuyant sur la connaissance de la bande basse fréquence et de données complémentaires préalablement extraites lors de l'opération de codage et transmises à faible débit (de l'ordre de 2 kbits/s) dans un canal auxiliaire.

Le principe général de la reconstruction du spectre haute fréquence par un décodeur à extension de bande est illustré par la figure 1.9. Le processus se résume en deux étapes successives : une régénération de la structure fine de la partie haute fréquence par des opérations de translation spectrale de la bande basse fréquence en bande haute fréquence et un ajustement de l'énergie de la bande synthétisée par l'enveloppe spectrale transmise par le codeur. Ainsi, les traitements liés au processus SBR sont présents aux deux extrêmités de la chaîne de transmission : côté codage et côté décodage.

Codage SBR

Bien que la technique SBR soit principalement considérée comme étant un post-traitement au niveau du décodeur, des paramètres de contrôle cruciaux sont extraits par le codeur, permettant une regénération d'un signal haute fréquence perceptivement

le plus similaire possible à la bande haute fréquence du signal original.

Le processus de codage SBR est représenté par la figure 1.10 . La combinaison SBR avec le codeur cœur (soit par exemple le codeur AAC pour le profil High Efficiency AAC, (HE-AAC[7])) est un système à double vitesse où le codeur cœur, chargé du codage du contenu basse fréquence, fonctionne à la moitié de la fréquence d'échantillonnage du codeur SBR.

Dans le système SBR où le signal à large bande est disponible, les paramètres complémentaires à extraire sont des données relatives à la représentation de l'enveloppe spectrale de la partie haute fréquence non transmise par le codeur cœur. Les descripteurs de l'enveloppe spectrale sont calculés selon une résolution temps-fréquence variable permettant de bien suivre l'évolution temporelle du signal et offrant la possibilité de bien contôler son contenu fréquentiel. Les paramètres de contrôle visent principalement à contrôler le rapport tonales à bruit des bandes hautes fréquences à synthétiser.

Décodage SBR

D'un point de vue un peu plus technique, le processus de décodage SBR est composé d'un ensemble de modules représentés par la figure 1.11. Tout le traitement est fait dans le domaine fréquentiel, par conséquent, à la sortie du décodeur cœur, le signal est injecté dans un banc de filtres d'analyse QMF (Quadrature Mirror Filter) de 32 canaux [Ekstrend 2002]. Les sous-bandes basses fréquences sont translatées en sous-bandes hautes fréquences par l'intermédiaire du Générateur haute fréquence permettant ainsi de synthétiser la structure fine des bandes hautes fréquence non transmises. L'ajustement de l'amplitude du spectre haute fréquence regénéré est assurée par l'enveloppe spectrale et le facteur de gain de contrôle transmis via le flux SBR permettant ainsi de maintenir l'energie et l'enveloppe spectrale du signal synthétisé similaires à celles du signal original. Outre l'ajustement de l'enveloppe, les composantes tonales et bruitées sont ré-équilibrées conformément au signal original en utilisant les paramètres de contrôle transmis par le flux SBR. Toutes les opérations précitées sont effectuées dans le domaine fréquentiel, l'étape finale de l'opération de décodage est une synthèse QMF pour revenir au domaine temporel.

7. Le codeur AAC assosié à la technologie SBR, mis au point par Coding Technologies, constitue l'un des systèmes de codage perceptif les plus puissants, assurant des signaux audio de haute qualité à des débits de l'ordre de 24 kbits/s en mono et de l'ordre de 48 kbits/s en stéréo. Le format MP3Pro est également un autre exemple d'application de la technique SBR associée au codeur MP3 (MPEG-1 couche 3).

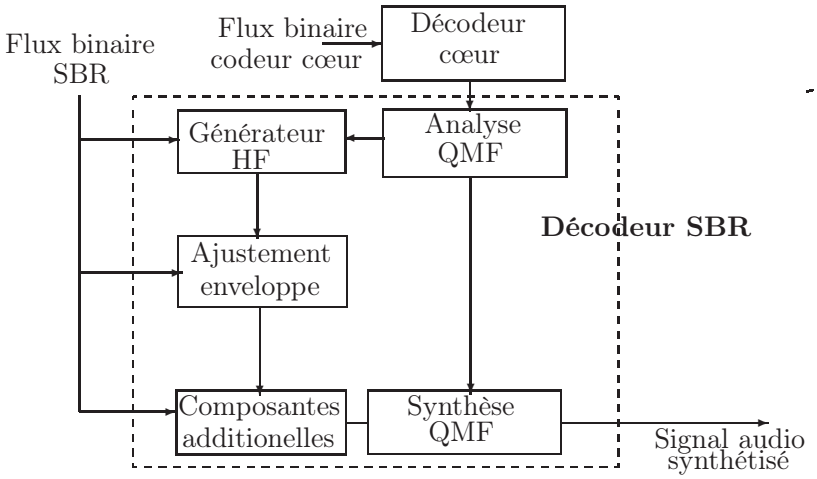


FIGURE 1.11 – Principe d'un décodeur à extension de bande [Dietz 2002].

1.3.2 Limites des codeurs à extension de bande

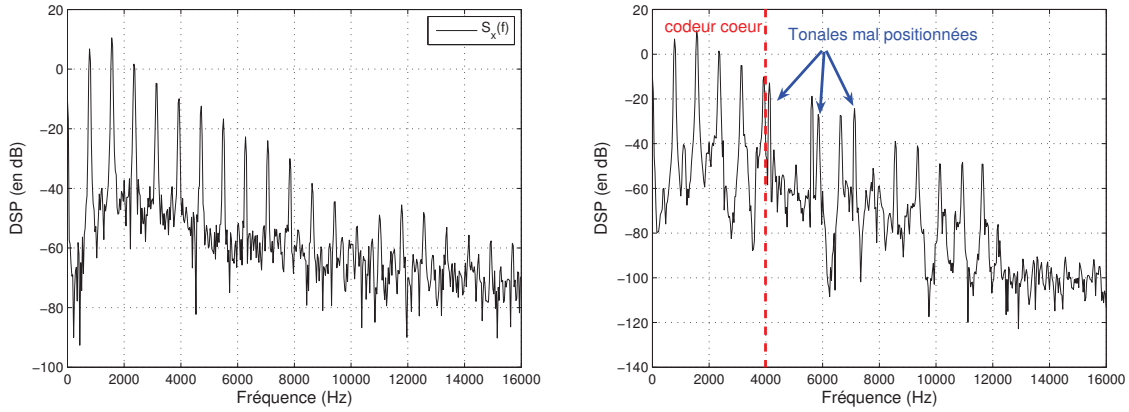
La technique SBR associée au codeur perceptif réduit efficacement le nombre de bits nécessaire pour le codage de la bande haute fréquence, tout en maintenant un signal perceptivement similaire à l'original. Cependant, cette technique présente des inconvénients majeurs concernant essentiellement le module d'extension de la structure fine. En effet, la reconstruction de la structure fine haute fréquence ne garantit ni le respect de l'harmonicité du signal original, ni la reconstruction des tonales isolées. Deux principaux phénomènes en découlent : phénomène de rugosité et synthèse de tonales isolées.

1.3.2.1 Rupture d'harmonicité et phénomène de rugosité

En présence de signaux bruités et/ou faiblement harmoniques, le module de regénération de la bande haute fréquence étend d'une manière efficace la structure fine du spectre sans dégradation gênante. Toutefois, pour les signaux fortement harmoniques, la technique développée se trouve limitée en ce sens que les tonales synthétisées sont mal placées dans le spectre du signal regénéré. Les duplications sont réalisées indépendamment de la fréquence fondamentale. Il en découle dans la majorité des cas des ruptures d'harmonicité.

La figure 1.12 illustre un exemple de rupture d'harmonicité sur une séquence de trompette échantillonnée à 44.1 kHz. Le spectre du signal original présente une harmonicité régulière avec une fréquence fondamentale de 780 Hz et 20 tonales étalées sur une largeur de bande de 16 kHz. Sur la version du signal synthétisée par le codeur aacPlus à 16 kbits/s, l'harmonicité n'est plus respectée. On note une déviation des composantes harmoniques à partir de la sixième tonale.

Les composantes tonales résultant de cette déviation peuvent entrer en dissonance et



(a) (b)

FIGURE 1.12 – (a)- DSP d'une séquence de 1024 ech. d'un signal de trompette échantillonné à 32 kHz, (b)- DSP de la séquence codée/décodée par le aacPlus à 16 kbits/s.

génèrent ainsi des artefacts sonores perceptivement gênants. En effet, l'emplacement des pics sonores synthétisés à proximité immédiate les uns des autres peut conduire à une modulation d'amplitude donnant naissance au phénomène de rugosité.

Selon Plomb [Plomb 1965], le phénomène de rugosité est considéré gênant si l'écart entre deux composantes tonales est approximativement contenue dans 5 à 50% de la largeur de la bande critique dans laquelle elles sont situées. D'après [Plomb 1965], la largeur de bande critique pour une fréquence donnée peut être approchée par :

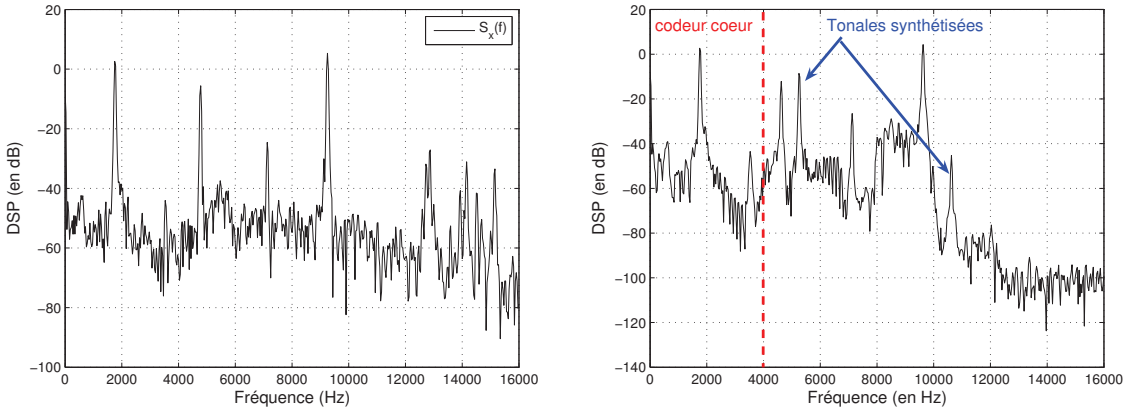
$$cb(f) = 25 + 75 \left(1 + 1.4 \left(\frac{f}{1000} \right)^2 \right)^{0.69} \tag{1.4}$$

Selon [Helmholtz 1954], l'écart entre deux composantes tonales se traduit perceptivement par :

- une qualité fortement dégradée pour des écarts faibles inférieures à 20 Hz. En effet, les tonales interagissent entre elle et génèrent un phénomène de battement gênant ;
- un phénomène de rugosité est perçu pour un écart entre 20 et 200 Hz ;
- Pour un écart supérieur à 200 Hz, la phénomène de rugosité est moins perceptible, amenant à une qualité acceptable.

1.3.2.2 Synthèse de tonales isolées

Le problème de la synthèse des tonales isolées concerne toujours le module d'extension de la structure fine du spectre haute fréquence et plus précisément la recontruction du spectre par duplication. La technique développée est actuellement limitée en ce sens qu'elle peut générer des tonales isolées en haute fréquence complètement différentes de



(a) (b)

FIGURE 1.13 – (a)- DSP d'une séquence de 1024 ech. d'un signal de glockenspiel échantillonné à 32 kHz, (b)- DSP de la séquence codée/décodée par le aacPlus à 16 kbits/s.

celles du signal original. Ce problème touche essentiellement les signaux faiblement harmoniques.

La figure 1.13 illustre ce phénomène pour une séquence de glockenspiel échantillonnée à 32 kHz et sa version codée/décodée par le aacPlus à 16 kbits/s. Le rapport tonale à bruit des sous-bandes hautes fréquences regénérées est complètement différent de celui du signal original. Sur la version codée/décodée, on note une regénération de trois composantes tonales, respectivement aux fréquences 2,6, 5,2 et 10,6 kHz, qui n'existent pas dans le signal original.

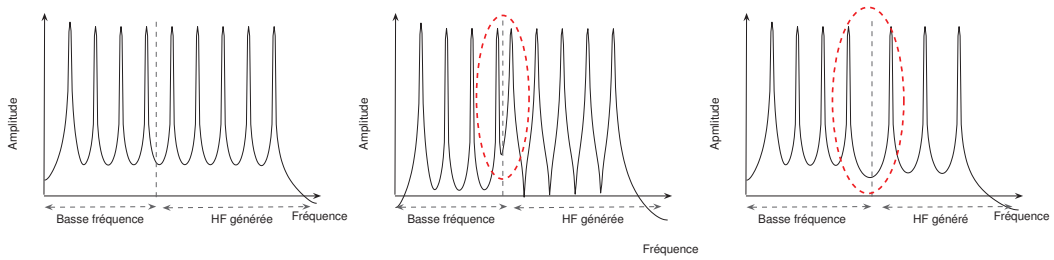
1.3.3 Approches existantes pour la correction d'harmonicité

Pour préserver l'harmonicité lors du codage à extension de bande, plusieurs solutions de différentes complexités ont été proposées dans la littérature :

Extension de la bande passante par étalement spectral

La solution de l'extension de bande par étalement spectral, (*Harmonic Bandwidth Extension*, HBE), a été proposé [Nagel 2009]. La technique repose sur la regénération de la structure fine du spectre haute fréquence par de multiple opérations d'étalement spectral de la bande basse fréquence réalisées par l'utilisation des vocodeurs de phase. Comme illustré sur la figure 1.14, un traitement avec un tel processus évite le phénomène de rugosité mentionné ci-dessus.

Le principe de fonctionnement d'un vocodeur de phase est composé des étapes suivantes [Allen 1977, Dolson 1986]. La première étape consiste à transformer le signal dans



(a) (b) (c)

FIGURE 1.14 – Synthèse de la bande haute fréquence par : (b) SBR et (c) HBE.

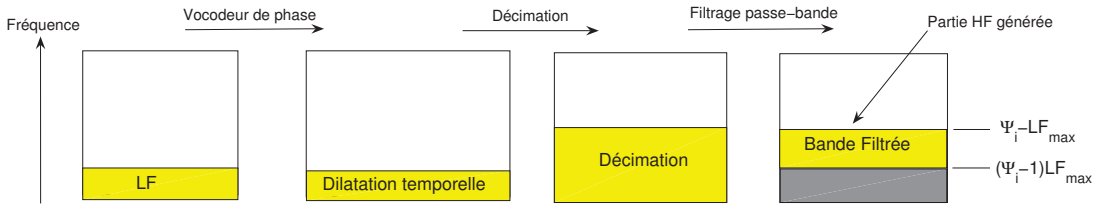


FIGURE 1.15 – Etapes de traitement de la technique d'extension de la bande passante harmonique (HBE) [Nagel 2009].

le domaine fréquentiel par DFT [8]. Dans une étape suivante, toutes les phases des composantes spectrale de la bande basse fréquence sont multipliées par un facteur d'étalement Ψ. Cette opération se traduit également par une dilatation dans le domaine temporel. Afin de ne pas altérer la distribution temporelle du signal, une opération de décimation par le facteur Ψ suit directement l'opération du passage au domaine temporel par IDFT [9].

La méthode HBE utilise de multiples vocodeur de phases, fonctionnant en parallèle afin de synthétiser la partie haute fréquence (HF). Les étapes de traitement d'une telle approche sont illustrées par la figure 1.15. La partie basse fréquence du signal de largeur de bande LF_{max} subit plusieurs étalements spectraux de facteur Ψ_i. En une seconde étape, le signal ainsi reconstruit est sous-échantillonnés par le facteur Ψ_i puis filtré par un filtre passe-bande de largeur de bande $[(\Psi_i - 1)LF_{max} : \Psi_i LF_{max}]$. La valeur maximale du facteur d'étalement Ψ_{max} est déterminée en fonction de la largeur de bande maximale LF_{max} à synthétiser.

Comme le cas de la technique SBR, la partie haute fréquence (HF) ainsi générée subit un traitement correctif, permettant de maintenir l'enveloppe spectrale et la tonalité de partie haute fréquence regénérée similaire à celle du signal original.

La technique HBE a été jugée bénéfique pour de nombreux types de signaux de mu-

8. Discrete Fourier Transform.
9. Inverse Discrete Fourier Transform.

sique, en particulier fortement harmoniques. Cependant, la technique présente deux inconvénients majeurs :

- Pour les signaux de parole, la méthode SBR reste la mieux adaptée à ces types de signaux. Ceci pourraît être expliqué par le fait que la technique HBE ne permet pas de garantir un respect exacte de la périodicité spectrale des signaux voisés.
- Pour les signaux audio, un nombre important des harmoniques résultantes est supprimée par l'application de la HBE, ce qui conduit à la non préservation du timbre.

Extension de bande par modulation continue

Dans le même objectif de la préservation de l'harmonicité du signal audio, Nagel [Nagel 2010] a proposé une deuxième méthode d'extension de bande appelée extension de bande par modulation continue (CM-BWE [10]). La méthode proposée, opérant dans le domaine temporel, est fondée sur le calcul de la fonction d'autocorrélation du spectre d'amplitude. Nous présentons dans ce qui suit le principe de cette technique.

Un signal harmonique est constitué d'une fréquence fondamentale F_0 et de ses multiples entiers $S = l.F_0$. Si on translate une partie du spectre d'un décalage de fréquence fixe LF_{max}, comme effectué par la technique SBR, on obtient $\widehat{S} = l.F_0 + LF_{max}$. Ce décalage est défini comme étant égal à la fréquence de transition entre la partie basse fréquence (BF) et la partie haute fréquence (HF) générée. Afin de conserver la structure harmonique originale, un réglage fin de ce décalage est nécessaire. Par conséquent, il s'agit de trouver un minimum δ vérifiant la relation :

$$i.F_0 = j.F_0 + LF_{max} + \delta; \quad i, j, \delta \in \mathbb{N}, \quad \delta < LF_{max}. \tag{1.5}$$

Ceci se traduit par l'égalité suivante :

$$\delta = k.F_0 - LF_{max}. \tag{1.6}$$

Si F_0 est connue, δ peut être déterminée. Dans le cas contraire, une manière simple de maximiser l'intercorrélation $R(\nu)$ entre le spectre BF et sa version translatée de $\frac{LF_{max}}{2}$, comme illustré sur la figure 1.16, définie par :

$$R(\nu) = \sum_{\omega=\omega_{min}}^{\omega=\omega_{max}} \Big| \; S(w) \; \Big| \Big| \; S(w + \tfrac{LF_{max}}{2} + \nu) \; \Big|. \tag{1.7}$$

La valeur de δ se déduit alors par :

$$\delta = arg \max_{\nu \in Z} R(\nu) \tag{1.8}$$

10. CM-BWE : Continuously Modulated Bandwidth Extension.

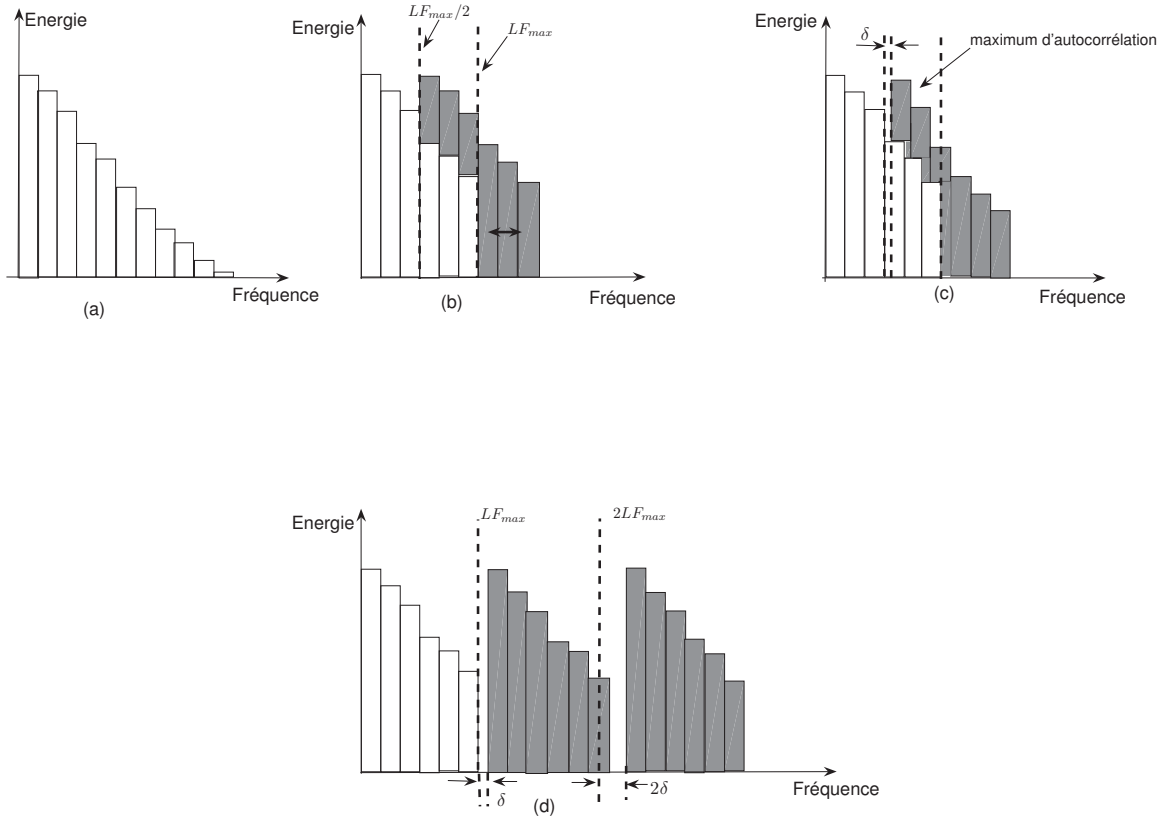


FIGURE 1.16 – Etapes de traitement de la technique CM-BWE.

Le décalage de fréquence δ est utilisé pour générer la structure fine de la partie haute fréquence conformément à la méthode illustrée par la figure 1.16d. La translation est réalisée par une modulation à bande latérale unique. Une mise en forme spectrale est par la suite appliquée sur la partie haute fréquence générée, similaire à celle utilisée par la technique SBR.

Comparée à la technique SBR, la méthode d'extension de bande par modulation continue garantit une préservation du caractère harmonique des signaux audio avec une complexité réduite par rapport à la solution HBE proposée dans [Nagel 2009]. Cependant, pour les signaux non harmoniques, aucun avantage n'est enregistré par la HBE par rapport à la SBR.

1.4 Conclusion

Dans ce chapitre, nous avons exposé les techniques de codage perceptif dédiées à la compression large bande. Ces techniques de compression exploitent des modèles psychoacoustiques, dont principalement la propriété de masquage. Ces techniques réduisent

considérablement le débit tout en maintenant une bonne qualité de reconstruction. En exploitant la corrélation entre les différentes parties du spectre, nous avons montré également qu'il est possible de restituer la bande haute fréquence à partir de la bande basse fréquence. Une réduction significative de débit est par conséquent obtenue.

Pour certains types de signaux, les stratégies de compression, présentées dans ce chapitre, introduisent des artefacts audibles :

- pré-écho : se produit essentiellement sur les signaux audio à caractère percussif ;
- rugosité : touche en particulier les signaux audio fortement harmoniques et découle du processus de la restitution des hautes fréquences par les duplications spectrales.
- synthèse des tonales isolées en haute fréquence : se traduit par le non respect du rapport tonal à bruit sur les signaux générés. Ce problème est lié essentiellement aux signaux à caractère tonal non harmonique.

Plusieurs techniques de correction des artéfacts précités sont présentées dans la littérature. Ces méthodes présentent cependant quelques limites. Nous proposons dans les chapitres qui suivent, deux nouvelles méthodes de réduction de pré-écho et de correction d'harmonicité.

Chapitre 2

Réduction de pré-écho par correction d'enveloppe temporelle

Sommaire

2.1 Introduction

La technique de réduction de pré-écho proposée repose sur la correction du signal à la sortie du décodeur audio. Elle vise la remise en forme de l'enveloppe temporelle du signal décodé afin de corriger le pré-écho résiduel et l'amolissement des attaques. Notons que cette technique diffère de la technique TNS présenté dans le paragraphe 1.2.4.2 du chapitre 1. Cette dernière consiste essentiellement en la mise en forme du bruit de quantification par l'enveloppe temporelle transmise à travers les paramètres AR (modélisation AutoRégressive) [Herre 1999].

Notons que l'estimation de l'enveloppe temporelle doit être réalisée :

- sur le signal original : la technique requiert une estimation d'enveloppe sur le signal original à l'entrée du codeur et une transmission (à très bas débit) des paramètres décrivant cette enveloppe sur un canal auxiliaire (voir Figure 2.1).
- sur le signal codé/décodé : la technique requiert une estimation d'enveloppe sur le signal décodé et une correction du signal décodé en exploitant les paramètres envoyés sur le canal auxiliaire.

  Le signal issu du décodeur subit une correction, fondée sur le rapport entre les enveloppes comme suit :

$$\widehat{x}(t) = x_{decode}(t)\frac{\hat{e}(t)}{\hat{e}(t)_{dec}}, \tag{2.1}$$

  avec
  - $\hat{e}(t)$: estimée de l'enveloppe temporelle du signal source ;
  - $\hat{e}(t)_{dec}$: estimée de l'enveloppe temporelle du signal décodé.

Ainsi, l'approche de correction proposée nécessite des traitements aux extrémités de la chaîne, côté codage et côté décodage. Ceci est illustré par le schéma de la figure 2.1.

Nous développons dans ce qui suit les détails de chaque module, à savoir la modélisation de l'enveloppe temporelle par prédiction linéaire dans le domaine fréquentiel, le codage de cette enveloppe et la correction.

2.2 Modélisation de l'enveloppe temporelle par prédiction linéaire fréquentielle (FDLP)

La modélisation d'enveloppe temporelle se fonde essentiellement sur la prédiction linéaire dans le domaine fréquentiel (FDLP). Nous reprenons dans ce qui suit les détails

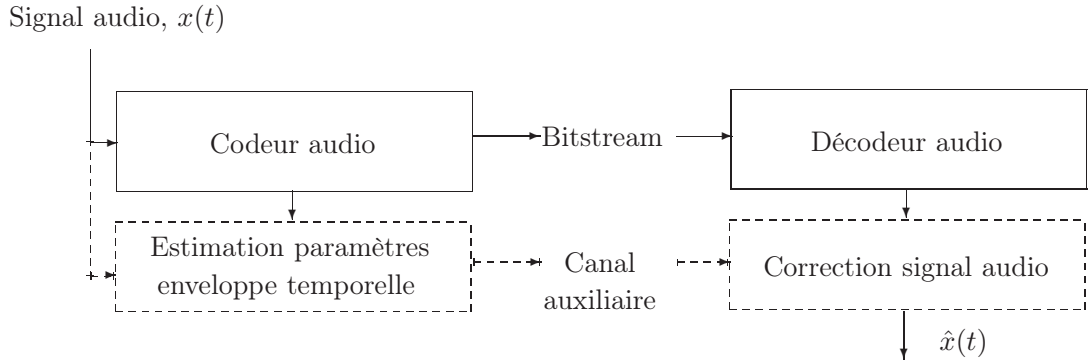


FIGURE 2.1 – Schéma général du système proposé.

de cette technique développés dans [Athineos 2007] après avoir présenté quelques rappels théoriques.

2.2.1 Signal analytique discret et enveloppe temporelle

Le signal analytique a été introduit par Gabor [Gabor 1946]. Sa propriété fondamentale est que son spectre est nul pour les fréquences négatives, en d'autres termes, il est «causal» dans le domaine fréquentiel.

Dans le cas discret, le spectre est «périodiquement causal» [Oppenheim 1999] ce qui signifie que la moitié de chaque répétition périodique du spectre est forcée à zéro. Marple [Marple 1999] a utilisé cette définition afin de dériver un signal analytique discret dans le temps en utilisant la Transformée de Fourier Discrète (TFD).

Dans le domaine temporel, le signal analytique est complexe avec sa partie réelle étant le signal d'origine et sa partie imaginaire étant la transformée de Hilbert du signal original [Cohen 1995] :

$$x_a(t) = x(t) + \mathrm{j}\ \widehat{x}(t). \tag{2.2}$$

$\widehat{x}(t)$ est la transformée de Hilbert du signal $x(t)$, il est donné par :

$$\widehat{x}(t) = x(t) * \frac{1}{\pi t}, \tag{2.3}$$

où $*$ représente le produit de convolution. **Le module du signal analytique représente l'enveloppe temporelle du signal** $x(t)$. Ce résultat est illustré sur la figure 2.2 où on visualise l'évolution temporelle d'un signal $x(t)$ et l'enveloppe temporelle correspondante.

Soit $\mathtt{x}$ un vecteur colonne réel de dimension impaire M, $M = 2N - 1$, représentant une durée finie d'un signal réel à temps discret $x(n)$:

$$\mathtt{x} = [x(0)\ \ x(1)\ \ \ldots\ \ x(M-1)]^T, \tag{2.4}$$

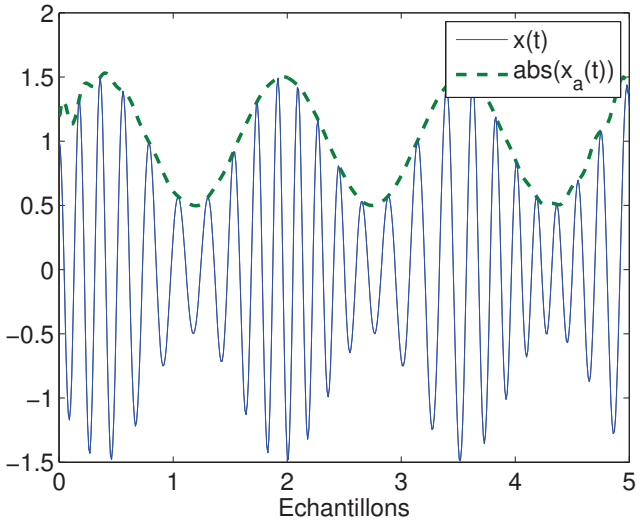


FIGURE 2.2 – Exemple d'un signal temporel et enveloppe temporelle correspondante.

où $\{.\}^T$ désigne la transposition. La conversion du signal $\mathbf{x}$ vers sa version analytique peut exprimée par une transformation matricielle donnée par [Athineos 2007] :

$$\mathbf{x}_a = A\mathbf{x}, \tag{2.5}$$

où la matrice $A_{M\times M}$ est definie par :

$$A = F^H(2W^2)(ZZ^T)F, \tag{2.6}$$

$W_{M\times M}$ est une matrice diagonale, définie par :

$$W = \begin{pmatrix} 1/\sqrt{2} & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & I_{M-1} \end{pmatrix}, \tag{2.7}$$

$Z_{M\times N}$ est une matrice zero-padding à droite , définie par :

$$Z_{M\times N} = [I_N 0_{N\times N-1}]^T, \tag{2.8}$$

où $I_{N\times N}$ est la matrice identité et $F_{M\times M}$ est la matrice TFD, définie par :

$$F(m,n) = \frac{1}{\sqrt{M}} \exp\left(-\mathrm{j}\tfrac{2\pi mn}{M} \right), \qquad m,n = 0,1,...,M-1, \tag{2.9}$$

La matrice F vérifie les relations suivantes : $F^{-1} = F^H$ et $F^H = F^*$.

La matrice A est interprétée comme suit : après avoir pris la DFT (F) du signal d'entrée, les fréquences négatives seront mises à zéro en multipliant par ZZ^T forçant ainsi

le spectre à être périodiquement causal. Ensuite, pour assurer l'orthogonalité des parties réelle et imaginaire, une mise à échelle par les poids appropriés ($2W^2$) [Marple 1999] est appliquée. Enfin nous prenons la TFD inverse (F^H) pour revenir au domaine temporel.

L'enveloppe temporelle peut être approchée par une modélisation Autorégressive (AR). En effet, les modèles AR sont couramment utilisés pour obtenir une estimation tout-pôles du spectre de puissance du signal. En exploitant la dualité temps/fréquence, une relation entre l'autocorrélation du spectre d'un signal et l'enveloppe temporelle correspondante est déduite. Nous présentons dans ce qui suit le principe de l'approche.

2.2.2 Transformation en Cosinus Discrète et enveloppe temporelle

La Transformée en Cosinus Discrète (TCD) est largement utilisée dans les opérations de compression des signaux. Elle est une transformation à valeurs réelles possédant de bonnes propriétés fréquentielles. Sur les 16 transformations trigonométriques discrètes, on s'intéresse dans cette partie à la TCD impaire type I (notée TCD-Io) qui est la seule liée à la TFD par l'opérateur WSHS SEO [1][Martucci 1994].

Soit $\mathtt{x}$ un signal à N échantillons temporels. Le spectre, X_{TCD}, obtenu par TCD-Io de $\mathtt{x}$ est donné par l'écriture matricielle suivante :

$$X_{TCD} = C\mathtt{x}, \tag{2.11}$$

où $C_{N\times N}$ est la matrice de Transformée en Cosinus Discrète (TCD-Io), définie par :

$$C(m,n) = \frac{2}{\sqrt{M}} k_m k_n \cos\left(\frac{2\pi mn}{M}\right) \qquad M = 2N-1, \tag{2.12}$$

avec $m, n = 0, 1, ..., N-1$ et les coefficients k_j valent :

$$k_j = \begin{cases} 1/\sqrt{2} & j \neq 0 \\ 1 & j = 0 \end{cases} \tag{2.13}$$

C est une matrice orthogonale qui peut être factorisée comme suit :

$$C = W(Z^T F S)W^{-1}, \tag{2.14}$$

1. WSHS SEO : left Whole-sample Symetric, right Half-sample Symetric Symetric Extension Operator. D'après [Martucci 1994], un vecteur $\overline{x}$ de longueur M, $M = 2N-1$, est dit WSHS symétrique si :

$$\overline{x}(n) = \begin{cases} x(n) & n = 0, 1, .., N-1 \\ x(M-n) & n = N, ..., M-1 \end{cases} \tag{2.10}$$

où $W_{N\times N}$ est une matrice diagonale, définie par :

$$W = \begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & I_{N-1} \end{pmatrix}, \tag{2.15}$$

et $S_{M\times N}$, $M = 2N-1$, est une matrice donnée par :

$$S = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ & & & 1 \\ & & & \\ & 1 & & \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{N-1} \\ 0 & J_{N-1} \end{pmatrix}. \tag{2.16}$$

J_{N-1} est la matrice d'identité inversée de dimension $N-1\times N-1$. Les détails de cette factorisation sont exposés dans l'annexe C.

Notons que la matrice $F_{M\times M}$ est complexe alors que la matrice $C_{N\times N}$ est réelle. La matrice $W_{N\times N}$ est inversible puisque elle est non singulière et diagonale. L'interprétation de l'équation 2.14 est que la TCD-Io correspond aux N premiers éléments (Z^T) de la TFD (F) de la matrice d'entrée WSHS-symétrique S. Afin de rendre les colonnes et les lignes orthogonales, on multiplie à gauche et à droite respectivement par W et W^{-1}.

Bien que la matrice C soit orthogonale, on peut en tirer une matrice non orthogonale, notée C_F, et son inverse correspondant, noté C_I. En effet, nous avons :

$$\begin{aligned} I_N &= CC^T && (2.17)\\ &= W(Z^TFS)W^{-1}W^{-1}(S^TF^HZ)W && (2.18)\\ &= W^2(Z^TFS)W^{-2}(S^TF^HZ) && (2.19) \end{aligned}$$

Les matrices C_F et C_I seront définies alors par :

$$\begin{aligned} C_F &= 2W^2(Z^TFS), && (2.20)\\ C_I &= \frac{1}{2}W^{-2}(S^TF^HZ). && (2.21) \end{aligned}$$

et on $C_FC_I = I_N$.

On note que, comme la matrice orthogonale C, la matrice C_F inclut le terme Z^TFS, *i. e.* une troncature de la transformée de Fourier de la séquence symétrique WSHS. La principale différence est que, dans C_F, le vecteur avant traitement est exactement la séquence d'entrée, alors qu'en C, l'intervention du facteur W^{-1} modifie forcement le vecteur d'entrée avant la transformation. On note également que W ne concerne que le premier élément du signal.

On définit $\mathtt{y}$ la partie non orthogonale de la TCD-Io du signal $\mathtt{x}$ par :

$$\mathtt{y} = C_F\mathtt{x}. \tag{2.22}$$

On pose $\widehat{y}$ la Transformée de Fourier Discrète inverse (TFD^{-1}) du signal $\mathtt{y}$ zéro paddé. En utilisant la relation 2.20 et la matrice A de la transformation analytique donnée par l'équation 2.6, on peut écrire $\widehat{\mathtt{y}}$ comme suit :

$$\begin{aligned}
\widehat{y} &= F^H Z y && (2.23)\\
&= F^H Z C_F \mathtt{x} \\
&= F^H Z (2W_N^2 Z^T F S)\mathtt{x} && (2.24)\\
&= F^H (2W_M^2)(ZZ^T) F S\mathtt{x} && (2.25)\\
&= AS\mathtt{x} && (2.26)
\end{aligned}$$

Le passage de l'équation 2.24 vers 2.25 vient du fait que $ZW_N^2 = W_M^2$ moyennant un changement de dimension approprié de W^2 de $N \times N$ à $M \times M$.

L'interprétation importante déduite de cette formule est que la Transformée de Fourier Discrète inverse (TFD^{-1}) du signal TCD-Io zéro paddé de $\mathtt{x}$ ($C_F\mathtt{x}$) est égale au signal analytique WSHS symétrisé.

Autocorrélation et TCD

Soit $\mathtt{x}$ un signal à N échantillons temporels. Le vecteur autocorrélation linéaire de $\mathtt{x}$, $\widetilde{r}_{\mathtt{x}}$, est donné par :

$$\widetilde{r}_{\mathtt{x}} = [\widetilde{r}_x(-N+1) \ \ldots \widetilde{r}_x(-1)\widetilde{r}_x(0)\widetilde{r}_x(1) \ \ldots \widetilde{r}_x(N-1)]^T, \tag{2.27}$$

avec

$$\widetilde{r}_x(m) = \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)x(n+|m|), \qquad |m| \leq N-1. \tag{2.28}$$

$\widetilde{r}_{\mathtt{x}}$ est un vecteur de longueur impaire M, $M = 2N-1$, et symétrique par rapport au point $m = 0$ ($\widetilde{r}_x(m) = \widetilde{r}_x(-m)$). On définit $\widehat{\mathtt{x}} = FZ\mathtt{x}$, F de dimension $M \times M$ et Z de dimension $M \times N$, la TFD du signal d'entrée $\mathtt{x}$ après zero-padding, l'autoccorrélation de l'équation 2.28 devient [Papoulis 1985] :

$$\widetilde{r}_{\mathtt{x}} = \frac{\sqrt{M}}{N} F^H(\widehat{\mathtt{x}} \odot \widehat{\mathtt{x}}^*), \tag{2.29}$$

où $\{.\}^*$ désigne le complexe conjugué et $A \odot B$ désigne le produit d'Hadamard (produit élément par élément, voir annexe C). La relation 2.29 correspond au théorème de Wiener-Khinchine.

Par analogie avec l'expression de l'autocorrélation $\widetilde{r}_x$ du signal $\mathtt{x}$ et en se référant à la relation 2.29, l'autocorrélation du signal $\mathtt{y}$ est donnée par :

$$\widetilde{r}_{\mathtt{y}} = \frac{\sqrt{M}}{N} F^H(\widehat{\mathtt{y}} \odot \widehat{\mathtt{y}}^*), \tag{2.30}$$

où $\widehat{\mathtt{y}} = F^H Z\mathtt{y}$. Cela signifie que le Produit de Hadamard $\widehat{\mathtt{y}} \odot \widehat{\mathtt{y}}^*$ dans l'équation 2.30 représente le carré de l'enveloppe de Hilbert du signal $\mathtt{x}$ WSHS-symétrisé. Ainsi, de même que l'enveloppe spectrale de $\mathtt{x}$ peut être approchée, selon 2.29, par un modèle AR dont l'autocorrélation r_x fournit les coefficients, l'enveloppe temporelle de $\mathtt{x}$ peut être approchée par un modèle AR de la DCT-Io de $\mathtt{x}$.

2.2.3 Structure du système de modélisation de l'enveloppe temporelle

Le principe de la méthode d'estimation de l'enveloppe temporelle est illustré par la figure 2.3. Le système est constitué de trois étapes successives :

1. La trame d'analyse est tout d'abord transformée dans le domaine fréquentiel par une Transformée en Cosinus Discrète (TCD).
2. Une prédiction linéaire est effectuée sur la représentation TCD pour obtenir une modèlisation paramétrique de l'enveloppe temporelle. Dans le contexte général, le modèle AR est utilisé, tel le cas de la modélisation de l'enveloppe temporelle utilisé dans la technique TNS [Herre 1999]. Dans le cadre de notre recherche, le modèle ARMA a été préféré. En effet, ce dernier permet d'assurer une meilleure représentation de l'enveloppe temporelle avec un nombre réduit de coefficients. L'ordre de prédiction dépend de la nature du signal ainsi que de la longueur de la trame d'analyse. Plus l'ordre est élevé, meilleure est l'approximation mais plus la quantité d'information à transmettre est importante. L'ordre de prédiction sera discuté ultérieurement.
3. L'enveloppe temporelle est approchée par la réponse "fréquentielle" du filtre défini par les coefficients ARMA. Elle est donnée par :

   $$\hat{e}(t) = |H(e^{\mathrm{j}t})|, \tag{2.31}$$

   où H est la fonction de transfert en z définie par :

   $$H(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{H_b(z)}{H_a(z)}. \tag{2.32}$$

   Une estimation des coefficients optimaux du modèle ARMA est réalisée par la méthode Prony détaillée en annexe A.

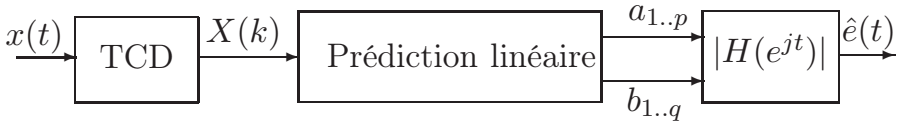


FIGURE 2.3 – Estimation de l'enveloppe temporelle par FDLP.

2.3 Codage de l'enveloppe temporelle

2.3.1 Représentation des paramètres du prédicteur

Dans le système proposé, seuls les paramètres ARMA modélisant l'enveloppe temporelle du signal d'origine doivent être codés et transmis au décodeur à travers un canal auxiliaire. Les coefficients AR et MA sont caractérisés par une grande dynamique qui ne facilite pas le codage. De plus, une faible erreur de quantification peut conduire à de fortes variations dans l'enveloppe temporelle reconstituée et génère souvent des problèmes d'instabilité du filtre de synthèse.

Les paramètres LSF (Line Spectral Frequencies), appelés encore paramètres LSP (Line Spectral Pairs), sont les plus appropriés à la quantification, puisqu'ils sont répartis dans l'intervalle $]0, \pi[$ et permettent un meilleur contrôle de la stabilité. Notons que la conversion est réversible et sans perte. Elle est fournie en annexe A.

La conversion en paramètres LSF est généralement appliquée aux paramètres (AR), correspondant à la fonction de transfert :

$$H(z) = (1 + \sum_{k=1}^{p} a_k z^{-k})^{-1} = \frac{1}{A(z)}. \tag{2.33}$$

Les coefficients MA sont transformés en coefficients LSF de la même manière que les paramètres AR. Toutefois, avant l'étape de la conversion, la fonction de transfert correspondant à la partie MA, définie par $H_b(z)$ de l'équation 2.32, est normalisée par b_0. On obtient alors :

$$H_b(z) = 1 + \sum_{i=1}^{q} b'_i z^{-i}, \tag{2.34}$$

où $b'_i = \frac{b_i}{b_0}$, $i = 1, .., q$. La valeur de b_0 détermine un gain global qui change peu avec le codage/décodage à l'inverse de la forme précise de l'enveloppe temporelle caractérisée par les autres coefficients AR. Pour cette raison, une estimation de la valeur de b_0 peut être déduite directement du signal codé/décodé. Afin d'illustrer la ressemblance entre ces deux paramètres (b_0 calculé sur le signal original et b_0 estimé sur le signal décodé), nous présentons dans le figure 2.4 une comparaison entre ces deux valeurs calculées sur des trames de 2048 échantillons. Compte-tenu de la similarité des b_0, les enveloppes temporelles calculées respectivement avec b_0 original et avec b_0 estimé sont très proches,

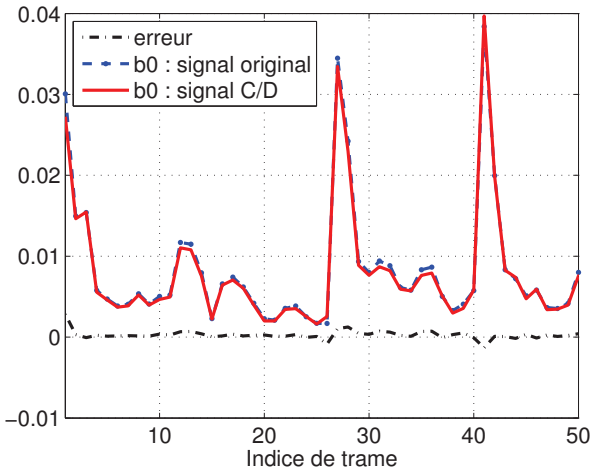


FIGURE 2.4 – Comparaison de b_0 et b_0 estimé calculés sur des trames de 2048 échantillons respectivement du signal original et de sa version codée/décodée par le codeur MP3 à 56 kbits/s .

comme illustré à la figure 2.5 pour un signal de castagnettes.

Condition de conversion

La conversion ARMA en LSF exige que toutes les racines de $H_a(z)$ (respectivement $H_b(z)$) se trouvent à l'intérieur du cercle unité. Cette propriété est garantie pour $H_a(z)$ seulement. Cependant, pour $H_b(z)$, selon la théorie de factorisation spectrale [Hayes 1996], cette contrainte peut être assurée par le remplacement de tous les pôles z situés à l'extérieur du cercle unité par leurs complexe réciproque $1/z$, ce qui ne change pas le module du filtre, et donc la forme de l'enveloppe codée.

2.3.2 Quantification et codage des paramètres du prédicteur

La technique de quantification utilisée tout au long de cette thèse est la quantification vectorielle (QV). Dans ce cas, les coefficients LSF, représentant les pôles et les zéros du modèle ARMA d'une trame, sont regroupés en un vecteur de dimension $p + q$.

Chaque vecteur est associé au vecteur le plus proche dans un dictionnaire de LSF à K entrées représentables sur N bits chacune. La génération de ces tables est réalisée par un algorithme classique LBG (Linde-Buzo-Gray) [Linde 1980] de minimisation d'erreur sur une base d'apprentissage (voir annexe B).

Les descripteurs de l'enveloppe ainsi quantifiés sont ensuite transmis, à un débit avoi-

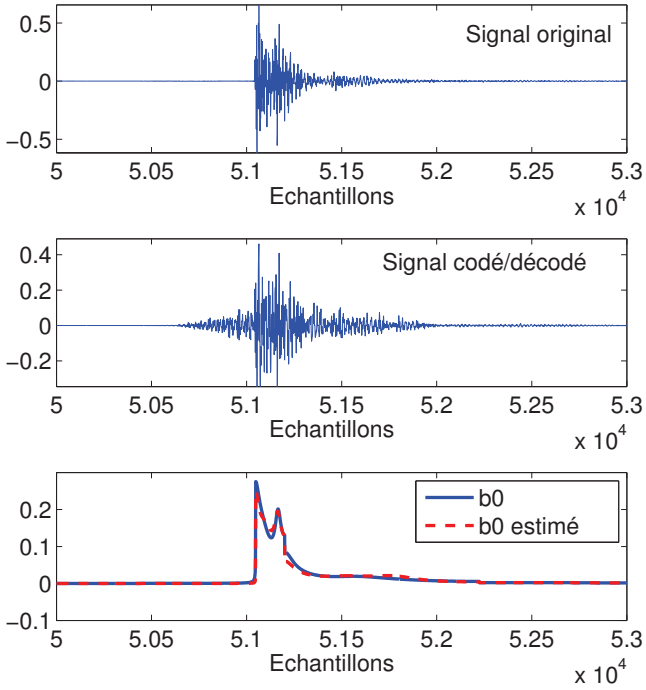


FIGURE 2.5 – Enveloppes temporelles estimées par b_0 calculé sur le signal original et b_0 calculé sur le signal codé/décodé.

sinant 500 bit/s, dans un flux distinct de celui du codeur. Ce débit nous est imposé par la perspective de l'utilisation du tatouage audio comme canal auxiliaire (voir chapitre 5).

2.3.3 Discussion de l'ordre de prédiction

L'ordre de prédiction à appliquer dépend de la nature de la trame d'analyse. En effet, pour les signaux localement stationnaires, les variations moyennes de l'énergie du signal sont assez lentes et il est assez aisé d'en modéliser l'enveloppe : un ordre relativement faible modélisera efficacement l'envelope temporelle. En revanche, pour des signaux présentant des transitions marquées, la modélisation de l'enveloppe devient plus difficile à cerner et necéssite un ordre de prédiction élevé.

Nous présentons sur la figure 2.6 différentes estimations de l'enveloppe temporelle d'un signal de violon échantillonné à 44.1 kHz. Trois ordres de prédiction ont été appliqués : un ARMA(2,3) représenté par l'enveloppe 1, un ARMA(5,3) représenté par l'enveloppe 2 et un ARMA(7,3) représenté par l'enveloppe 3. On remarque que pour la zone stationnaire, les trois enveloppes estimées sont comparables. En revanche, pour la zone non stationnaire, on note une différence entre les enveloppes estimées :

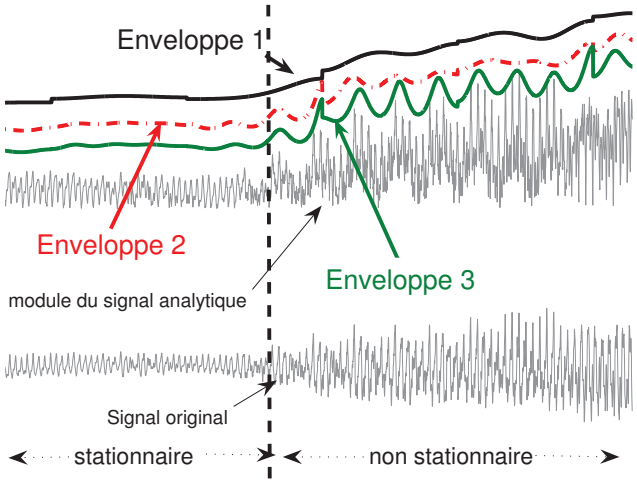


FIGURE 2.6 – Différentes estimations d'enveloppe temporelle d'un signal de violon calculé avec trois modèles ARMA d'ordres différents : (2,3), (5,3) et (7,3).

- L'enveloppe 1 modélise "globalement" l'énergie du signal. Cette enveloppe n'épousant pas suffisamment les variations dynamiques du signal audio.
- L'enveloppe 2 passe par les maxima du signal et modélise correctement les variations dynamiques du signal, en particulier le signal analytique. Pour des raison de débit, c'est vers ce modèle que nous tacherons de tendre dans la suite du document.
- L'enveloppe 3 présente une variation plus fine et plus précise, tend à accrocher efficacement les variations dynamiques du signal.

2.3.4 Problèmes liés à la modélisation FDLP dans le cas des signaux percussifs

Le signal représenté par la Figure 2.7 résulte de la superposition d'un signal de castagnette échantillonné à 44.1 kHz et l'enveloppe temporelle correspondante estimée avec un modèle ARMA d'ordre (7,3). Le signal de la castagnettes est non stationnaire et présente une discontinuité à l'instant de l'attaque (échantillon 485). En effet, à cet instant, une variation brutale de l'énergie du signal est notée.

La sélection de l'enveloppe temporelle à cet instant représentée par la Figure 2.7 met en évidence le problème de l'estimation de l'enveloppe temporelle sur ce type de signal. En effet, l'énergie varie fortement lors de l'attaque de la castagnettes alors que la variation de l'enveloppe temporelle associée est lente. Même avec un ordre de prédiction relativement élevé, la technique de modélisation de l'enveloppe temporelle exposée précédemment

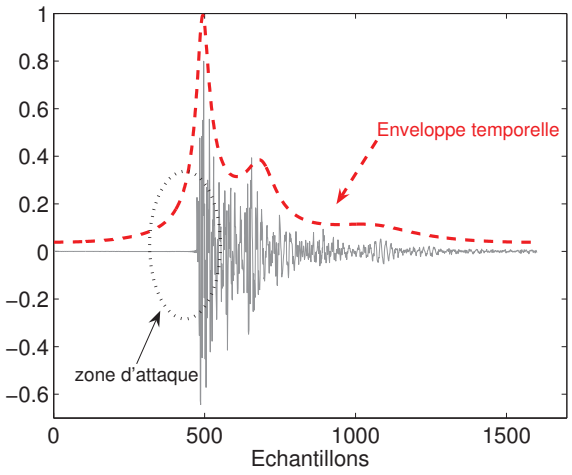


FIGURE 2.7 – Séquence audio de castagnettes échantillonnée à 44.1 kHz et enveloppe temporelle corresponadante estimée avec un modèle ARMA d'ordre (7,3).

devient dès lors totalement inefficace en présence d'une forte variation énergétique : la fidélité à l'original ne peut être assurée.

La solution adoptée consiste à subdiviser les trames d'analyse à caractère percussif en deux sous-trames selon la position de l'attaque. Une modélisation FDLP est par la suite appliquée sur chaque sous-trame. On développe dans ce qui suit la technique de détection des trames transitoires ainsi que la localisation de l'attaque.

2.4 Détection et localisation des transitions

Une attaque correspond à une brusque variation d'énergie de tout, ou d'une partie du spectre du signal. Elle correspond, la plupart du temps, à l'apparition d'un instrument ou à un changement de note au cours du temps. La détection d'attaque est requise dans le codage par transformée afin d'ajuster la taille des fenêtres d'analyse et de synthèse. L'amélioration de la résolution temporelle permet de limiter les phénomènes d'étalement de bruit et de pré-écho abordés dans le paragraphe 1.2.1 du chapitre 1.

Dans le cadre de notre étude, la détection des trames à attaque et la localisation de la position de la transition servent uniquement à définir la subdivision adéquate de la trame d'analyse en deux sous-trames. Dans ce contexte, nous envisageons deux hypothèses, à savoir :

– Hypothèse 1 : le récepteur est capable de localiser la position de l'attaque. Dans ce cas, le flux binaire à transmettre contiendra uniquement les descripteurs de l'enveloppe temporelle relatifs à chaque sous-trame comme illustré sur la Figure 2.8. Le récepteur doit également détecter les trames à attaque.

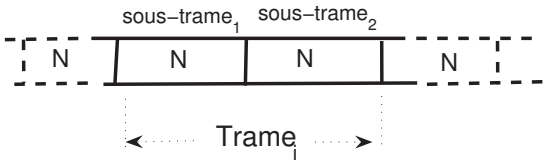


FIGURE 2.8 – Segmentation du flux binaire sans transmission de la position d'attaque.

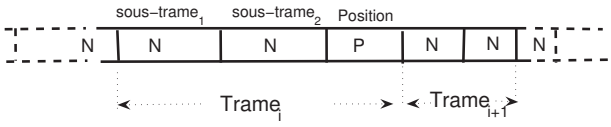


FIGURE 2.9 – Segmentation du flux binaire avec transmission de la position d'attaque.

– Hypothèse 2 : le récepteur n'est pas capable de localiser la position de l'attaque. Dans ce cas, deux segmentation du flux binaires se présentent :
  - en présence d'attaque, on transmet les enveloppes temporelles relatives aux deux sous-trames de part et d'autre de l'attaque ainsi que l'instant de la transition (voir Figure 2.9) ;
  - en absence d'attaque, on transmet uniquement les descripteurs de l'enveloppe temporelle relatifs à chaque sous-trame comme illustré sur la Figure 2.9.

  Le rôle du récepteur se limite à la détection des trames à attaque. En présence d'attaque, une subdivision de la trame en deux sous-trames selon la position de la transition fournie par le flux binaire est réalisée. Dans le cas contraire, une subdivision de la trame en deux sous-trames de même taille.

2.4.1 Détection des trames à attaque

La détection des trames à attaque dans un signal musical, c'est à dire la décision de classifier une trame en trame à attaque ou pas, peut passer par l'analyse trame à trame de l'énergie. Un changement soudain dans l'énergie ainsi qu'un changement soudain du contenu spectral peut s'apparenter à l'introduction d'une transitoire dans la trame. Sur ce principe se fonde la technique de détection des trames transitoires développée dans [3GPP 2004]. Cette technique est illustrée par la Figure 2.10, elle fait appel à plusieurs étapes d'analyse :

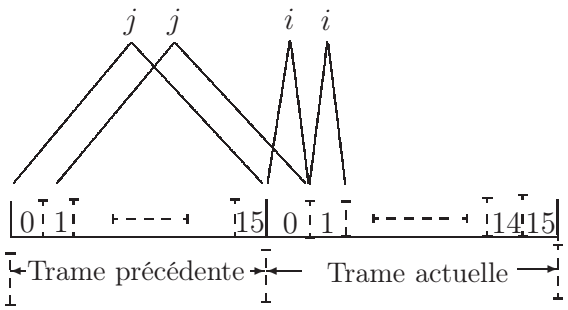


FIGURE 2.10 – Bloc de détection des trames transitoires

- Le signal audio est segmenté en trames d'analyse de 2048 échantillons, noté X_j.
- La trame d'analyse est filtrée par un filtre passe haut de fonction de transfert :

$$H(z) = \frac{0.7548(z-1)}{z-0.5095}, \tag{2.35}$$

  visant à éliminer la composante continue ;
- Les trames filtrées sont ensuite découpées en 16 sous-trames x_i de 128 échantillons. L'énergie de chaque sous-trame d'indice i définie par :

$$E_i(x) = \sum_{k=0}^{128} x_i(k)^2, \tag{2.36}$$

  est calculée.
- Pour chaque sous-trame i, on calcule par la suite le coefficient $attackR$ définit par :

$$attackR(i) = \frac{E_i(x)}{\text{moyenne}(E_j(x)|_{j=i-8:i-1})} \tag{2.37}$$

La valeur, $attackRatio = \max(attackR(i)|_{i=0,\ldots,15})$, est utilisée comme un indicateur de présence de transition. En effet, la probabilité d'avoir une attaque est d'autant plus élevée que la valeur d'*attackRatio* est élevée.

La détection d'attaque repose sur la comparaison d'*attackRatio* à un seuil *thr*. En se référant à la norme AAC [3GPP 2004], pour les signaux originaux, la valeur de *thr* est fixée à 10. Pour les signaux codés/décodés et en tenant compte de la présence du pré-écho en zone d'attaque, la valeur de *thr* est légèrement modifiée. En se fondant sur des mesures empiriques, la valeur du *thr* est fixée, pour les signaux codés/décodés comme suit :

$$thr = \begin{cases} 50 \text{ pour un débit } \leq 24 \text{ kbits/s} \\ 10 \text{ pour un débit } > 24 \text{ kbits/s} \end{cases} \tag{2.38}$$

On représente dans la Figure 2.11 le résultat de l'application de l'algorithme ci-dessus sur un signal audio (violon + castagnettes) échantilloné à 44.1 kHz. Les trames 1 et 3 sont considérées comme stationnaires par le détecteur d'attaque (coefficient d'attaque

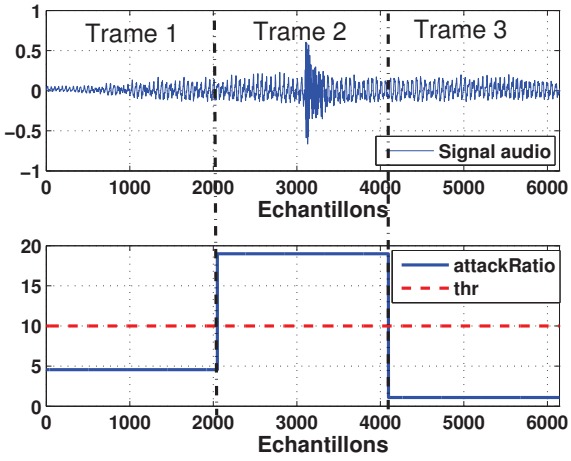


FIGURE 2.11 – Signal audio (violon+castagnettes) et coefficients d'*attackRatio* correspondants pour trois trames

ratio inférieur au seuil (10)). L'énergie de la trame 2 varie fortement, ce qui se traduit par un coefficient d'attaque ratio supérieur au seuil. Cette trame est alors considérée comme transitoire.

2.4.2 Méthodes pour la localisation d'attaque

2.4.2.1 Méthode fréquentielle : indice de stationnarité

La première méthode choisie pour la localisation de la position de l'attaque est fondée sur le calcul des indices de non stationnarité utilisé dans [Larbi 2005b, Laurent 1998]. Il s'agit d'une mesure de distance entre les Représentations Temps-Fréquence (RTF) du signal calculées à différents instants. Diverses mesures de cette distance sont proposées dans la littérature : la distance de Kolmogorov, la distance de Bhattacharyya et la divergence de Küllback [Laurent 1998, Sadok 2006], toutes inspirées de la théorie des probabilités.

La représentation temps-fréquence est calculée sur toute la durée du signal. A chaque instant n d'analyse, deux imagettes $I_1(n;\tau,f)$ et $I_2(n;\tau,f)$ de même durée L, sont extraites de la RTF globale, de part et d'autre de cet instant n avec :

$$\begin{aligned} I_1(n;\tau,f) &= RTF(n-L+\tau,f) \\ I_2(n;\tau,f) &= RTF(n+\tau,f), \end{aligned} \tag{2.39}$$

où L est la largeur des deux imagettes (en temps), τ est dans $[0,L]$ et f est la fréquence (voir figure 2.12). Nous avons opté pour une RTF utilisant le spectrogramme lissé par une fenêtre de Hamming, qui présente l'avantage d'être simple à utiliser et qui s'exprime

par :

$$S_x(t,f) = \left| \int_{-\infty}^{+\infty} x(u)h^*(u-t)e^{-j2\pi fu}du \right|^2, \tag{2.40}$$

où x est le signal analysé et h est la fenêtre de Hamming de longueur N_h.

Les modules des deux imagettes sont ensuite normalisées pour avoir une énergie unité selon :

$$NI_k(n,\tau,f) = \frac{|I_k(n,\tau,f)|}{\int_{\tau=0}^{L}\int_{-\infty}^{+\infty}|I_k(n,\tau,f)|dfd\tau}, k = 1,2 \tag{2.41}$$

Les nouvelles imagettes normalisées sont comparées entre elles en calculant une des distances suivantes :

– **La distance de Kolmogorov :**

$$SI_{ko}(n) = \int_{\tau=0}^{L}\int_{-\infty}^{+\infty}|NI_1(n;\tau,f) - NI_2(n;\tau,f)|dfd\tau \tag{2.42}$$

– **La divergence de Küllback :**

$$SI_{ku}(n) = \int_{\tau=0}^{L}\int_{-\infty}^{+\infty}(NI_1(n;\tau,f) - NI_2(n;\tau,f))\log\left(\frac{NI_1(n;\tau,f)}{NI_2(n;\tau,f)}\right)dfd\tau \tag{2.43}$$

– **La distance de Bhattacharyya :**

$$SI_{bh}(n) = -\log\left(\int_{\tau=0}^{L}\int_{-\infty}^{+\infty}\sqrt{NI_1(n;\tau,f).NI_2(n;\tau,f)}\right)dfd\tau \tag{2.44}$$

Dans les expériences de [Larbi 2005b], il a été démontré que la distance de Kolmogorov est plus sensible aux changements brusques des caractéristiques spectrales du signal en question. Ainsi, pour la localisation de la transition, le choix s'est porté sur la mesure de la distance de Kolmogorov.

Le comportement des indices de stationnarité s'interprète de la manière suivante : si les caractéristiques du signal étudié ne présentent pas de variations à l'instant n, l'indice

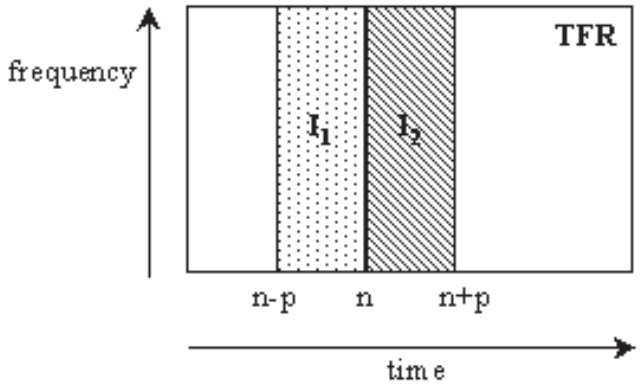


FIGURE 2.12 – Imagettes I_1 et I_2 [Larbi 2005b].

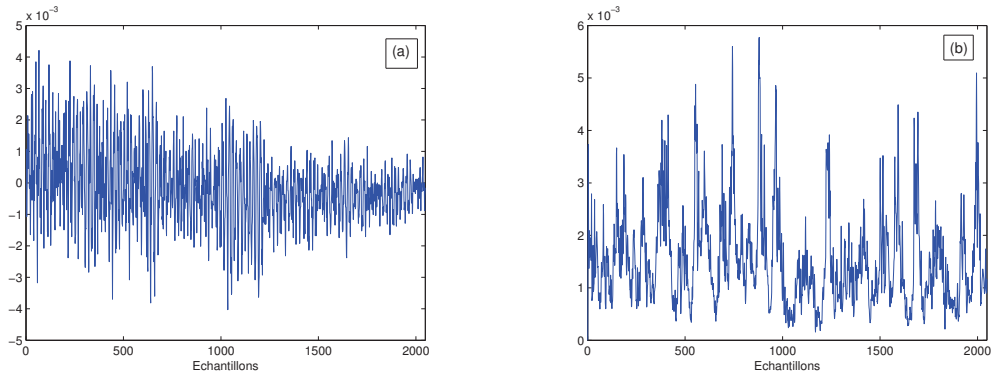


FIGURE 2.13 – (a) Variation temporelle d'une trame non transitoire, (b) indice de Kolmogorov correspondant.

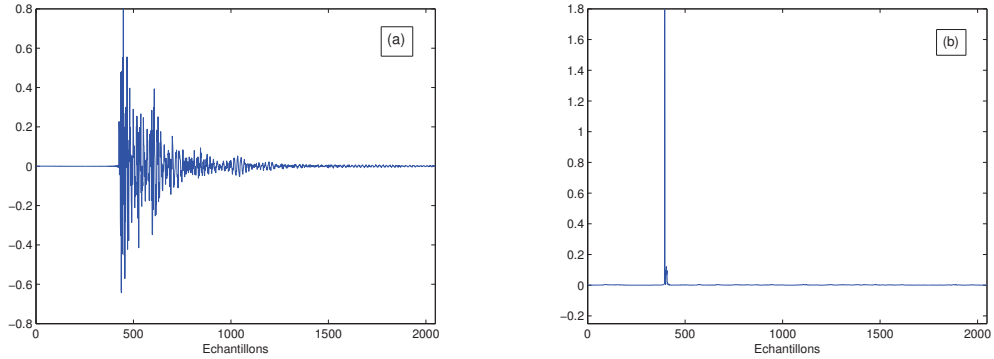


FIGURE 2.14 – (a) Variation temporelle d'une trame transitoire, (b) indice de Kolmogorov correspondant.

de stationnarité $SI(n)$ sera presque nul (voir figure 2.13). Dans le cas contraire, $SI(n)$ augmente sensiblement, indiquant la présence d'une variation plus au moins rapide du contenu fréquentiel du signal, autrement dit d'une transitoire (voir figure 2.14).

La Figure 2.15 superpose les indices de stationnarité d'un extrait de catagnette échantillonné à 44100 Hz et sa version codé/décodé avec le codeur MP3 à 48 kbps. Pour les zones de changement de puissance du signal audio, les pics de SI_{decode} de la version du signal codé/décodé sont nettement moins importants que ceux de $SI_{original}$ correspondant au signal original. Ainsi, les attaques risquent d'être difficiles à localiser dans le signal codé/décodé, de sorte qu'il sera nécessaire de transmettre au récepteur les positions des attaques pour permettre la correction de l'enveloppe (voir hypothèse 2 de l'introduction de la section 2.4).

Dans ce cas de figure, le module de réduction de pré-écho utilisé en entrée du décodeur prend en entrée :

– Le signal codé/décodé ;

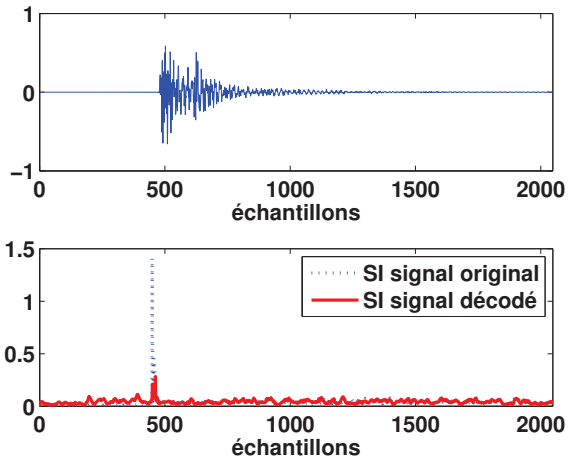


FIGURE 2.15 – Un extrait de musique (castagnettes), indice de Kolmogorov correspondant

- La position d'attaque pour les trames à attaque ;
- Les paramètres représentant les enveloppes temporelles des sous-trames, avant et après l'attaque, du signal original.

2.4.2.2 Méthode temporelle : détecteur algébrique

La méthode temporelle adoptée pour le calcul de la position d'attaque est fondée sur une modélisation algébrique proposée dans [Mboup 2008]. Cette dernière repose essentiellement sur une représentation polynomiale par morceaux du signal :

$$x(t) = \sum_{i=1}^{K} \chi_{[t_{i-1};t_i]} p_i(t - t_{i-1}) + n(t), \tag{2.45}$$

où

$$\begin{cases} \chi_{[t_{i-1};t_i]} & : \text{fonction caractéristique de l'intervalle } [t_{i-1}; t_i], \\ (p_i)_{i\in[1,K]} & : \text{une série de polynômes d'ordre } N, \\ n(t) & : \text{bruit additif.} \end{cases}$$

Soit T l'intervalle de temps tel que dans chaque intervalle $I_\tau^T = (\tau, \tau + T)$, au plus un changement se produit.

Soit $x_\tau(t) = x(t + \tau), \quad t \in [0, T]$, la restriction du signal dans l'intervalle I_τ^T. On définit le point de discontinuité, dit t_τ, relativement à l'intervalle I_τ^T avec :

- $t_\tau = 0$ si $x_\tau(t)$ est stationnaire
- $0 < t_\tau < T$ sinon

Soit $n(t) = 0$. Au sens de la théorie des distributions de L. Schwartz, la dérivée d'ordre

N du signal s'écrit :

$$\frac{\mathrm{d}^N}{\mathrm{d}t^N}x_\tau(t) = [x_\tau^{(N)}(t)] + \sum_{k=1}^{N} \mu_{N-k}\delta(t-t_\tau)^{k-1}, \tag{2.46}$$

où :

- $[x_\tau^{(N)}]$ représente la partie régulière de la dérivation d'ordre N du signal.
- μ_k est le saut de la dérivation d'ordre k au point t_τ.

$$\mu_k = x^{(k)}(t_{\tau^+}) - x^{(k)}(t_{\tau^-}) \tag{2.47}$$

Si :

$-\ \ \mu_0 = \mu_1 = ... = \mu_k = 0, \ \ 0 \leq t_\tau \leq T$, il n'y aucune transition dans l'intervalle donné.

$-\exists\ \ k\ \ /\ \ \mu_k \neq 0, \ \ 0 \leq t_\tau \leq T$, il y a une détection de transition dans l'intervalle donné au point t_τ.

Le problème de localisation des transitions est maintenant orienté vers la localisation des discontinuités du signal. Plusieurs estimateurs peuvent être tirés de l'équation 2.46.

En supposant que l'ordre du polynôme est inférieur à N, $x_\tau^{(N)} \equiv 0$, et par suite, équation 2.46 devient :

$$\frac{d^N}{dt^N}x_\tau(t) = \sum_{k=1}^{N} \mu_{N-k}\delta(t-t_\tau)^{k-1}. \tag{2.48}$$

Pour réduire la complexité de la résolution temporelle, l'équation 2.48 est transférée au domaine opérationnel en utilisant la transformée de Laplace :

$$s^N\widehat{x}_\tau(s) - \sum_{m=0}^{N-1} s^{N-m-1}\frac{d^m}{dt^m}x_\tau|_{t=0} = e^{-t_\tau s}(\mu_{N-1} + s\mu_{N-2} + ... + s^{N-1}\mu_0). \tag{2.49}$$

Compte tenu du fait que la condition initiale et les sauts de dérivation de $x_\tau(t)$ sont des paramètres inconnus, on annule les sauts $\mu_0, ..., \mu_{N-1}$ en appliquant N dérivations de l'équation 2.49 dans le domaine opérationnel.

Après quelques étapes de calcul, on obtient finalement :

$$\sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} t_\tau^{N-k}(s^N\widehat{x}_\tau(s))^{(N+k)} = 0, \tag{2.50}$$

où $\begin{pmatrix} N \\ k \end{pmatrix} = \frac{N!}{k!(N-k)!}$ désigne le coefficient binomial.

Rappelons que, par les règles classiques du calcul opérationnel, la multiplication par s^ν, $\nu > 0$, correspond à une dérivation d'ordre ν dans le domaine temporel. Une dérivée dans le domaine temporel est équivalent à filtrage passe-haut qui peut amplifier l'effet de

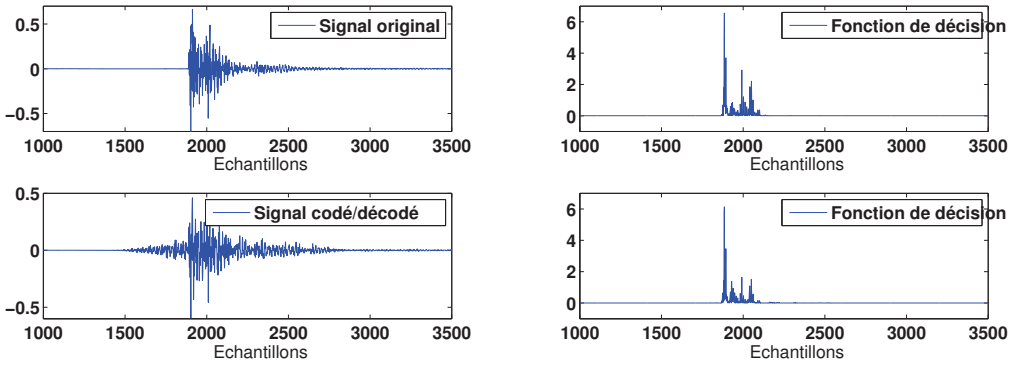


FIGURE 2.16 – Fonction de décision calculée par un détecteur algébrique d'ordre 2 pour un signal de castagnettes (en haut) et sa version codé/décodé par un MP3 à 48 kbits/s (en bas)

bruit. Afin d'éviter le filtrage passe-haut, toute l'équation 2.50 est divisée par s^ν, $\nu \geq N$, ce qui, dans le domaine temporel, sera équivalent à une intégration. Ainsi, on obtient :

$$\sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} t_\tau^{N-k} \frac{(s^N \widehat{x}_\tau(s))^{(N+k)}}{s^\nu} = 0. \tag{2.51}$$

Comme il n'y a plus de paramètres inconnus à gauche, l'équation 2.51 est transformée à nouveau au domaine temporel en utilisant la règle suivante :

$$L^{-1}(s^{-v}\hat{u}) = \frac{1}{(v-1)!} \int_0^t (t-\tau)^{v-1} u(\tau) d\tau.$$

La transformée de l'équation 2.51 dans le domaine temporel est donnée par :

$$\sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} t_\tau^{N-k} \varphi_{k+1}(t) = 0, \tag{2.52}$$

où

$$\varphi_{k+1}(t) = L^{-1}\left(\frac{(s^N \widehat{x}_\tau(s))^{(N+k)}}{s^\nu} \right) = \int_0^\infty h_{k+1}(\tau) x(t-\tau) d\tau \tag{2.53}$$

et

$$h_{k+1}(\tau) = \begin{cases} \left(\frac{\tau^{\nu-1}(T-\tau)^{(N+k)}}{(\nu-1)!} \right)^{(N)}, & 0 \leq \tau < T \\ 0, & \text{sinon} \end{cases} \tag{2.54}$$

Ce qui conduit à un calcul simple de la localisation de la transition t_τ. A titre d'exemple, pour $N = 2$, l'équation 2.52 se résume à une équation second degré à résoudre.

L'équation 2.52 est le cœur de l'approche algébrique [Mboup 2008]. En effet, un intervalle I_τ^T est libre de transition si et seulement si $\mu_1 = \mu_2 = ... = \mu_k = 0 \quad \forall\, t_\tau$, *i.e.* $\varphi_{k+1}(t) = 0 \quad \forall\, k$ d'après la transformation de 2.49 en 2.52. Ainsi, chacun des φ_{k+1} peut

être considéré comme une fonction de décision. Cependant, en présence de bruit, $n(t) \neq 0$ (non pris en compte dans les équations précédentes), certaines valeurs de φ_{k+1} peuvent être différentes de zéro, même en absence de transitoires. Par conséquent, une fonction de décision est donnée par :

$$J(t) = \prod_{i=0}^{N} \varphi_i(t). \tag{2.55}$$

La probabilité d'avoir une discontinuité est d'autant plus élevée que $J(t)$ est élevée. Ainsi, une discontinuité est détectée si $J(t)$ est supérieure à un seuil $\lambda > 0$, dépendant du niveau de bruit $n(t)$. Dans le cadre de nos expériences, pour chaque trame, on fixe λ comme suit :

$$\lambda = \frac{\max_{t\in[0;2047]} J(t)}{20}. \tag{2.56}$$

Pour illustrer les performances du détecteur algébrique et sa robustesse au phénomène du pré-écho, nous présentons dans la Figure 2.16 la fonction de décision pour un signal original et sa version codé/décodé en présence d'un pré-écho. La Figure 2.16 montre qu'un détecteur algébrique d'ordre 2 est largement suffisant pour localiser la transition. En outre, comme illustré dans la Figure 2.16, la méthode est robuste au bruit de codage/décodage. Les multiples intégrations effectuées par le détecteur algébrique ont réduit le bruit, ce qui conduit à une fonction de décision similaire pour les deux signaux : original et codé/décodé.

Le détecteur algébrique présente ainsi l'avantage de la localisation de la position de l'attaque directement sur le signal décodé. Ainsi, la capacité offerte par le canal auxiliare sera réservée uniquement à la transmission de l'enveloppe temporelle (cas de l'hypothèse 1).

2.4.3 Performances des deux détecteurs de transition

Pour comparer les performances des deux détecteurs, nous proposons d'évaluer la robustesse et la complexité de ces deux algorithmes.

• Robustesse à la compression MPEG

La figure 2.17 présente l'erreur d'estimation de la position de l'attaque en fonction des indices des attaques, en utilisant respectivement le détecteur fréquentiel (IS) et le détecteur algébrique (DA). La position réelle de l'attaque est mesurée directement sur le signal original. Ces tracés sont obtenus en utilisant deux séquences audio, castagnettes et triangle, échantillonnées à 44100 Hz, et leurs versions codées/décodées par le codeur MP3 à 56 kbits/s.

L'erreur de l'estimation de la postion d'attaque donnée par le détecteur fréquentiel est beaucoup plus importante que celle obtenue par le détecteur algébrique. On peut conclure alors que le détecteur algébrique est plus robuste à la compression MPEG.

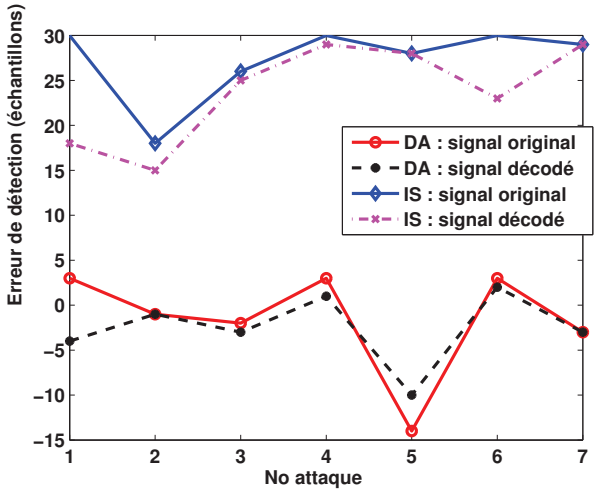


FIGURE 2.17 – Erreur de détection.

• **Complexité**

Bien que le cadre mathématique du détecteur algébrique paraisse complexe, la fonction de décision (2.53 et 2.55) est simple à calculer. Finalement, la complexité est plus faible que celle du détecteur fréquentiel. En effet, le DA nécessite $N + 1$ opérations de convolution, avec ici $N = 2$, sachant que le calcul des coefficients de chaque filtre h_{k+1} est réalisé une seule fois.

La complexité du calcul pour le détecteur fréquentiel correspond aux nombre d'opérations essentielles pour la transformation temps-fréquence. Pour chaque échantillon n, la complexité est de l'ordre de $N \log N$ où N est le nombre de point fft.

2.5 Système complet pour la réduction de pré-écho

Bien que cette technique soit principalement considérée comme étant un post-traitement au décodeur, des paramètres cruciaux, permettant la réduction du pré-echo après réception, sont extraits au niveau du codeur. L'utilisation du détecteur algébrique dans le système complet de réduction de pré-écho présente l'avantage de transmettre les informations relatives à l'enveloppe temporelle d'une manière plus précise pour un même débit binaire offert par le canal auxiliaire.

2.5.1 Traitement côté codeur

Le traitement côté encodage est représenté par la Figure 2.18. Le signal original est segmenté en trames d'analyse de 2048 échantillons. Une détection d'attaque est réalisée

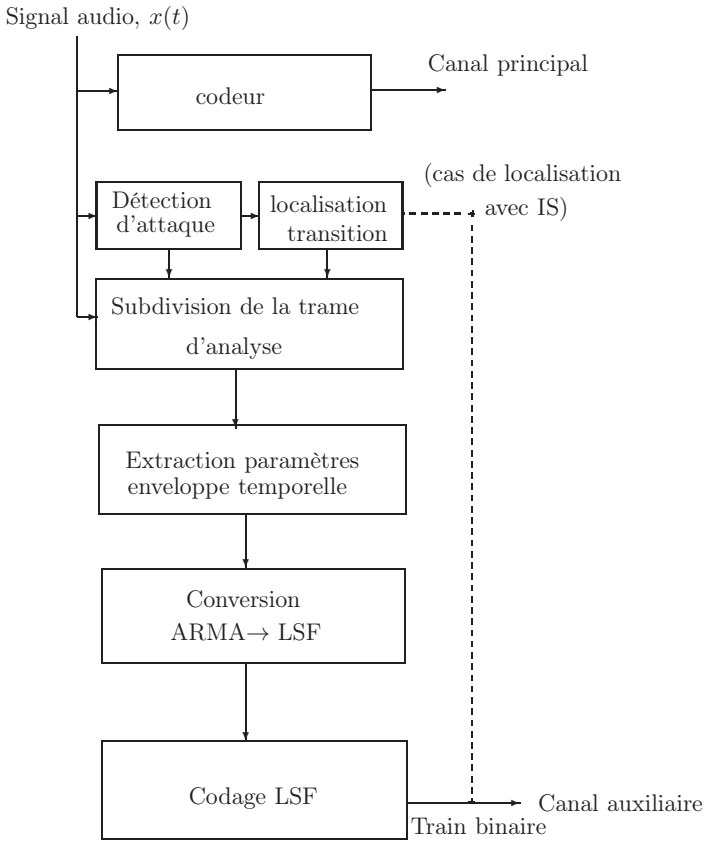


FIGURE 2.18 – Diagramme de fonctionnement du système côté codeur.

sur la trame d'analyse, permettant de déterminer la subdivision à utiliser pour bien suivre l'évolution temporelle du signal original :

- pour les trames stationnaires, la trame d'analyse est divisée en deux sous-trames de 1024 échantillons. Sur chaque sous-trame, on réalise une estimation de l'enveloppe temporelle ;
- pour les trames transitoires, la trame est subdivisée en deux sous-trames selon la position de la transition. Cette dernière est estimée par le module de localisation de transition. La transmission de l'instant de l'attaque est réservée uniquement dans le cas d'une localisation avec IS (trait pointillet de la Figure 2.18). Pour le cas du détecteur DA, l'instant de l'attaque est estimé directement sur le signal décodé. Une fois le découpage réalisé, il y a extraction des paramètres de l'enveloppe temporelle pour chacune des deux sous-trames.

Les coefficients modélisant l'enveloppe temporelle (ARMA) sont convertis en coefficients LSF. Ces descripteurs, ainsi que la position de l'attaque (uniquement dans le cas d'une localisation avec IS), sont ensuite quantifiés avant d'être transmis, à un débit avoisinant 500 bps, dans un flux distinct de celui du codeur. Ce débit nous est imposé par les contraintes liées à l'utilisation du tatouage audio comme canal auxiliaire dans la suite de

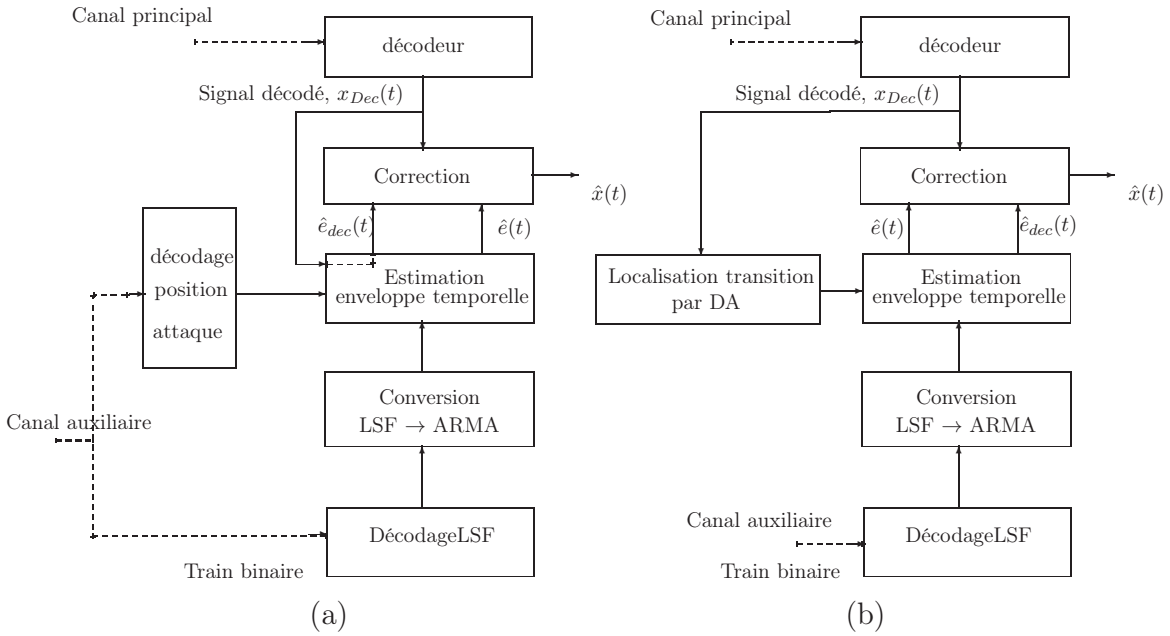


FIGURE 2.19 – Diagramme de fonctionnement du système côté décodeur : (a) cas IS (hypothèse 2), (b) cas DA (hypothèse 1).

ce travail (partie 2).

2.5.2 Traitement côté décodeur

Le train binaire auxiliaire et le signal issu du décodeur audio portent les informations requises pour corriger les attaques, à savoir :

- la position de la transition fournie soit dans le train binaire (cas de localisation avec IS , voir Figure 2.19 -(a)), soit par détection algébrique directement sur le signal codé/décodé (voir Figure 2.19 -(b)) ;
- les indices correspondant aux paramètres LSF quantifiés. Les coefficients ARMA déduits de ces paramètres permettent la remise en forme de l'enveloppe temporelle du signal décodé.

Une fois les enveloppes temporelles des signaux original et décodé estimées, le signal issu du décodeur subit une correction, fondée sur le rapport entre les enveloppes, comme suit :

$$signal_{corrige}(t) = signal_{decode}(t)\frac{\hat{e}(t)}{\hat{e}(t)_{dec}} \tag{2.57}$$

avec

- $\hat{e}(t)$: estimée de l'enveloppe temporelle du signal source obtenue par décodage du flux tatoué ;
- $\hat{e}(t)_{dec}$: estimée de l'enveloppe temporelle du signal décodé obtenue par modélisation FDLP.

2.6 Analyse et évaluation des performances du système proposé

2.6.1 Protocole expérimental

Le système de réduction de pré-écho et de mise en forme de l'attaque proposé a été testé sur des extraits de musique mono à caractère percussif, en particulier des séquences de castagnettes et de triangle échantillonnés à 44100 Hz. Les codeurs audio utilisés sont les implémentations des codeurs AAC et MP3 du projet Lame[2]. Ces versions offrent divers débits de compression allant de 8 à 64 kbits/s pour le AAC et de 24 à 96 kbits/s pour le MP3.

Les coefficients de prédiction ARMA décrivant l'enveloppe temporelle sont convertis en coefficients LSF, conformément à la technique fournie en annexe C, puis quantifiés (quantification vectorielle décrite au paragraphe 2.3.2). L'ordre du modèle de prédiction adopté est fixé à : ARMA(2,3) et ARMA(5,3). Rappelons que l'ordre de prédiction est limité par le débit du canal auxiliaire.

- Pour un ordre du modèle égal à (2,3), le choix s'est porté sur une location de l'instant de transition par les IS et une transmission de la position. Nous adoptons une quantification vectorielle des LSF de 256 vecteurs en entrés (quantification des LSF sur 8 bits) et une quantification de la position d'attaque sur 4 bits. Le dictionnaire est généré par en utilisant une grande base de données de signaux de caractéristiques différentes (harmoniques et percussifs). Ainsi, le débit du canal auxiliaire varie entre 344 et 431 bps.
- Pour un ordre de modèle égal à (5,3), le choix s'est porté sur une détection d'attaque par un DA au niveau du décodeur (pas de transmission de l'instant de l'attaque) et une quantification vectorielle des LSF avec 1024 vecteurs en entrée (quantification des LSF sur 10 bits). Le debit du canal auxiliaire est fixé à 431 bps, réservé uniquement à la transmission des paramètres de l'enveloppe temporelle du signal original.

Le système proposé est évalué aussi bien pour le codage simple et pour le codage multiples. En effet, d'après [Reiss 2004], la distorsion de la qualité de compression s'accentue dans le cadre du codage multiple : le pré-écho s'accumule pour chaque cycle de codage/décodage. Dans ce cadre, le système de réduction de pré-écho est appliqué après chaque cycle de codage/décodage.

Le détecteur d'attaque intégré dans le système de réduction de pré-écho proposé utilise les deux techniques présentées ci-dessus à savoir les indices de stationnarités et le détecteur algébrique. La comparaison des deux techniques est aussi envisagée dans la partie

2. LAME version 3.98 (http ://www.mp3dev.org/).

ODG	Qualification de la dégradation
0	Imperceptible
-1	Perceptible mais non gênante
-2	Légèrement gênante
-3	Gênante
-4	Très gênante

TABLE 2.1 – Echelle de dégradation à cinq notes et valeurs de l'ODG associée.

évaluation. La qualité perçue des signaux après correction est évaluée par des mesures objectives de qualité.

2.6.2 Critère de mesure objective

Les mesures subjectives mettant en œuvre des tests d'écoute pour l'évaluation de la qualité de restauration des signaux audio sont très coûteux en terme de temps et d'argent. Pour une évaluation plus rapide et moins coûteuse, différentes mesures objectives intègrant les particularités du système auditif humain ont été proposées dans la littérature et normalisées [1387 1998]. Les mesures objectives fournies par le logiciel PEMO-Q (Perception Model of Quality assessment)[Huber 2006] permettent d'obtenir une mesure relativement fiable, ce qui nous a conduit à le choisir comme critère d'évaluation objectif.

PEMO-Q fournit un score appelé ODG (Objective Difference Grade) approchant la dégradation perceptive SDG [3] (Subjective difference Garde) entre le signal original x et sa version traitée y.

Les valeurs de ODG varient sur une échelle d'évaluation variant dans l'intervalle [-4,0], présentée dans la table 2.1. La valeur 0 correspond à une dégradation imperceptible et la valeur -4 correspond à une dégradation très gênante.

Comparé à l'algorithme PEAQ [4] proposé par la recommandation ITU-R B.S. 1387 [1387 1998], il a été démontré dans [Huber 2006] que les qualités audio prédites par PEMO-Q présentent une bonne corrélation avec les mesures subjectives SGD. De plus, l'algorithme adopte un modèle cognitif beaucoup plus simple que celui utilisé par PEAQ et offre des performances meilleures que celui-ci.

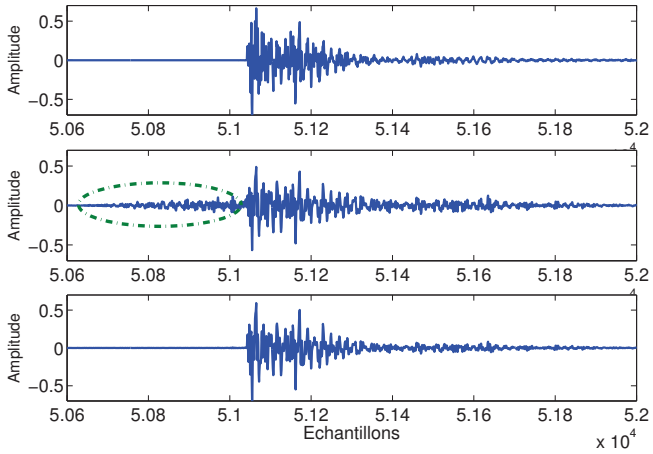


FIGURE 2.20 – Signal original (en haut), signal codé avec le MP3 à 56 kbps (au milieu) et signal décodé reconstitué avec la méthode proposée (en bas).

2.6.3 Illustration de la réduction de pré-écho sur un signal audio transitoire : "castagnettes"

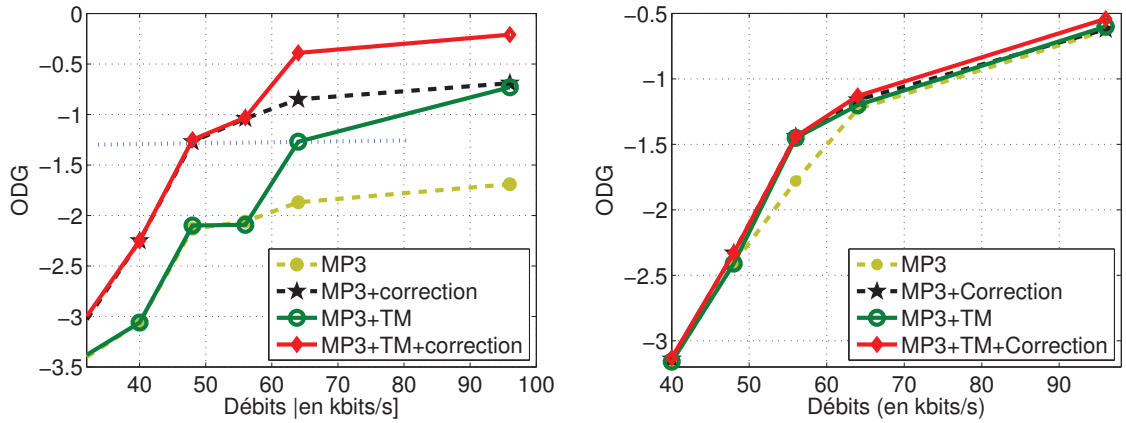
Dans ce qui suit, la restauration du signal par enveloppe temporelle est comparée à la méthode *Temporal Masking (TM)* utilisée dans le codeur MP3. Le signal original, un extrait de castagnettes échantilloné à 44100 Hz, est présenté en figure 2.20 -(a). Sur la figure 2.20-(b), on relève les défauts du codeur MP3 à 56 kbps qui se matérialisent par deux aspects : la présence de pré-écho et la mauvaise restitution de la dynamique du signal. Enfin, le résultat de la restauration d'attaque est présenté par la figure 2.20-(c). On peut aisement remarquer la réduction du pré-écho grâce à la solution proposée. En outre, on peut observer aussi qu'une meilleure restitution de la dynamique du signal est obtenue.

2.6.4 Evaluation objective : contexte d'un simple codage/décodage

- ***Cas de la détection d'attaque par IS***

Dans un premier temps, le système de réduction de pré-écho considéré a été testé en intégrant la méthode fréquentielle (IS) pour la localisation d'attaque. Sur les figures 2.21 et 2.22, on compare l'évolution des ODGs en fonction du débit de codage, pour un signal de castagnettes et un signal de triangle échantillonnés à 44100 Hz et codés/décodés

3. Test subjectif proposé par la recommandation UIT-R BS. 562.
4. Perceptual Evaluation of Audio Quality.



(a) (b)

FIGURE 2.21 – Evolution de l'ODG en fonction du débit de compression pour le système proposé, le codeur MP3+TM et le codeur MP3 seul : (a) castagnettes, (b) triangle.

respectivement par les codeurs MP3 et AAC. Les résultats de simulations indiquent que :

- la technique TM de réduction de pré-écho associé au codeur MP3 offre une qualité légèrement meilleure que celle du MP3 seul uniquement pour des débits > 50 kbits/s. L'approche proposée associée au codeur MP3+TM permet d'améliorer la qualité de la compression, en particulier pour le signal de la castagnettes. En effet, pour ce signal, le système proposé offre une qualité transparente pour un débit de 56 kbits/s. On note aussi que la correction proposée à 48 kbits/s offre la même qualité qu'à 64 kbits/s obtenue par le MP3+TM seul, offre ainsi un gain de 24 kbits/s.
- Associé au codeur AAC+TNS, pour des débits de compression variant de 32 à 64 kbits/s, la solution proposée offre une qualité légèrement meilleure que celle de l'AAC+TNS seul.

- ***Cas de la détection d'attaque par DA***

Maintenant, on se place dans une situation où le système proposé utilise la méthode algébrique pour la localisation d'attaque. Rappelons que dans ce cas, la position de l'attaque sera détectée directement sur le signal codé/décodé. Le codeur transmet alors uniquement les paramètres modélisant l'enveloppe temporelle.

Le débit du canal auxiliaire est ici réservé uniquement pour la transmission de l'enveloppe temporelle. Ce gain en terme de débit laisse entrevoir un système de réduction de pré-écho plus performant. Le protocole expérimental adopté consiste en la comparaison

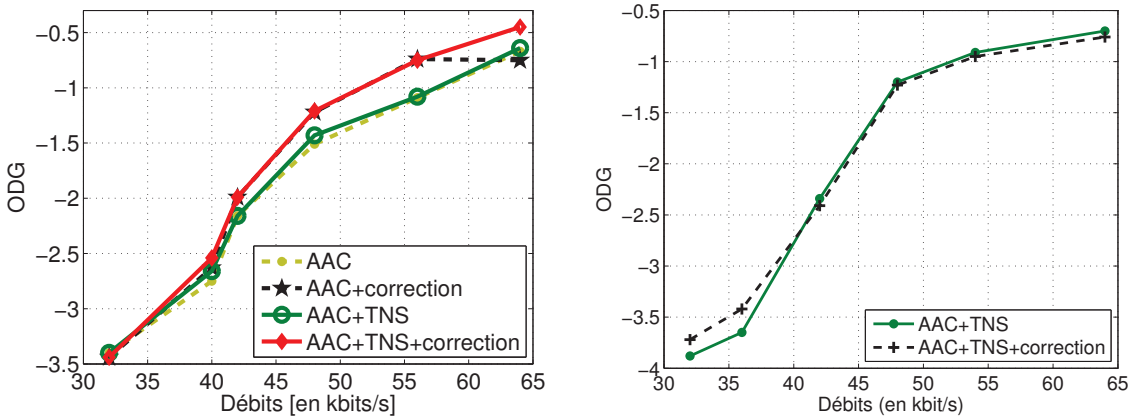


FIGURE 2.22 – Evolution de l'ODG en fonction du débit de compression pour le système proposé associé au codeur AAC+TNS et par le codeur AAC+TNS seul : (a) castagnettes, (b) triangle.

de la performance du système proposé en utilisant le détecteur fréquentiel (IS) avec un modèle ARMA d'ordre (2,3) codé sur 8 bits et le détecteur algébrique (DA) avec un modèle ARMA d'ordre (5,3) codé sur 10 bits (la non transmission de la position des attaques dans le cas du DA permet une meilleure modélisation de l'enveloppe temporelle pour le même débit de canal auxiliaire). On compare sur la figure 2.23 l'évolution de l'ODG en fonction du débit variant de 32 à 96 kbits/s (codeur MP3). Pour le même débit d'insertion, ces tracés montrent que les performances du système utilisant DA sont meilleures que celle avec le IS. Pour certains débits de compression, un gain d'une note de score ODG est atteint. Un gain en terme de débit sera par conséquent obtenu : à 64 kbits/s, on obtient la même qualité audio que le MP3+TM à 96 kbits/s.

2.6.5 Evaluation objective : contexte d'un codage multiple

Nous présentons sur la figure 2.24 le schéma du système de correction pour le cas du codage multiple, appelé aussi "codage en tandem". Cette structure comporte trois blocs :

- extraction des paramètres décrivant l'enveloppe temporelle du signal original $x(t)$. Ces paramètres serviront pour chaque cycle de codage/décodage ;
- C/D : codage et décodage ;
- correction de l'enveloppe temporelle. Nous adoptons ici le détecteur d'attaque algébrique et une modélisation ARMA(5,3) de l'enveloppe temporelle.

Afin de juger les performances du système proposé dans le cadre du multi-codage, l'algorithme Pemo-Q est utilisé au récepteur après chaque cycle de codage/décodage. Cet algorithme permet de mesurer la dégradation avant et après correction du signal audio.

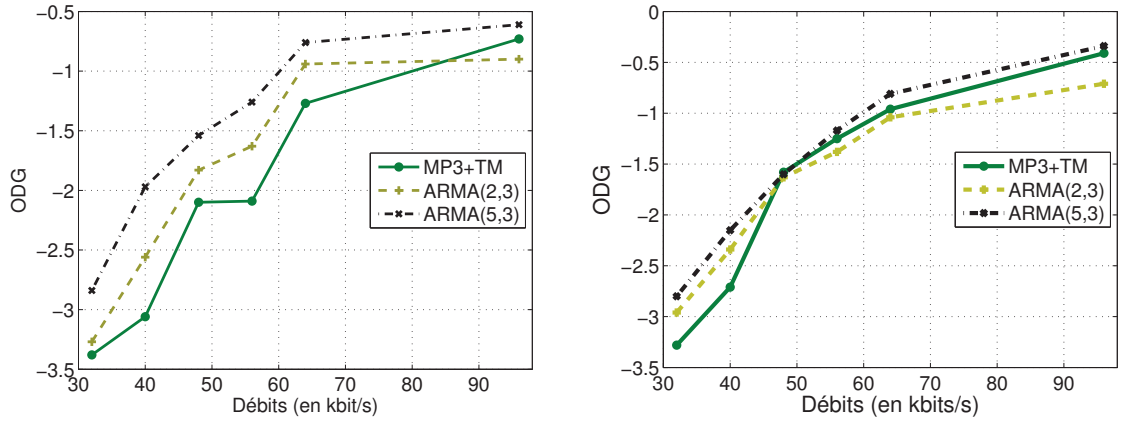


FIGURE 2.23 – Evolution de l'ODG en fonction du débit de compression MP3 pour différents ordres du modèle ARMA : (a) castagnettes, (b) triangle.

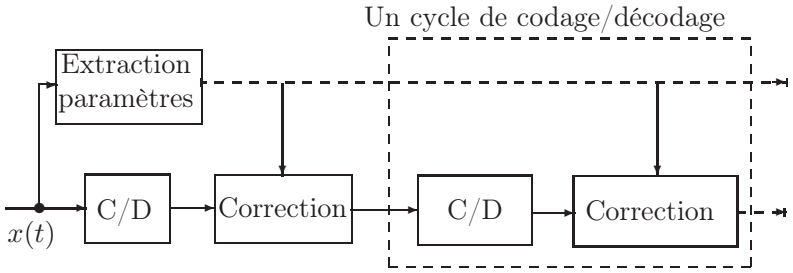


FIGURE 2.24 – Diagramme de fonctionnement du système de correction dans le contexte du codage multiple.

La figure 2.25 illustre les performances de l'approche proposée dans le contexte du multi-codage/décodage en utilisant le codeur MP3 avec les même signaux castagnettes et triangle que précédement. Sans correction, la qualité des signaux audio se dégrade davantage après chaque cycle de codage/décodage. L'option TM, intégrée dans le codeur MP3, apporte une légère amélioration de la qualité des signaux codés/décodés. Cependant, on remarque une nette amélioration de la qualité sur les signaux restaurés par l'approche proposée. En particulier, pour le signal de la castagnettes, la qualité reste presque transparente (ODG ≥ -1).

2.7 Conclusion

Dans ce chapitre, nous avons proposé un système de réduction de pré-écho adapté aux codeurs audio MPEG à bas débits. La solution consiste à corriger l'enveloppe temporelle du signal décodé en exploitant la connaissance *a priori* de l'enveloppe temporelle du signal

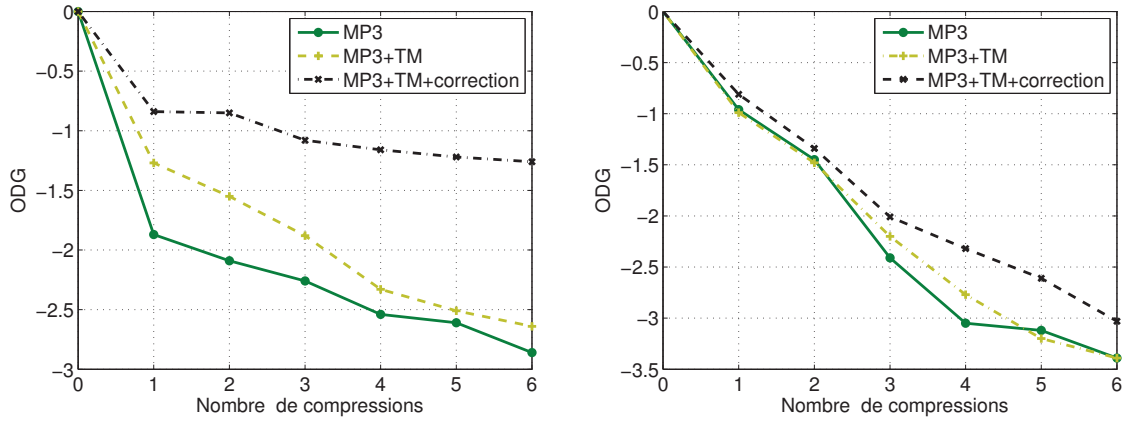


FIGURE 2.25 – Evolution de l'ODG en fonction du nombre de compressions MP3 : (a) castagnettes, (b) triangle.

original. Cette dernière est modélisée par la technique FDLP (Frequency Domain Linear Prediction) permettant de bien suivre l'évolution temporelle du signal tout en réduisant le nombre de paramètres qui la décrivent.

Le système proposé a été évalué par des critères objectifs dédiés à la mesure de la qualité audio perçue. Les résultats obtenus montrent une amélioration de la qualité des signaux audio restaurés. Ceci a été noté en particulier pour le cas du codeur MP3+TM et MP3 seul dans le contexte du codage simple et multiple. Associé au codeur AAC+TNS et AAC seul, les résultats obtenus sont légèrement meilleurs que ceux obtenus avec le AAC seul.

Chapitre 3

Restauration d'harmonicité/tonalité par translations spectrales

Sommaire

3.1 Introduction

La tâche du module d'extension de bande, appelé aussi SBR, est la synthèse de la bande haute fréquence non transmise par le codeur cœur. Principalement, deux opérations sont effectuées : la regénération de la structure fine du spectre haute fréquence par duplications spectrales et la mise en forme spectrale de la bande synthétisée.

Comme nous avons vu au chapitre 1, les positions des composantes tonales ne sont pas respectées lors de la regénération des hautes fréquences à partir du spectre basse fréquence. Il en découle dans la majorité des cas un phénomène de rugosité (cas des signaux à caractère tonal) et rupture d'harmonicité (pour les signaux harmoniques).

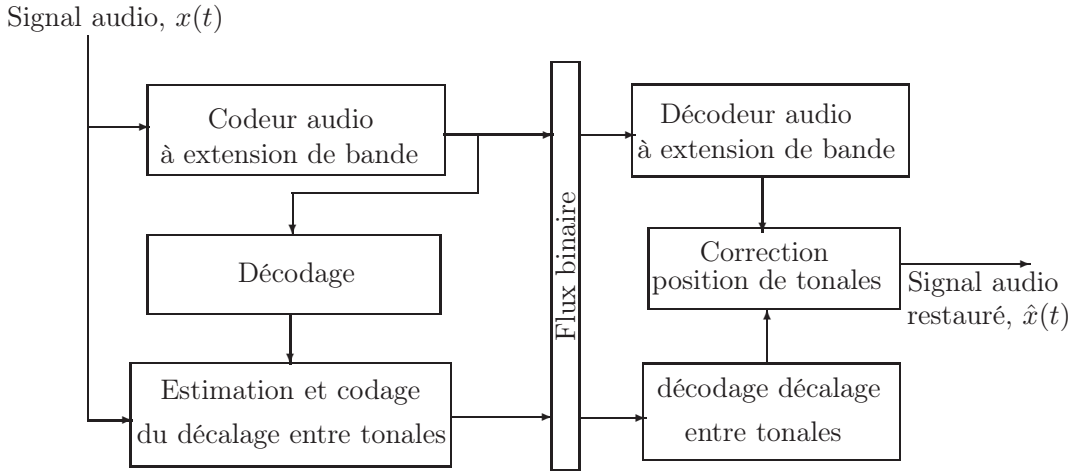


FIGURE 3.1 – Schéma général du système de correction de tonales proposé.

Débit (kbits/s)	fréquence d'échantillonnage (kHz)	Largeur de bande (kHz)
8	16, 22.05, 24, 32	7.5, 7.9, 8.2, 8
10	16, 22.05, 24, 32	8.0, 10.5, 10.9, 11
12	16, 22.05, 24, 32	8, 11, 11.4, 11
16	22.05, 24, 32, 44.1, 48	11.04, 11.4, 12.3, 11.7, 12
20	32, 44.1, 48	14.5, 14.8, 15
24	32, 44.1, 48	15.3, 14.8, 15.4

TABLE 3.1 – Configuration du HE-AAC en mono pour les débits de 8 à 24 kbits/s [Technologies 2007].

Nous proposons dans ce chapitre une technique de correction des positions de tonales, dédiée principalement aux signaux tonals et/ou harmoniques, par des translations spectrales. La technique proposée repose sur l'architecture fournie par la figure 3.1. Le signal original à large bande est codé par le codeur à extension de bande à bas débit, en particulier le codeur HE-AAC, appelé aussi aacPlus [1]. Notons que le codeur HE-AAC offre une grande souplesse d'utilisation concernant la fréquences d'échantillonnage. En effet, il fonctionne avec des signaux audio d'entrée de fréquences d'échantillonnage variées (16, 32, 44.1 et 48 kHz) et délivre également en sortie des signaux audio de fréquences d'échantillonnage variées en fonction de la fréquence d'échantillonnage en entrée et du débit considéré (voir tableau 3.1). Afin de simplifier le traitement, nous travaillons avec des signaux originaux échantillonnés à 44.1 kHz, et des signaux codés/décodés échantillonnés à 32 kHz. Les débits de compression considérés sont 16 et 20 kbits/s.

1. http ://aacplus-evaluation-package.software.informer.com/8.0/download/

Débit (kbits/s)	fréquence d'échantillonnage (kHz)	Largeur de bande (kHz)
8	8, 11.025, 12	3.8, 3.8, 3.8
10	8, 11.025, 12	3.8, 3.8, 3.8
12	8, 11.025, 12, 16	4.0, 5.2, 5.2, 5.2
16	8 à 24	4.0 à 5.2
20	11.025 à 24	5.5 à 7.2
24	11.025 à 32	5.5 à 10

TABLE 3.2 – Configuration du codeur cœur AAC en mono pour les débits de 8 à 24 kbits/s.

Dans le but de réduire au maximum le débit de transmission des données relatives à la position des tonales, nous proposons de transmettre les décalages Δf entre les positions des tonales détectées sur le signal original et celles détectées sur le signal codé/décodé dans la bande de fréquence [4 - 11.7 kHz]. Ce choix dépend de la largeur de bande synthétisée par le codeur SBR (voir tableau 3.1). Pour cela, nous réalisons un décodage au niveau du codeur.

Le décodeur audio à extension de bande synthétise le signal pleine bande. Les positions de tonales synthétisées par le décodeur SBR sont par la suite corrigées par des opérations de translation spectrale en utilisant le décalage de position transmis et décodé. La translation fréquentielle des tonales étant réalisée après la correction d'enveloppe spectrale du décodeur SBR, il faudrait en toute rigueur également corriger l'amplitude des tonales translatées. Cependant, les déplacements en fréquence étant faibles et l'enveloppe spectrale présentant des variations lisses, l'erreur commise sur l'amplitude est a priori négligeable. Ce problème pourrait également être évité en corrigeant la position des tonales avant la correction d'enveloppe spectrale du décodeur SBR, mais cette solution ne correspond pas à notre objectif d'un traitement post-décodeur.

Comme illustré sur la figure 3.1, les systèmes de traitement se présentent aux deux extrémités de la chaîne de transmission : côté codage et côté décodage. Nous développons dans la section 3.2 la technique de détection et de transmission des positions des tonales. Dans la section 3.3, nous présentons la méthode de translation spectrale retenue qui vise à corriger les positions de tonales mal synthétisées.

3.2 Traitement côté codeur

Comme annoncé au paragraphe 3.1, le traitement principal au niveau du codeur consiste à identifier les composantes tonales sur le signal original ainsi que sur le signal décodé, à calculer le décalage entre les positions à transmettre au décodeur dans un

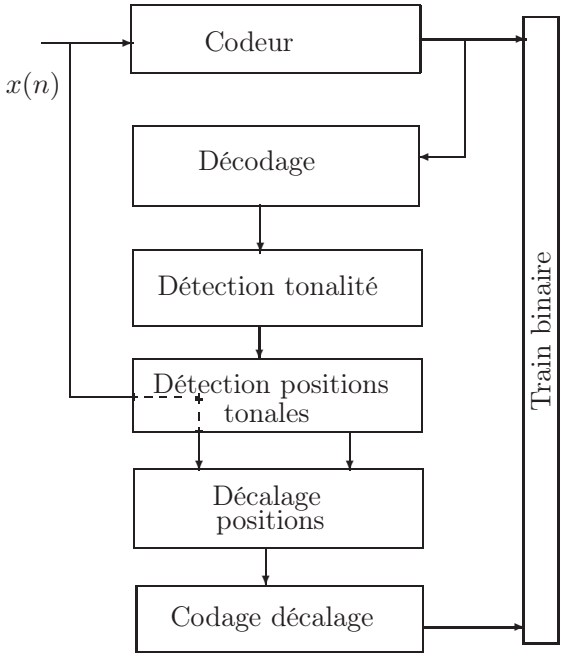


FIGURE 3.2 – Diagramme de fonctionnement côté codeur.

flux distinct de celui du codeur audio. Ce principe est illustré sur la figure 3.2.

Les traitements décrits dans ce paragraphe requièrent des modules auxiliaires d'analyse des trames. Nous développons dans ce qui suit la technique de détection de tonalité et le module de localisation des tonales.

3.2.1 Détection des signaux à caractère tonal

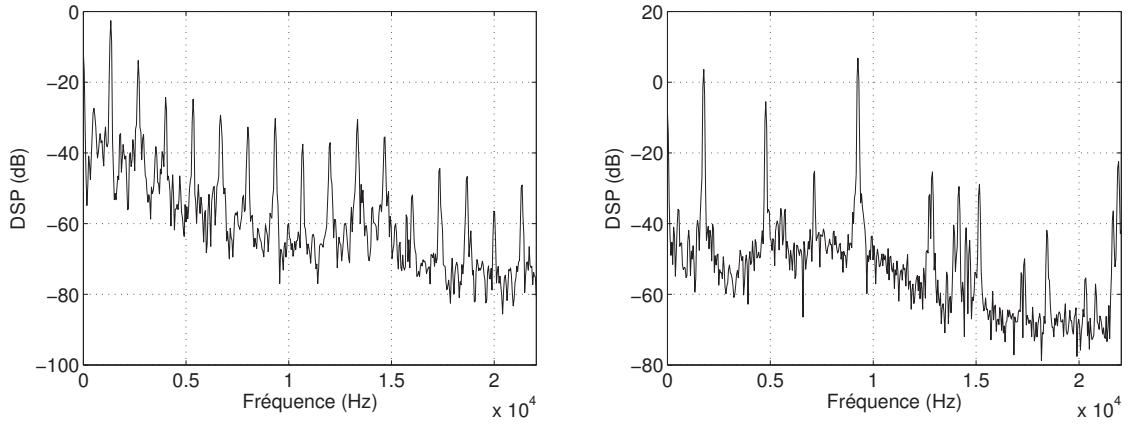
3.2.1.1 Définition

Les signaux à caractère tonal peuvent être dissociés en deux classes : harmoniques et non harmoniques.

Les signaux harmoniques : composés d'une série d'harmoniques déterminées par la fréquence fondamentale f_0, leur amplitude A_i et leur phase ϕ_i (voir figure (3.3-a) et d'une éventuelle composante de bruit ($e(t)$). Ils s'écrivent sous la forme :

$$x_{harm}(t) = \sum_{i=1}^{N} A_i \cos\Big(2\pi i f_0 + \phi_i \Big) + e(t), \tag{3.1}$$

où N détermine le nombre des harmoniques.



(a) (b)

FIGURE 3.3 – (a) Trame harmonique d'un signal de violon (2048 éch.), (b) Trame non harmonique d'un signal de glockenspiel (2048 éch.).

Les signaux non harmoniques : composés du bruit ($e(t)$) et de tonales isolées non harmoniquement liées. Chacune de ces tonales est déterminée par sa fréquence f_i, son amplitude A_i et sa phase ϕ_i (3.3-b). Ils s'écrivent sous la forme :

$$x_{tonal} = \sum_{i=1}^{N} A_i \cos\left(2\pi f_i + \phi_i \right) + e(t), \tag{3.2}$$

où N détermine le nombre des harmoniques.

3.2.1.2 Méthode de détection de tonalité

Une façon simple de déterminer la nature du signal de type bruit ou tonal est la mesure de la Platitude Spectrale, en anglais *Spectral Flatness Measure* (SFM). Cette mesure est définie par le rapport entre la moyenne géométrique G_m et la moyenne arithmétique A_m du spectre du signal :

$$SFM_{dB} = 10 \log_{10}\left(\frac{G_m}{A_m} \right) \tag{3.3}$$

où $Gm = \sqrt[N]{\prod_{k=0}^{N-1} |X(k)|^2}$ et $Am = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2$.

La mesure de la platitude spectrale SFM_{dB} s'interprète de la manière suivante : une valeur de SFM proche de 1 indique que le spectre est de puissance est réparti sur toutes les bandes de fréquence (cas d'un bruit). A l'inverse, une valeur de SFM_{dB} proche de 0 indique que la puissance est relativement concentrée sur des fréquences particulières (cas d'une somme de sinusoïdes).

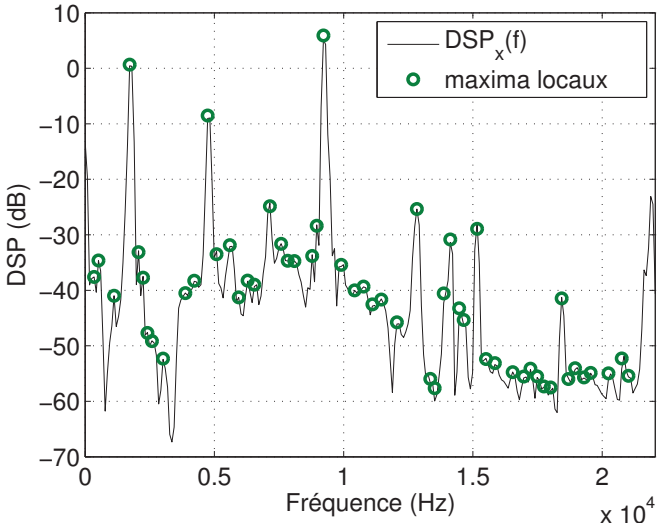


FIGURE 3.4 – Maxima locaux sur le spectre de fréquence d'une trame de glokenspiel de 512 éch. échantillonnée à 44.1 kHz.

Ainsi, la mesure de SFM_{dB} permet de classer les signaux en signal tonal ou bruit par le calcul d'un indice de tonalité défini par :

$$\alpha = min\left(\frac{SFM(dB)}{SFM_{min}}, 1 \right) \tag{3.4}$$

où $SFM_{min} = -60$ dB [Johnston 1988].

Les valeurs de l'indice de tonalité α sont dans l'intervalle [0, 1]. Plus α est proche de 0, plus le son est de nature bruit. A l'inverse, plus α est proche de 1, plus le signal est à caractère tonal.

Ainsi, l'indice de tonalité est une mesure qui peut être utilisé comme un détecteur de tonalité. Afin d'avoir une décision finale, la valeur de α est comparée à un seuil τ. Nous avons fixé empiriquement τ à 0.2. Ainsi, nous proposons, pour chaque trame d'analyse d'indice i, un indicateur de tonalité booléen M_i défini par :

$$M_i = \begin{cases} 1 & \text{si } \alpha_i \geq 0.2 \\ 0 & \text{sinon} \end{cases} \tag{3.5}$$

où α_i est l'indice de tonalité correspondant à la trame d'indice i.

3.2.2 Estimation de la position des tonales

Une méthode de détection des composantes tonales consiste, en première étape, à repérer les pics, ou les maxima locaux, qui ressortent du spectre de puissance $X(k)$ (voir Figure 3.4). Une composante $X(k)$ est dite tonale si elle est supérieure à ses voisines

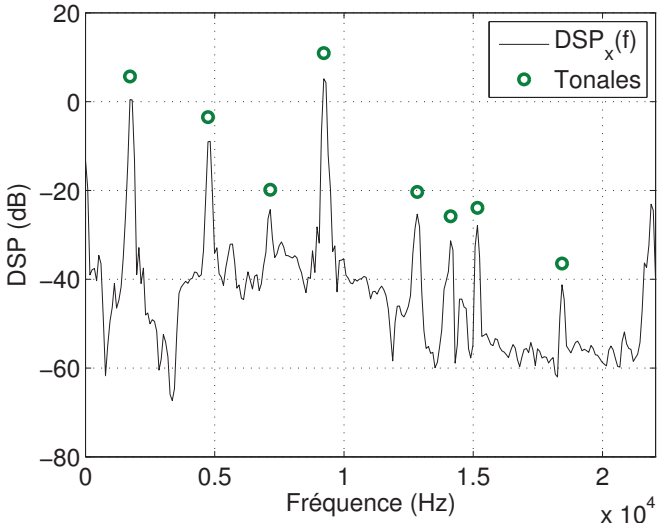


FIGURE 3.5 – Détection des tonales.

immédiates ($f \pm 1$) (maximum local) et si elle est supérieure de 7 dB à ses autres voisines écartées d'elle d'une distance au plus de Δf. Ceci se traduit par les inégalités suivantes :

$$X(k) > X(k-1) \tag{3.6}$$

$$X(k) \geq X(k+1) \tag{3.7}$$

$$X(k) - X(k+j) \geq 7 \;\; dB \;\; \forall \;\; |j| \leq \Delta f \tag{3.8}$$

où k représente les fréquences discrètes. Pour une résolution fréquentielle de 512 points, le modèle psycho-acoustique de la norme MPEG-1 [ISO/IEC 1993] définit Δf comme suit :

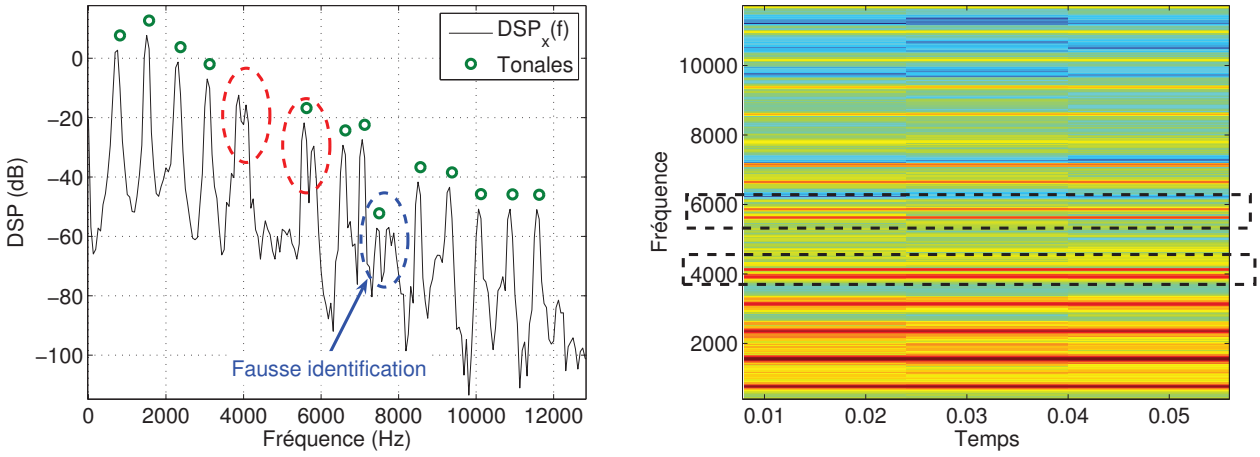
$$\Delta f \begin{cases} = 2 & \text{si} \quad 2 < k < 63 \\ \in [2,3] & \text{si} \quad 63 \leq k < 127 \\ \in [2,6] & \text{si} \quad 127 \leq k < 250 \end{cases} \tag{3.9}$$

Le domaine de variation de plus en plus large pour Δf est dû au fait que la résolution fréquentielle de l'oreille humaine est importante pour les basses fréquences.

Le résultat de la détection des composantes tonales sur une trame de glokenspiel échantillonnée à 44.1 kHz par l'algorithme décrit précédement est illustré par la figure 3.5. Les positions de tonales détectées coincident avec les pics qui ressortent du spectre.

Problème d'identification des tonales dans les signaux codés/décodés

L'estimation des positions de tonales par l'algorithme décrit précédement utilise une résolution fréquentielle de $N = 512$ points. Ce type de résolution est cependant insuffisant pour le cas des signaux présentant un phénomène de dissonance car les déviations



(a) (b)

FIGURE 3.6 – (a) Spectre d'une trame de 512 échantillons de trompette codée/décodée par le aacPlus à 16 kbits/s (2048 éch.), (b) Spectrogramme de la trame correspondante.

harmoniques sont faibles (inférieures à 200 Hz). Il est par exemple impossible, avec un tel critère, de détecter deux tonales espacées de moins de 200 Hz comme illustré sur la figure 3.6 (a) où nous présentons le spectre d'une trame de 512 échantillons d'un signal de trompette codé/décodé par le codeur aacPlus à 16 kbits/s.

Sur le spectrogramme de la figure 3.6 (b), on note la présence de deux raies tonales en 4 kHz espacées d'une déviation harmonique inférieure à 200 Hz. Egalement, on note une seconde déviation en 5.6 kHz. Ces deux séries d'harmoniques ne sont pas toutes détectées par l'algorithme. En effet, en raison de la faible résolution fréquentielle ($N = 512$), les tonales marquées par le cercle rouge ne vérifient pas le critère de tonalité imposé par l'algorithme présenté précédement (écart entre composantes voisines supérieur à 7 dB). En outre, on note une fausse identification de la tonale marquée par le cercle bleu.

Enveloppe spectrale et identification des tonales

Dans le but d'éviter les problèmes liés à la fausse ou la non détection de composantes tonales, nous proposons une légère modification de l'algorithme d'identification des tonales présenté précédemment, qui consiste à réduire l'écart entre les composantes voisines (Δf) ainsi que le rapport de puissance entre ces dernières (4 dB au lieu de 7 dB). Afin de s'assurer que les composantes détectées représentent bien des composantes tonales, nous proposons d'utiliser un seuillage fondé sur l'enveloppe spectrale.

La stratégie d'identification proposée comprend alors différentes étapes de calcul que nous détaillons dans ce qui suit.

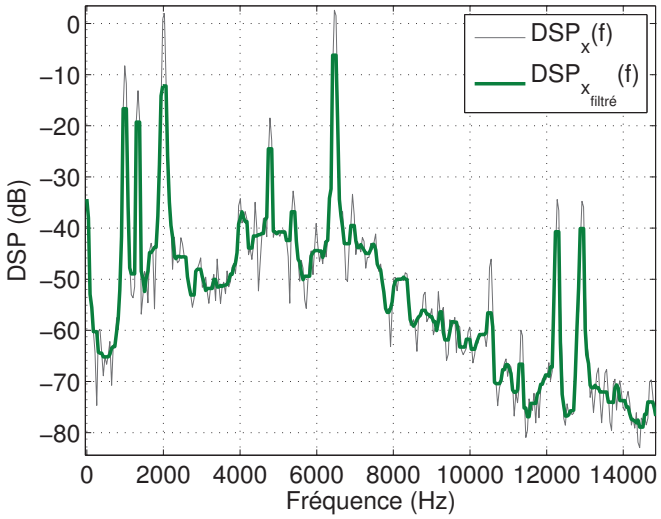


FIGURE 3.7 – DSP d'une trame de 1024 éch. de glokenspiel et DSP filtrée par un filtre médian.

.

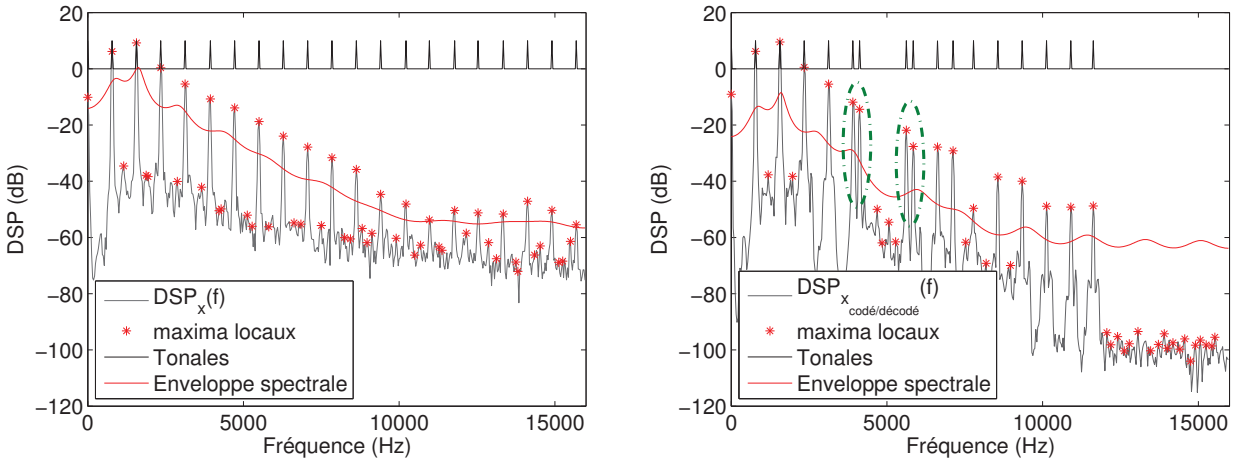
Passage dans le domaine fréquentiel et lissage du spectre : La méthode est fondée sur l'analyse par Transformée de Fourier Discrète (TFD) sur 1024 échantillons. Ce choix ayant un effet, d'une part, sur la réduction de la complexité du traitement et d'autre part sur la résolution fréquentielle dans le cas du codage à bas débit.

Sur le spectre du signal ainsi calculé, on applique un lissage fréquentiel par un filtre médian, avant de procéder au processus de détection de tonale. Ce choix de filtrage est guidé par le souci de diminuer le nombre de pics qui ressortent du spectre du signal (voir figure 3.7).

Détection des pics : la deuxième étape consiste à repérer les pics qui ressortent du spectre de puissance lisse. L'ensemble S_T des maxima locaux est défini par :

$$S_T = \{X(k)|X(k) > X(k \pm 1) \text{ et } X(k) > X(k + j|_{j\in\{-3,-2,2,3\}}) + 4\ dB\}. \tag{3.10}$$

Seuillage et enveloppe spectrale : A cette étape, on s'intéresse à éliminer les pics qui ressortent du spectre indiquant une fausse détection de tonales (comme illustré par le figure 3.6 -(b)). Pour cela, nous proposons d'éliminer tous les éléments de S_T inférieur à un certain seuil. Le seuil considéré est une estimation de l'enveloppe spectrale par une modélisation AutoRegressive (AR) d'ordre p. L'enveloppe que nous cherchons ne doit pas trop varier en fréquence mais doit plutôt donner l'allure générale de la distribution de l'énergie du signal. Pour cela, nous avons opté pour un ordre de prédiction $p = 15$.



(a) (b)

FIGURE 3.8 – Identification des tonales par la solution proposée : (a) sur signal original, (b) sur le signal codé/décodé à 16 kbits/s.

Nous présentons sur la figure 3.8 le résultat de l'implémentation de l'algorithme précité sur une séquence de trompette échantillonnée à 44.1 kHz et sa version codée/décodée avec le aacPlus à 16 kbits/s échantillonnée à 32 kHz. La méthode proposée permet une réduction notable des maximas locaux inutiles, aussi bien sur le spectre du signal original que sur sa version codé/décodé. En outre, le système proposé détecte correctement toutes les composantes tonales, en particulier la série des tonales en 4 et 5.6 kHz du signal décodé.

3.2.3 Quantification et codage de décalage de position

Les positions des tonales, réparties sur la bande de fréquence synthétisée par le décodeur SBR ([4, 11.7 kHz]), couvrent une large dynamique. Pour réduire la quantité d'information à transmettre pour la correction des tonales après décodage, nous proposons d'effectuer, au niveau du codeur, un décodage "à blanc" et de transmettre simplement les différences entre les positions originelles des tonales et celles détectées sur le signal décodé. Le vecteur des décalages est noté Δf. Il est de taille variable, dépendante du nombre de tonales dans la bande répliquée. Pour déterminer ce vecteur Δf, chaque tonale de la bande SBR du signal codé-décodé est appariée à la tonale la plus proche du signal original, qui ne doit être appariée qu'à une seule tonale du signal codé-décodé, la plus proche. Pour les tonales du signal codé-décodé non appariées, une valeur spéciale est fixée dans le vecteur Δ_f (voir étape codage). Quant aux tonales du signal original n'ayant pas d'équivalente dans le signal codé-décodé, elles ne sont pas traitées.

La technique de quantification scalaire uniforme est adoptée pour le codage des éléments du vecteur Δf. Le processus de codage se compose des deux étapes suivantes :

– Une première étape de quantification de décalage Δf. On réalise ici une quantification scalaire uniforme de chacune des valeurs discrètes du vecteur Δf sur n bits, couvrant une dynamique de $\pm$ f_0. Le choix de f_0 dépend de la nature du signal selon harmonique ou tonal :
  - pour les signaux harmoniques : f_0 représente la valeur de la fréquence fondamentale discrète. Ce choix se justifie par le fait que les tonales sont harmoniquement liées et sont espacées d'une distance égale à la fréquence fondamentale. L'avantage de l'utilisation de cette plage est que la valeur de la fréquence fondamentale peut être déduite directement du signal décodé puisque la partie basse fréquence est complètement préservée par le codeur/décodeur cœur.
  - pour les signaux non harmoniques : f_0 représente le décalage maximal des tonales situées dans la bande haute fréquence générée par le codeur SBR par rapport aux tonales du signal original.
– Une seconde étape de codage de Δf quantifiée. Pour un codage sur n bits, les nombres -2^{n-1} à $2^{n-1}-1$ sont codés selon un codage de Gray de manière à limiter l'impact d'une erreur binaire sur le canal auxiliare (dans la perspective de l'utilisation du tatouage comme canal auxiliaire). Comme la différence de position entre les harmoniques ne peut jamais atteindre la valeur de la fréquence fondamentale, le mot représentant -2^{n-1} sera utilisé pour coder les tonales à éliminer.
  Le débit nécessaire à la transmission de l'erreur de position dépend fortement du nombre de tonales identifiées dans la sous-bande [4, 11.7 kHz] ainsi que du nombre de bits alloués pour coder chaque élément du vecteur Δf. Par exemple, pour la note "LA" de fréquence fondamentale 440 Hz échantillonnée à 44.1 kHz, sur une trame d'analyse de 2048 échantillons (46.4 ms), on compte 19 tonales situées dans la sous-bande [4, 11.7 kHz]. Si on réserve 5 bits pour représenter chaque élément du vecteur Δf, le débit de transmission maximal est de l'ordre de 2 kbits/s.

3.3 Traitement côté décodeur

La solution développée dans cette section a pour but de repositionner les tonales mal synthétisées par le codeur SBR à partir du signal codé/décodé et des données auxiliaires relatives aux positions originelles des composantes tonales. Le repositionnement des tonales est réalisé par des opérations de translation spectrale. Le principe de la correction est illustré par la figure 3.9.

Le signal est décodé puis segmenté en trames d'analyse de 2048 échantillons recouvrantes de 50% avant d'être injecté dans le module de détection de tonalité. La détection de tonalité, réalisée conformément à la technique présenté dans le paragraphe 3.2.1.2, permet de définir le traitement à appliquer sur la trame. Pour les trames non tonales, aucun

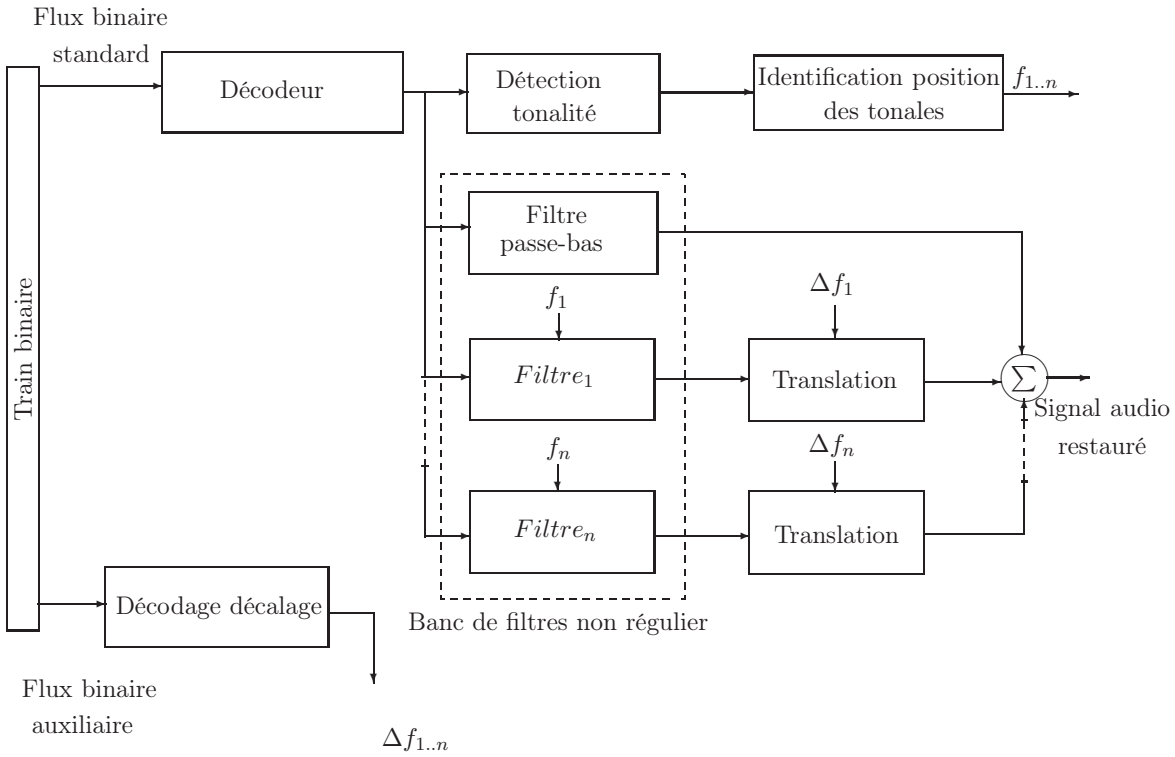


FIGURE 3.9 – Diagramme de fonctionnement côté décodeur.

traitement n'est associé. Dans le cas contraire, les composantes tonales situées dans la bande fréquentielle [4, 11.7 kHz] sont identifiées conformément à la technique présentée au paragraphe 3.2.2. La gamme de fréquence choisie correspond à la largeur de bande synthétisée par le décodeur SBR dans les configurations choisies.

La correction des positions de tonales repose sur des translations spectrales selon les erreurs Δf transmises et décodées. A l'aide d'un banc de filtres non régulier, on subdivise la partie haute fréquence, $X_{HF}(f)$, synthétisée par le décodeur SBR, en sous-bandes en fonction des positions de tonales f_i détectées sur le signal décodé. On définit deux cataégories de sous-bandes :

- des sous-bandes tonales de largeur 100 Hz centrées autour des composantes fréquentielles tonales comme illustré sur la Figure 3.11.
- des sous-bandes de différentes largeurs qui représentent la partie basse fréquence générée par le codeur cœur AAC et les sous-bandes hautes fréquences non tonales (voir Figure 3.11).

Les sous-bandes ainsi définies sont réalisées à partir d'un banc de de 160 filtres sélectifs modulés de largeur 100 Hz réalisant une partition uniforme de l'axe des fréquences comme le montre le tracé de la figure 3.10. Les réponses impulsionnelles des filtres h_k

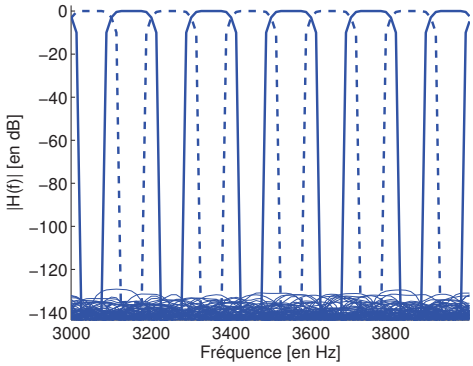


FIGURE 3.10 – Réponses fréquentielles d'un banc de filtres régulier.

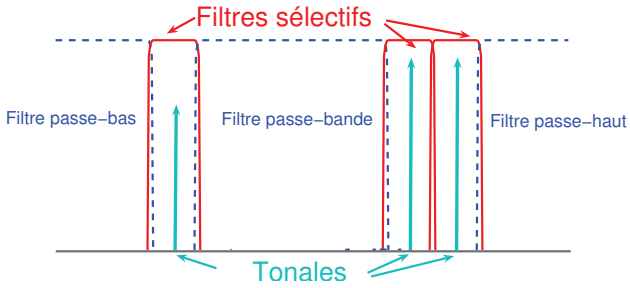


FIGURE 3.11 – Exemple d'un banc de filtres non régulier.

sont contruites à partir d'un filtre prototype passe-bas idéal h de la façon suivante :

$$h_k(n) = h(n) * \cos(\frac{2k+1}{2M} n\pi), \quad n = 0...N-1. \tag{3.11}$$

où M représente le nombre de filtres et N est la longueur du filtre prototype. Ce banc de filtres n'est pas à reconstruction parfaite mais le RSB est supérieur à 90 dB.

Pour chaque sous-bande tonale d'indice i, on réalise une transation spectrale d'un décalage de Δf_i. Les sous-bandes hautes fréquences ainsi remises en position sont ajoutées au reste du signal (partie basse fréquence [0, 4 kHz] et haute fréquence non modifiées).

La technique de translation retenue pour la correction des positions de tonale repose sur une modulation à bande latérale unique. Nous développons dans ce qui suit les détails de cette technique.

3.3.1 Translation par modulation d'amplitude

La modulation d'amplitude, appelée aussi modulation à double bande latérale (DBL), se traduit par la translation du spectre du signal modulant $m(t)$ autour des fréquences $-f_p$ et $+f_p$, où f_p désigne la fréquence de la porteuse. La translation est assurée par multiplication du signal $m(t)$ par une forme d'onde porteuse $A\cos(2\pi f_p t)$. Le signal modulé

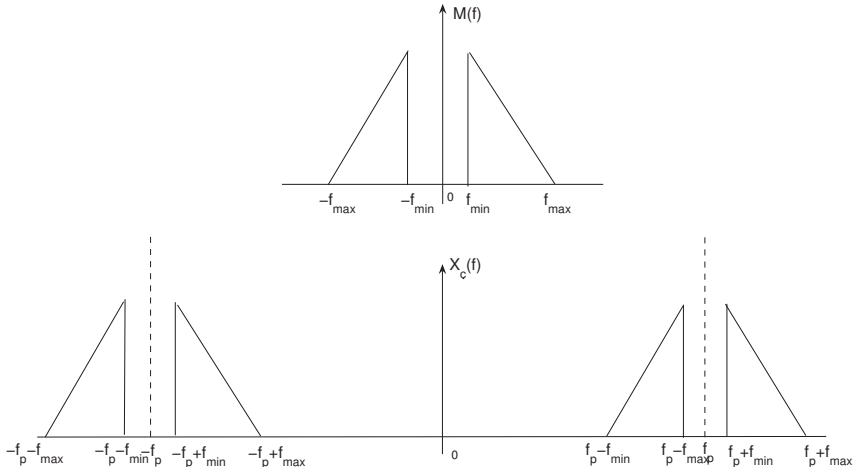


FIGURE 3.12 – Illustration schématique d'une modulation d'amplitude à porteuse sinusoïdale supprimée dans le domaine spectral : spectre du signal modulant (en haut), spectre du signal modulé (en bas).

$\phi_{DBL}(t)$ résultant est donné par :

$$\Phi_{DBL}(t) = Am(t)\cos(2\pi f_p t) \tag{3.12}$$

où $m(t)$ désigne le signal modulant et f_p est la fréquence de la porteuse.

La transformée de fourier de l'équation 3.12 s'écrit :

$$\Phi_{DBL}(f) \;=\; \frac{A}{2}\Big[\; M(f+f_p) + M(f-f_p) \;\Big] \tag{3.13}$$

où $\Phi_{DBL}(f)$ et $M(f)$ représentent respectivement le spectre du signal modulé et le spectre du signal modulant.

Comme illustré sur la figure 3.12, le spectre du signal modulé se compose de deux copies du spectre du signal modulant, $M(f)$, autour de la fréquence $\pm f_p$. Bien que la modulation à double bande offre une liberté dans le choix de la partie du spectre à translater, la translation d'une ou plusieurs parties du spectre par un tel processus requiert des opérations de filtrage. Toutefois, le processus de filtrage de la bande latérale (supérieure ou inférieure) est difficile à appliquer à cause des filtres très sélectifs qui sont exigées [Fontolliet 1983].

Pour une simple translation en fréquence, la sinusoïde peut être remplacée par une exponentielle complexe, $\exp(2\pi f_p t)$. Cependant, cette solution se traduit par un signal modulé complexe.

3.3.2 Translation par modulation à Bande Latérale Unique

La modulation à Bande Latérale Unique (BLU) est similaire à la modulation à double bande, mais au lieu d'utiliser l'ensemble du spectre du signal à moduler, une sélection de

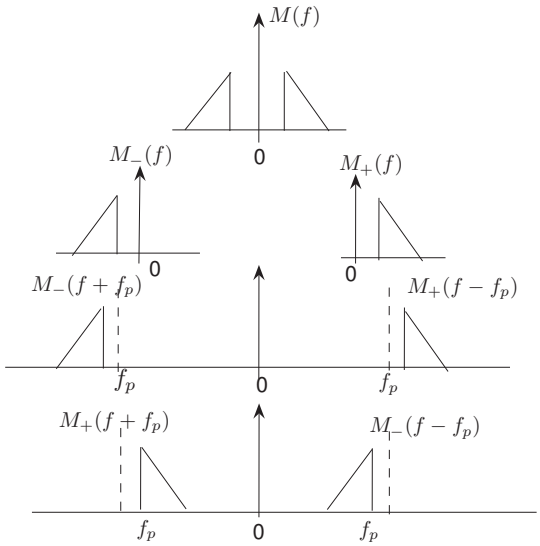


FIGURE 3.13 – Représentation fréquentielle schématique d'un modulateur à bande latérale unique.

la bande latérale inférieure ou supérieure est appliquée. La sélection des bandes latérales inférieure ou supérieure forme respectivement la modulation à la bande latérale inférieure ou la modulation à bande latérale supérieure.

Pour éliminer une des bandes latérales, le modulateur BLU met en œuvre une méthode progressive fondée sur l'utilisation d'une transformée de Hilbert. Nous détaillons dans ce qui suit le principe de cette méthode.

Analyse mathématique de la modulation BLU

Comme illustré sur la figure 3.13, le spectre du signal modulé à bande latérale unique supérieure est donné par :

$$\Phi_{BLUS}(f) = M_+(f - f_p) + M_-(f + f_p) \tag{3.14}$$

où $M_+(f)$ et $M_-(f)$ représentent respectivement les sous-bandes latérales correspondant aux fréquences positives et négatives du spectre $M(f)$ et f_p est la fréquence de la porteuse. La transformée de l'équation 3.14 dans le domaine temporel est donnée par :

$$\begin{aligned} \Phi_{BLUS}(t) &= \frac{1}{2}\Big(\, m_a(t)\exp(\mathrm{j}2\pi f_p t) + m_a^*(t)\exp(-\mathrm{j}2\pi f_p t) \, \Big) && (3.15) \\ &= \mathrm{R}[m_a(t)\exp(\mathrm{j}2\pi f_p t)] && (3.16) \end{aligned}$$

R désigne la partie réelle, $m_a(t)$ est le signal analytique correspondant à $m(t)$ (défini par l'equation 2.2 du chapitre 2) et $*$ désigne le complexe conjugué. Les détails de ce calcul sont présentés en annexe E.

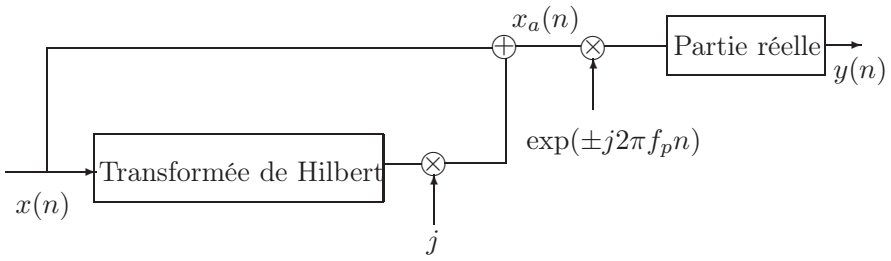


FIGURE 3.14 – Schéma général d'un modulateur à Bande Latérale Unique.

Le procédé de la translation spectrale par modulation a bande latérale unique se résume donc par le schéma de la figure 3.14. La translation par modulation à bande latérale unique est une technique facile à mettre en oeuvre et peu complexe. Contrairement à la technique de translation par modulation d'amplitude qui requiert une opération de filtrage pour l'élimination des bandes inutiles, la BLU offre la liberté du choix de la partie du spectre à translater sans la mise en œuvre d'opérations de filtrage. Dans le cadre de notre étude, nous avons adopté la technique BLU pour la correction des positions des composantes tonales.

3.4 Evaluation des performances du système proposé

Le système de correction de tonales proposé a été testé sur des séquences de signaux audio mono à caractère fortement harmonique, en particulier des séquences de trompette, violon et pipe échantillonnées à 44100 Hz et sur un extrait de musique à caractère tonal non harmonique, le glokenspiel échantillonné à 44100 Hz. Le codeur à extension de bande utilisé est la version normalisée du codeur HE-AAC (aacPlus). Cette version offre des débits de compression allant de 8 à 160 kbits/s, avec une qualité jugée transparente à 24 kbits/s en mono.

Afin de faciliter le traitement, nous travaillons dans cette partie avec des signaux originaux échantillonnés à 44.1 kHz et des signaux codés/décodés échantillonnés à 32 kHz. Le débit de compression considéré est 16 kbits/s. Les paramètres du vecteur Δf, décalage à appliquer, sont codés sur 6 bits et transmis par un canal auxiliaire à faible débit (inférieur à 500 bps).

Comme illustré dans le tableau 3.2, le codeur cœur AAC génère des signaux audio à bande passante de 4 à 7.2 kHz pour les débits de compression allant de 16 à 20 kbits/s. Pour les fréquences inférieurs à ces bandes, aucun traitement n'est associé.

3.4.1 Illustration de correction d'harmonicité/tonalité

Nous présentons sur la figure 3.15 (a, b) une illustration du phénomène de rupture d'harmonicité détecté sur une séquence de trompette échantillonnée à 44.1 kHz et co-

dée/décodée par le codeur HE-AAC à 16 kbits/s. Ces composantes entrent en dissonance et génèrent des artefacts perceptivement gênants qui se traduisent à l'écoute par un sifflement. On note sur la version synthétisée par le codeur HE-AAC une rupture d'harmonicité à partir de la sixième tonale autour de 4 kHz et qui s'étend jusqu'à 8 kHz. On observe une nette correction d'harmonicité sur la version restaurée du signal par l'appoche proposée (figure 3.15 (c)). En effet, les composantes tonales repositionnées sont harmoniquement liées. Une correction d'harmonicité est également observée sur la version synthétisée du signal de violon par le codeur HE-AAC à 16 kbits/s

Sur la figure 3.16(a, b), nous présentons un signal tonal non harmonique de glockenspiel et sa version codée/décodée par le HE-AAC à 16 kbits/s. On note sur le signal codé/décodé des tonales isolées synthétisées autre que les composantes tonales du signal original (exemple tonale encadrée par le rectangle pointillé). Bien que le signal d'analyse soit fortement non stationnaire, une correction de certaines composantes tonales du signal est vérifiée sur la figure 3.15 (c).

3.4.2 Evaluation objective : mesure de rugosité

Les performances du système de correction de tonalité s'évaluent de préférence par des mesures subjectives mettant en œuvre des tests d'écoute. Pour des raisons de rapidité et de coût, nous avons eu recourt à des mesures objectives. D'après la norme [3GPP 2005], l'évaluation de la qualité audio des codeurs HE-AAC se fait à travers les mesures objectives PEAQ (Perceptual Evaluation of Audio Quality) fondé sur la norme UIT BS.1387 [ITU-R 2001]. Cependant, les mesures obtenues par la version libre de peaq [2] ne coïncident pas avec les mesures présentées dans la littérature et confirmées par les tests d'écoute : une qualité transparente à 24 kbits/s. Nous avons donc adopté d'autres critères d'évaluation dédiés principalement à la mesure de la rugosité.

Plusieurs méthodes de mesure de rugosité, fondées sur une approche spectrale, ont été proposées dans la littérature. Leur principe, suggéré par Plomb et Levet [Plomb 1965], consiste à additionner les rugosités partielles des couples de tonales formant la représentation spectrale du son.

Le principe de la mesure de rugosité est schématisé par la figure 3.17. La trame d'analyse présente une liste de tonales de fréquence et d'amplitude (f_i, A_i). Pour chaque paire possible de composantes (i,j), on définit la rugosité partielle $r_{i,j}$ calculée en fonction de leur différence fréquentielle selon [Vassilakis 2001] :

$$r_{i,j} = X^{0.1} * 0.5(Y^{3.11}) * Z \tag{3.17}$$

2. http ://www-mmsp.ece.mcgill.ca/Documents/Software/index.html

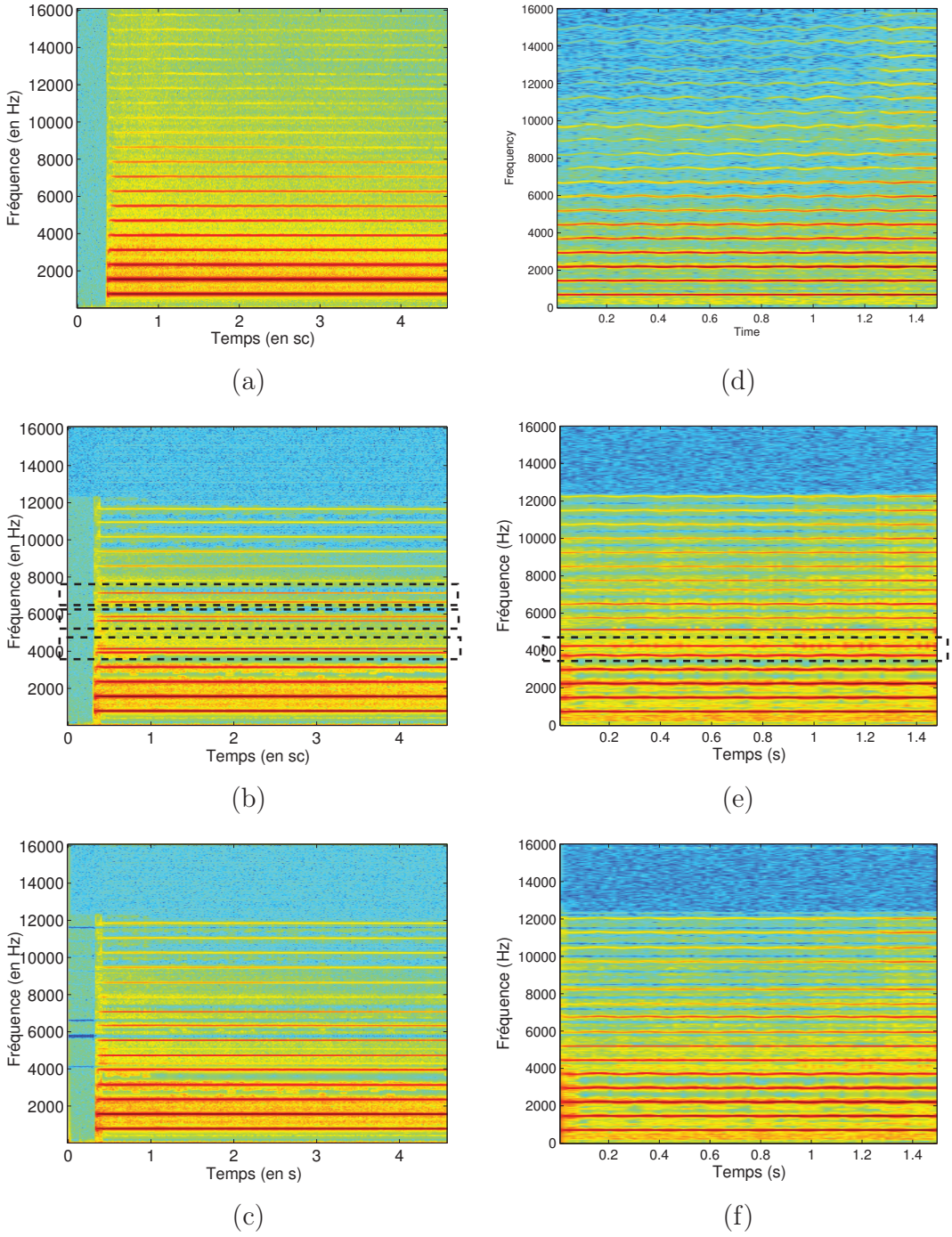


FIGURE 3.15 – Correction de l'harmonicité par le système proposé respectivement pour une séquence de trompette (à gauche) et une séquence de violon (à droite) : (a, d) spectrogrammes des signaux originaux, (b, e) spectrogrammes des signaux codés/décodés à 16 kbits/s, (c, f) spectrogrammes des signaux restaurés.

où

$$\begin{cases} X = & A_{min} * A_{max} \\ Y = & 2A_{min}/(A_{min} + A_{max}) \\ Z = & e^{b1s(f_{max}-f_{min})} - e^{b2s(f_{max}-f_{min})} \end{cases} \tag{3.18}$$

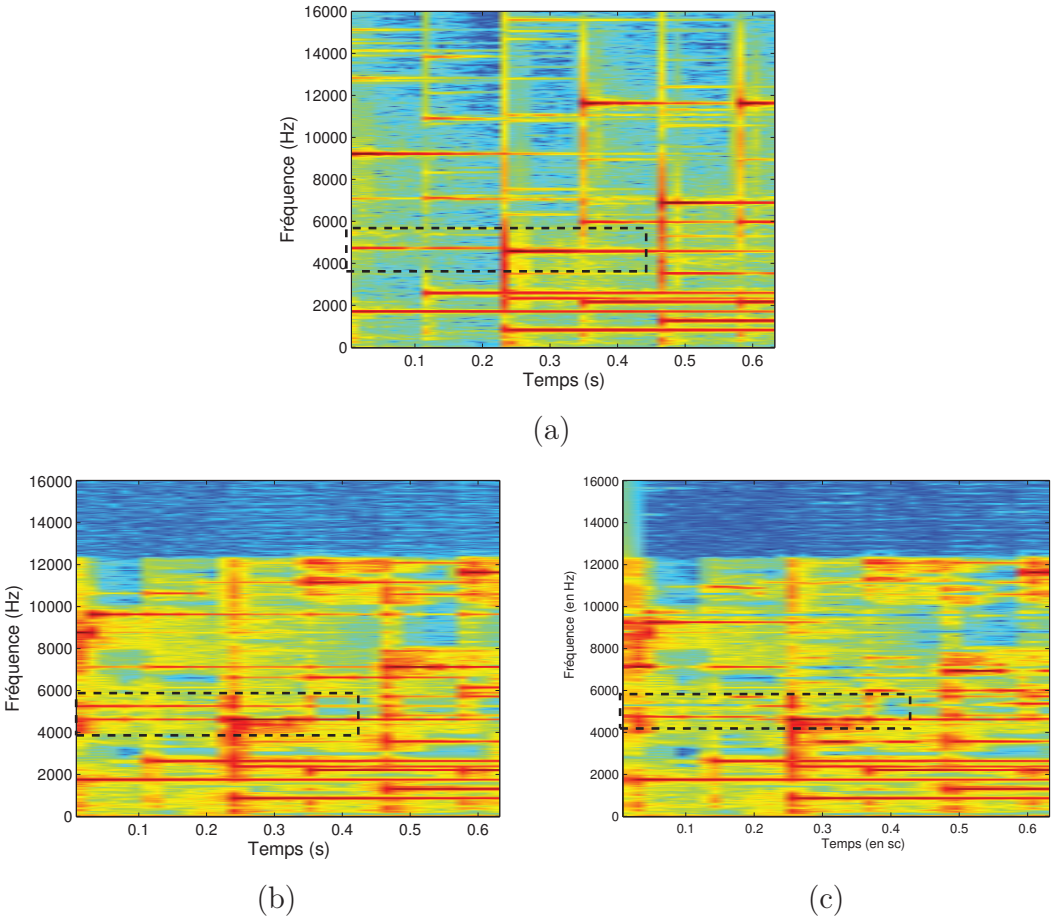


FIGURE 3.16 – Correction de tonalité par le système proposé pour une séquence de glokenspiel : (a) spectrogramme du signal original, (b) spectrogramme du signal codé/décodé à 16 kbits/s, (c) spectrogramme du signal restauré.

Avec :

$A_{min} = min\{A_i, A_j\}$; $A_{max} = max\{A_i, A_j\}$; $f_{min} = min\{f_i, f_j\}$; $f_{max} = max\{f_i, f_j\}$; $b1 = 3.5$; $b2 = 5.75$; $s = 0.24/(s1 f_{min} + s2)$; $s1 = 0.0207$ et $s2 = 18.96$.

La mesure de la rugosité partielle ne dépend pas uniquement de la déviation entre les composantes tonales mais également de l'intensité des dernières (le terme $X^{0.1}$ de l'équation 3.17 liée à l'amplitude des tonales) ainsi que du degré de la fluctuation d'amplitude (le terme $Y^{3.11}$ de l'équation 3.17 lié à la moyenne d'amplitude des tonales). Le calcul de ces deux paramètres est fondé sur [Terhardt 1974] et ajusté par [Vassilakis 2001].

Toutes les rugosités partielles, ainsi calculées, sont par la suite additionnées pour fournir la rugosité totale R. Pour un son comportant N composantes de fréquences et d'amplitudes notées $\{f_i, a_i\}$, la rugosité estimée R est donnée par :

$$R = \sum_{i,j=1,\quad i \neq j}^{N} r_{i,j}. \tag{3.19}$$

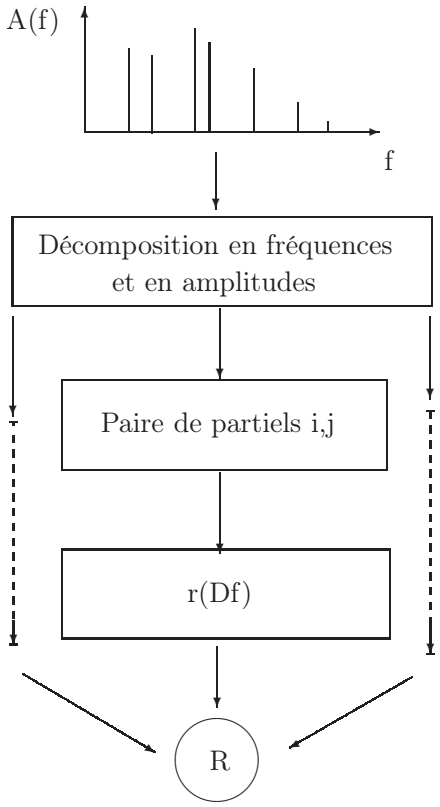


FIGURE 3.17 – Principe de la mesure de la rugosité R.

La mesure de rugosité présentée précédement a été retenue pour évaluer les performances du système proposé. Notons que cette mesure est une valeur intrinsèque qui dépend fortement de la nature du signal. Il s'agit donc de comparer la rugosité des deux séquences codé/décodé et corrigé par rapport à l'originale. On liste dans le tableau 3.3, les valeurs de la rugosité estimées relatives à 4 signaux fortement harmoniques suite à leur correction par le système proposé. Le but de la correction consiste donc à s'approcher au mieux à la valeur de la rugosité originelle.

Les résultats obtenus montrent que la solution de restauration des tonales proposée corrigent nettement pour trompette 2 et pipe les défauts du codeur standard HE-AAC à 16 kbits/s. En effet, une amélioration significative de la note de la rugosité a été observée pour les 4 extraits de musique considérés, en particulier pour la séquence pipe. Les corrections par ajustement fréquentiel d'une ou de plusieurs tonales réduit donc considérablement les phénomènes de rugosité affectant particulièrement les signaux harmoniques purs. Pour les signaux tonals tels que le glokenspiel, les variations du signal sont très rapides, elles se traduisent par des non stationnarités. Par conséquent, la mesure de la rugosité telle que présentée précédement ne peut être évalué que instantanément.

Signal audio	Original	codé/décodé	Restauré
Trompette1	71.32	67.17	67.34
Trompette2	107.49	103.99	108.14
Violon	82.21	49.9	56.74
Pipe	84.96	46.34	83.2

TABLE 3.3 – Evaluation objective de la performance du système de correction de tonalité par la mesure de rugosité. Les signaux tests sont des notes pures fortement harmoniques codées/déodées à 16 kbits/s.

3.5 Conclusion

Nous avons développé dans ce chapitre une technique de correction de l'harmonicité et de préservation de tonalité pour des signaux issus des codeurs à extension de bande, en particulier, le codeur HE-AAC. La solution proposée, dédiée principalement aux signaux tonals et fortement harmoniques, repose sur un ajustement fréquentiel d'une ou d'un ensemble de tonales par de multiples translations spectrales. Les translations spectrales sont réalisées dans le domaine temporel par le biais d'une modulation à bande latérale unique combiné à un filtrage des composantes.

Le système proposé a été évalué par un critère objectif dédié à la mesure de la rugosité perçue. Les résultats obtenus prouvent une amélioration significative de la qualité des signaux fortement harmonique. Pour les signaux tonals, le critère adopté est moins pertinent et l'évaluation de la correction se limite à des comparaisons des spectrogrammes. Pour des signaux complexes non stationnaire, une recherche d'autres critères objectifs dédiés à la perception de la qualité audio est envisagé.

Deuxième partie

Tatouage audio en traitement du son :

application à la réduction de pré-écho et à la correction d'harmonicité/tonalité

La première partie a été consacrée à la description des deux solutions proposées pour la restauration des signaux audio décodés. Dans cette partie, nous nous intéressons à l'intégration du tatouage audio dans les techniques de renforcement de la qualité des signaux audio précitées. Dans ce contexte, le tatouage audio remplace le canal auxiliaire précédent et œuvre comme une mémoire du signal originel, porteuse d'informations nécessaires pour la préservation de la tonalité et la réduction de pré-écho.

Chapitre 4

Tatouage audio, un canal de communication virtuel

Sommaire

4.1 Introduction

Les techniques de renforcement des signaux audio présentées dans les chapitres précédent exigent la présence d'un canal de transmission auxiliaire permettant de véhiculer des informations nécessaires pour le traitement correctif du son. Le tatouage audio peut jouer ce rôle en le considérant comme étant un "canal virtuel".

Notons que dans la littérature, des méthodes de tatouage robustes à la compression ont été proposées. Il s'agit soit d'insérer les informations utiles dans les espaces réservés dans le flux codé pour les méta-données [Geiger 2006], ou encore de tatouer dans le domaine compressé [Koukopoulos 2001]. Nous avons opté ici pour le tatouage additif introduit dans

les travaux de Larbi [Larbi 2005b] ce qui a l'avantage de ne pas modifier le format du flux audio, de ne pas dépendre de ce format et de ne nécessiter aucun débit supplémentaire.

Après avoir défini les objectifs du tatouage et les contraintes d'inaudibilité et robustesse, ce chapitre présente l'impact de la compression MPEG (MP3, AAC et aacPlus) sur les performances du système de tatouage utilisé. Le caractère percussif de certains signaux audio met en évidence les limites de la chaîne de tatouage utilisée. Une structure améliorée du système de tatouage est alors proposée.

4.2 Tatouage audio : objectifs et contraintes

Le tatouage (ou Watermarking en anglais) est l'art de cacher de l'information directement dans des données multimédia de façon robuste et imperceptible. Dans le contexte des signaux audio, le tatouage met à profit les imperfections du système auditif humain pour garantir l'inaudibilité du message inséré.

4.2.1 Objectifs usuels du tatouage audio

Initialement, le tatouage audio s'est fortement développé avec l'augmentation de documents sous format numérique. Les signaux audio, auxquels nous nous intéressons, sous leur forme numérique, sont très facilement reproductibles. Ainsi, des techniques de protection efficaces, telles que le tatouage, sont donc devenues indispensables pour permettre d'authentifier les auteurs des documents.

Le tatouage consiste alors à insérer un signal (une marque ou signature) dans un autre signal numérique, audio dans notre contexte. Le signal tatoué résulte de la superposition de ces deux signaux : le document tatoué contient alors des informations (également appelées données cachées) qui peuvent être utilisées à plusieurs fins [Gomes 2002b, Gomes 2003] :

- la stéganographie cherche à cacher un message secret entre un émetteur et un récepteur à travers un signal. La nature de l'information dissimulée ne revêt pas d'importance : il peut tout aussi bien s'agir d'un texte en clair que de sa version chiffrée.
- les applications de tatouage sécuritaire sont envisagées dans le contexte de la protection des documents numériques vis-à-vis des attaques d'une tierce personne, le pirate. Ce type d'application est utilisé pour l'insertion des marques de copyright, la protection des droits d'auteur, l'indexation des documents, etc.
- l'utilisation du tatouage audio comme canal auxiliaire de transmission pour véhiculer des informations additionnelles. Ces informations peuvent être destinées à l'auditeur (par exemple paroles d'une chanson, publicité ou informations sur le do-

cument) ou servir à une application cible : ainsi, dans le projet Artus [Bailly 2006], la transcription de la parole en geste du langage parlé complété est tatouée pour activer un avatar en réception. Dans un esprit similaire, de nouvelles conceptions du tatouage sont nées, où le tatouage facilite certains traitements audio en réception : annulation d'écho acoustique [Gilloire 1998, Larbi 2005b, Aicha 2008] et séparation de source [Liu 2007] par exemple et dans les travaux du projet DReaM [1]. C'est dans cette classe d'applications que nous nous situons.

4.2.2 Les principales contraintes en tatouage audio

Les principales contraintes que doit satisfaire le système de tatouage varient selon les applications. Elles sont :

L'inaudibilité

Le signal de tatouage ne doit pas être perçu par l'auditeur afin de ne pas altérer la qualité sonore de la musique. L'inaudibilité de ce signal est assurée en exploitant les propriétés psychoacoustiques de l'oreille humaine : le masquage auditif. Le principe de ce critère sera détaillé dans le paragraphe 4.3.2.

La robustesse

Le signal audio tatoué peut être affecté par le canal de communication qui le véhicule. Les performances de détection du message inséré peuvent donc se dégrader en présence d'attaques de divers types. Les dégradations peuvent être dissociées en deux catégories :

- Des perturbations illicites qui sont introduites par un pirate et qui visent à altérer le signal de tatouage inséré. On retrouve ces perturbations dans le cas où le tatouage est utilisé pour la protection de la propriété des droits d'auteur par exemple.
- Des perturbations licites qui affectent le signal audio porteur de l'information lors de son transfert ou de sa manipulation. Ce genre de perturbations inclut les modifications de format (compression MPEG), les opérations de filtrage (passe-bas, passe-haut, etc), etc.

Dans notre cas, modifier ou détruire le tatouage ne présente aucun intérêt. Les perturbations externes considérées se limitent donc aux perturbations licites, en particulier le codage audio, puisque les informations transmises visent à corriger le signal après codage/décodage.

Le débit

1. Projet DReaM : Disque Repensé pour l'Écoute Active de la Musique (http ://scrime.labri.fr/index.php/fr/researchprojects/themesderecherche/201-dream).

La contrainte de débit dépend de l'application envisagée. Lorsqu'il s'agit d'insérer une marque de copyright, le débit requis est faible, l'effort étant porté sur la robustesse. Dans le cas où le tatouage est utilisé comme un canal de communication virtuel, comme ici, le débit doit être suffisant pour transmettre les données nécessaires à la correction prévue au niveau du décodeur. Il est cependant limité par la contrainte de robustesse au codage/décodage.

La compléxité

La contrainte de compléxité est importante pour des systèmes fonctionnant en temps réel. En outre, la compléxité du système de tatouage est d'autant plus importante que les exigences en terme d'amélioration de débit de tatouage et de la robustesse aux attaques sont accrues.

4.3 Tatouage audio additif : principes de base

Plusieurs techniques de tatouage sont développées dans la littérature pour des contextes applicatifs divers [Cox 2002, Arnold 2003, Sagi 2007, Xiang 2011] : elles dépendent de la nature du document numérique (audio ou visuel) et de la robustesse souhaité du tatouage (qui ne peut pas s'enlever facilement ou au contraire fragile).

Nous considérons ici un tatouage audio additif par étalement de spectre [Kirovski 2003], en particulier proposé par Larbi [Larbi 2005a] [2]. Notre choix s'est porté sur ce système pour sa simplicité et sa robustesse aux perturbations introduites par le canal de communication, notamment le codage audio.

4.3.1 Principe du tatouage additif

Le principe du système est schématisé par la Figure 4.1. On distingue trois principaux éléments de cette chaîne, qui seront détaillés par la suite :

– L'émetteur qui consiste en la mise en forme du message binaire a_k en un signal de tatouage $w(t)$ et à l'insertion par addition dans le domaine temporel de $w(t)$ dans le signal audio $x(t)$. Ainsi, on obtient le signal tatoué $y(t)$,

$$y(t) = x(t) + w(t); \tag{4.1}$$

– Le canal de communication qui est à l'origine des perturbations apportées au signal audio tatoué, une compression audio à bas ou à très bas débit dans notre contexte ;

2. Le système de tatouage additif été élaboré initialement au département de Traitement du Signal et des Images à l'ENST (Ecole Nationale Supérieure des Télécommunications), à l'UFR Mathématiques et informatique de l'Université de Paris Descartes et à l'Unité de recherche Signaux et Systèmes de l'ENIT.

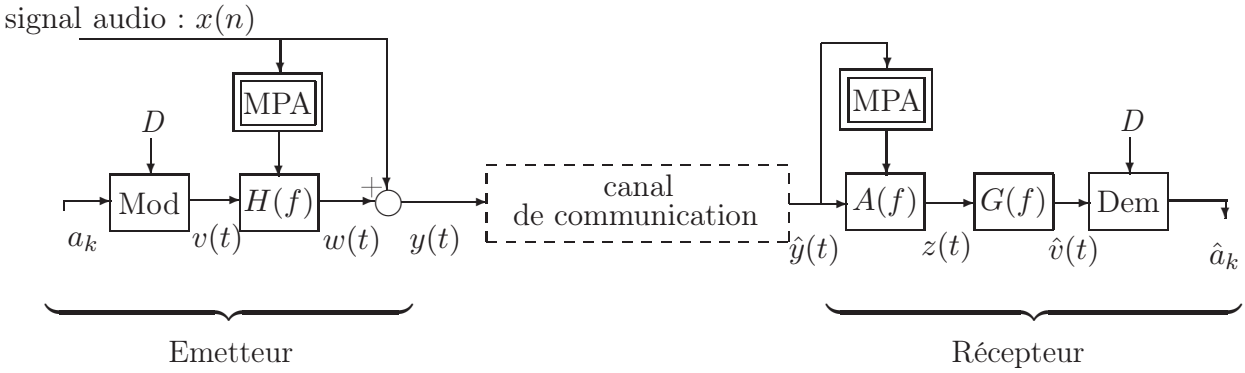


FIGURE 4.1 – Système de tatouage de référence en présence de perturbations externes.

– Le récepteur qui doit restituer l'estimation $\hat{a}_k$ du message binaire inséré à l'emission.

4.3.2 L'émetteur : générateur du signal tatoué

Le rôle de l'émetteur consiste à insérer le message binaire désiré dans le signal audio hôte. Le signal de tatouage doit être imperceptible et robuste aux différentes manipulations dont il pourrait être éventuellement l'objet.

Comme indiqué sur la figure 4.1, l'émission consite en deux opérations : une modulation et une mise en forme spectrale du signal de tatouage $w(t)$.

Modulation

Le message à transmettre est une séquence binaire constituée de symboles a_k qui sera modulée (MOD) à l'aide d'un dictionnaire D contenant M vecteurs d_k ($M = 2^l$ où l est le nombre de bits par symbole). Les éléments de chaque vecteur d_k de taille N_d suivent une loi gaussienne. La taille N_d du vecteur d_k dépend du débit de tatouge : plus N_d est élevée, plus le débit est faible. Le signal modulé $v(t)$ est donc la concaténation de ces séquences choisies en fonction des symboles représentant le message à insérer.

Mise en forme spectrale du signal de tatouage

L'inaudibilité du signal de tatouage $w(t)$ est assurée par la mise en forme spectrale du signal modulé $v(t)$ à l'aide d'un filtre perceptif $H(f)$ fondé sur un Modèle PsychoAcoustique (MPA). Ce MPA, détaillé en annexe D, exploite les propriétés perceptives du système auditif humain. Cette opération vise à donner la puissance maximale à ce signal modulé tout en respectant les contraintes d'inaudibilité. La puissance du signal $v(t)$ ainsi maximisée optimise la détection à la réception.

Le tatouage $w(t)$ est inaudible si sa Densité Spectrale de Puissance (DSP) est au dessous du seuil de masquage (fourni par le MPA) du signal audio sur chaque fenêtre de

traitement, i.e. :

$$S_w(f) \leq M_x(f), \quad |f| \leq \frac{Fe}{2}, \tag{4.2}$$

où $S_w(f)$ désigne la DSP du tatouage $w(t)$ et $M_x(f)$ le seuil de masquage du signal audio $x(t)$. Cette contrainte peut s'exprimer :

$$S_v(f)|H(f)|^2 \leq M_x(f) \tag{4.3}$$

où $S_v(f)$ désigne la DSP de $v(t)$.

En choisissant $v(t)$ un bruit blanc de puissance unité, $S_v(f) = \sigma_v^2 = 1$, il vient :

$$|H(f)|^2 = M_x(f) \tag{4.4}$$

Le filtre $H(f)$ vérifiant la relation 4.4 est implémenté comme un filtre tout-pôles :

$$H(z) = \frac{b_0}{1 + \sum_{i=1}^{P_h} h_i z^{-i}}. \tag{4.5}$$

4.3.3 Le récepteur : égalisation, débruitage et détection

En se référant aux résultats théoriques en communications numériques, dans le cas classique d'un canal Bruit Blanc Additif Gaussien (BBAG) et en absence d'interférences entre symboles (IES), le récepteur optimal est un détecteur recherchant dans D le mot le plus proche de celui reçu.

Si l'on considère le contexte de tatouage audio, le canal de communication introduit :

- un bruit très fort, corrélé et non gaussien, constitué par le signal audio $x(t)$;
- des IES dues au filtre de mise en forme H ;
- une non-stationnarité du canal liée à la variation d'une fenêtre à l'autre de la Réponse Impulsionnelle (RI) de H.

Ces contraintes pèsent sur le choix de récepteur optimal. S. Larbi [Larbi 2005a] a étudié l'apport des structures d'égalisation dans ce système de tatouage. Elle a montré en particulier que le problème posé par la chaîne de tatouage était équivalent à une égalisation aveugle d'un canal non causal et à minimum de phase.

Filtrages de Zero-Forcing et de Wiener

La détection du message tatoué est fondée sur une réalisation cascadée d'un filtre de Zero-Forcing $A(f)$ et d'un filtre de Wiener $G(f)$ (voir Figure 4.1).

a) Le filtre zero-forcing, $A(f)$, est utilisé pour inverser la mise en forme spectrale du signal modulé réalisée par $H(f)$ au niveau de l'émetteur,

$$A(f) = \frac{1}{\hat{H}(f)}. \tag{4.6}$$

Ce filtre utilise une estimation $\hat{H}(f)$ du filtre de mise en forme $H(f)$ utilisé par l'émetteur (voir Figure 4.1). En exploitant le fait que $x(t)$ et $y(t)$ sont perceptivement similaires, $H(f)$ est alors approché par $\hat{H}(f)$ qui correspond au seuil de masquage de $y(t)$.

b) A ce stade, le signal de tatouage est encore noyé dans le signal audio. Ainsi, un filtre de Wiener G est utilisé afin de réduire l'effet du signal hôte [Diniz 2005]. Le filtrage de Wiener est utilisé en communications numériques lorsque la transmission est perturbée par un bruit coloré en plus de l'IES. Cet égaliseur est optimal au sens de l'Erreur Quadratique Moyenne (EQM) entre le vecteur estimé $\hat{v}(t)$ et le vecteur émis $v(t)$.

Soit $z(t)$ l'entrée du filtre $G(z)$, la sortie du filtre $\hat{v}_t$ est donnée par :

$$\hat{v}(t) = \sum_{i=-\infty}^{+\infty} g(i)z(t-i) \tag{4.7}$$

où g représente la réponse impulsionnelle du filtre $G(z)$ et

$$\begin{aligned} z(t) &= \hat{y}(t) * a(t) && (4.8)\\ &= (v(t) * h(t) + x(t)) * a(t) && (4.9)\\ &\approx v(t) + x(t) * a(t) && (4.10) \end{aligned}$$

avec $a(t)$ la réponse implusionelle du filtre $A(f)$.

En minimisant la quantité : $EQM = E[(v(t) - \hat{v}(t))^2]$ et en exploitant le fait que $v(t)$ et $x(t) * a(t)$ sont non corrélés, la détermination de la RI g de ce filtre passe par la résolution des équations de Wiener-Hopf données par :

$$r_v(k) = \sum_{i=-\infty}^{+\infty} g(i)r_z(k-i) \quad \forall \ k \tag{4.11}$$

où $r_v(k) = E[v(t)v(t-k)]$ et $r_z(k) = E[z(t)z(t-k)]$ sont respectivement les fonctions d'autocorrélation d'ordre k de $v(t)$ et $z(t)$. Les valeurs de $r_v(k)$ sont connues grâce au dictionnaire à la réception et les valeurs de $r_z(k)$ sont estimées à partir du signal $z(t)$. Comme la RI $g(t)$ est de longueur infinie, elle est tronquée en un nombre fini de coefficients. Le vecteur obtenu à la sortie de filtre de Wiener est :

$$\hat{v}(t) = \sum_{i=-Q_w}^{P_w} g(i)z(t-i) \tag{4.12}$$

où P_w et Q_w sont les ordres des parties causale et non causale de G.

Détection du message inséré

La phase de démodulation et de décision est fondée sur un détecteur par corrélation, qui compare le vecteur $\hat{v}(t)$ issu du filtre de Wiener avec les éléments du dictionnaire. Le processsus de détection fait appel à deux opérations successives :

1. le démodulateur calcule la corrélation entre le signal reçu, sur des fenêtres de taille le temps symbole N_d, et chaque élément du dictionnaire,
2. on choisit parmi les M valeurs de corrélation obtenues la plus grande, ce qui donne le symbole associé à l'élément du dictionnaire choisi.

De proche en proche, on reconstruit ainsi la séquence des symboles reçue.

Dans une application temps réel, la désynchronisation constitue un des facteurs de perturbation auxquels un système de tatouage se doit d'être robuste. Ce problème nécessite de recourir à des mécanismes de synchronisation de l'information dans la chaîne de tatouage [Proakis 2001, Gomes 2002a, Baras 2002]. Dans notre cas, le problème de synchronisation est moins critique car le signal de tatouage est inséré au début du fichier audio et l'application envisagée n'est pas en contexte temps réel.

4.4 Evaluation des performances du système de tatouage en présence d'une compression MPEG

S. Larbi a étudié dans [Larbi 2005a] les performances du système de tatouage et l'efficacité des stratégies d'insertion pour le cas d'une compression MPEG 1 couche 1 pour les deux débits de compression : 96 et 64 kbits/s. Dans cette section, nous présentons une étude expérimentale des performances du système de tatouage dans le cas d'une compression MP3, AAC et aacPlus pour quatre signaux audio : deux signaux percussifs (castagnettes et tabla) et deux signaux non percussifs (violon et pop) échantillonnés à 44100 Hz. Deux objectifs sont visés par cette étude. Le premier est de confirmer l'efficacité du système de tatouage en termes de Taux d'Erreur Binaire (TEB) en présence d'un canal de communication représenté par une compression MPEG (MP3, AAC et aacPlus). L'influence de la largeur de bande des vecteurs du dictionnaire d'émission sera également étudié. Le deuxième objectif est d'évaluer les performances du système de tatouage de référence dans le contexte du multicodage.

Taux d'erreur binaire

La mesure de la fiabilité de détection de tatouage est calculée à travers la probabilité d'erreur de transmission. Cette probabilité d'erreur peut être estimée par le taux d'erreur binaire (TEB) donné par le ratio entre le nombre de bits erronés sur le nombre de bits total émis,

$$\text{TEB} = \frac{\text{nombre de bits erronés}}{\text{nombre total de bits émis}}. \tag{4.13}$$

4.4.1 Impact de la largeur de bande en présence d'une compression MPEG

Le but de la compression MPEG n'est pas d'obtenir un signal décodé égal au signal original, mais plutôt que ce signal soit perçu de la même manière que l'original par un auditeur humain. Ce principe est fondé sur propriété de masquage auditif. En effet, certaines composantes du signal sonore, particulièrement les parties hautes fréquences, ne sont pas perçues par l'oreille humaine en présence d'autres signaux. Ces parties masquées du signal sont donc considérées par le codeur comme insignifiantes. De ce fait, lors du procédé d'allocation binaire et sous la contrainte de débit de compression, peu de bits sont alloués à la quantification de ces composantes "presque" inaudibles et l'information qu'elles portent est, par conséquent, en grande partie perdue.

C.Barras [Baras 2005] a étudié dans ses travaux ce phénomène pour le cas d'une compression MPEG 1 layer 3 (MP3) pour deux débits de compression : 96 et 64 kbits/s. Des expériences similaires, réalisées dans le cas d'une compression MP3 et AAC, mettent en évidence ce problème, respectivement illustré par les figure 4.2 et 4.3 où nous présentons une comparaison des DSP du signal audio originel $x(t)$ et de sa version codée/décodée $y(t)$, par le codeur MP3 [3], pour différentes valeurs du débit de compression (96, 64 et 40 kbits/s) et par le codeur AAC [4] pour les débits de compression 64, 40 et 32 kbits/s. Ces DSP sont calculées sur une trame de 1024 échantillons d'un signal de violon échantillonné à 44.1 kHz.

On remarque que les codeurs MP3 et AAC conservent bien les basses fréquences du signal. Cependant, une grande partie des hautes fréquences est éliminée. La largeur de cette bande dépend essentiellement du type de codeur utilisé et du débit de compression appliqué. La compression MPEG effectue donc un filtrage passe-bas. La valeur de la fréquence de coupure f_c ne peut pas être déterminée précisément à partir de ces courbes, d'autant plus qu'elle dépend fortement du débit de compression et du type de codeur

3. Le MP3 a une qualité transparente pour un debit de 96 kbps en mono.
4. Le AAC a une qualité transparente pour un debit de 64 kbps en mono.

canal	sans perturbation	96 kbits/s	64 kbits/s	40 kbits/s
$f_c = fe/2$	$8.56\ 10^{-4}$	$1.77\ 10^{-2}$	$7.48\ 10^{-2}$	$1.65\ 10^{-1}$
$f_c = 11$ kHz	$3.3\ 10^{-4}$	$2.52\ 10^{-4}$	$1.1\ 10^{-3}$	$1.19\ 10^{-2}$
$f_c = 5$ kHz	$4.36\ 10^{-5}$	$2.30\ 10^{-4}$	$2.39\ 10^{-4}$	$4.43\ 10^{-4}$

TABLE 4.1 – Influence de la largeur de bande $[0, f_c]$ sur les TEB du système de tatouage considéré dans le cas d'un canal sans perturbation et avec une compression MP3 à 96, 64 et 40 kbits/s (débit de tatouage 78 bits/s).

canal	sans perturbation	64 kbits/s	40 kbits/s	32 kbits/s
$f_c = fe/2$	$8.56\ 10^{-4}$	$5.9\ 10^{-3}$	$2.26\ 10^{-1}$	$3.89\ 10^{-1}$
$f_c = 11$ kHz	$3.3\ 10^{-4}$	$1.2\ 10^{-3}$	$6.69\ 10^{-2}$	$2.59\ 10^{-1}$
$f_c = 5$ kHz	$4.36\ 10^{-5}$	$5.19\ 10^{-4}$	$5\ 10^{-5}$	$3\ 10^{-3}$

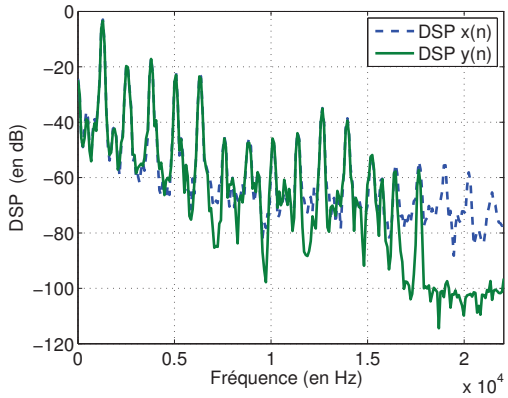
TABLE 4.2 – Influence de la largeur de bande $[0, f_c]$ sur les TEB du système de tatouage considéré dans le cas d'un canal sans perturbation et avec une compression AAC à 96, 64 et 40 kbits/s (débit de tatouage 78 bits/s).

canal	sans perturbation	24 kbits/s	20 kbits/s
$f_c = fe/2$	$8.56\ 10^{-4}$	$3.98\ 10^{-1}$	$4.49\ 10^{-1}$
$f_c = 11$ kHz	$3.3\ 10^{-4}$	$1.07\ 10^{-1}$	$2.17\ 10^{-1}$
$f_c = 5$ kHz	$4.36\ 10^{-5}$	$1.27\ 10^{-2}$	$3.69\ 10^{-2}$
$f_c = 3.5$ kHz	$1.29\ 10^{-5}$	$4\ 10^{-5}$	$1.8\ 10^{-2}$

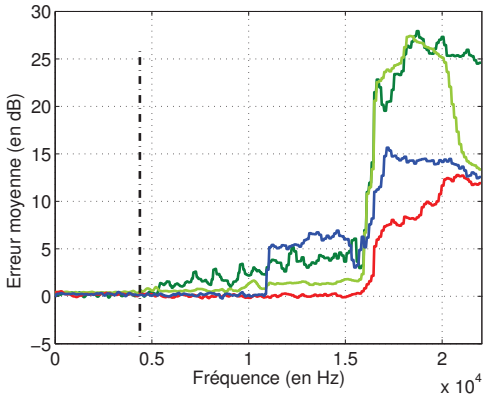
TABLE 4.3 – Influence de la largeur de bande $[0, f_c]$ sur les TEB du système de tatouage considéré dans le cas d'une compression aacPlus à 24 et 20 kbits/s (débit de tatouage 78 bits/s).

utilisé. D'une façon générale, pour les deux types de codeurs considérés et pour des débits de codage supérieurs à 32 kbits/s, la fréquence de coupure est supérieure à 5 kHz. Nottons qu'outre la limitation de la largeur de bande, la quantification des coefficients spectraux faite par la compression MPEG modifie le contenu spectral et est susceptible de modifier le signal de tatouage.

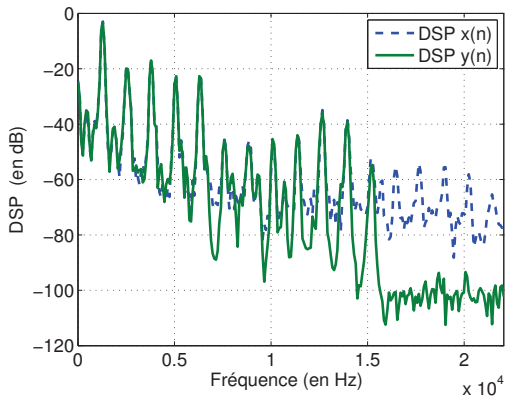
En présence d'une compression MPEG et sous la contrainte de la robustesse, le système de tatouage est assimilé à un canal de transmission opérant dans la bande $[0,\ f_c]$. De ce fait, la largeur de bande du signal tatouage $w(t)$ et du signal modulé $v(t)$ doit être inférieur à f_c. Ce filtrage passe-bas est également introduit en amont du récepteur : il permet de supprimer, du signal reçu, les composantes fréquentielles supérieures à f_c non significatives pour la détection du tatouage. Le schéma du système de tatouage est donc modifié comme présenté sur la figure 4.4.



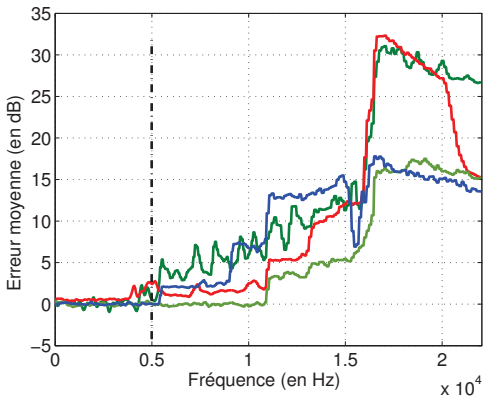
(a) MP3 à 96 kbits/s



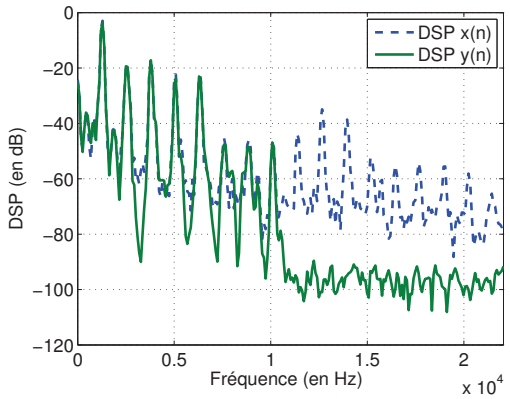
(d) MP3 à 96 kbits/s



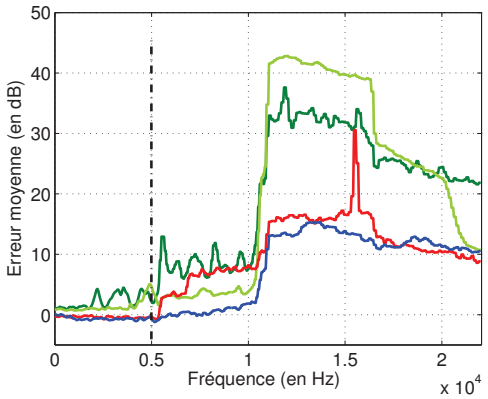
(b) MP3 à 64 kbits/s



(e) MP3 à 64 kbits/s

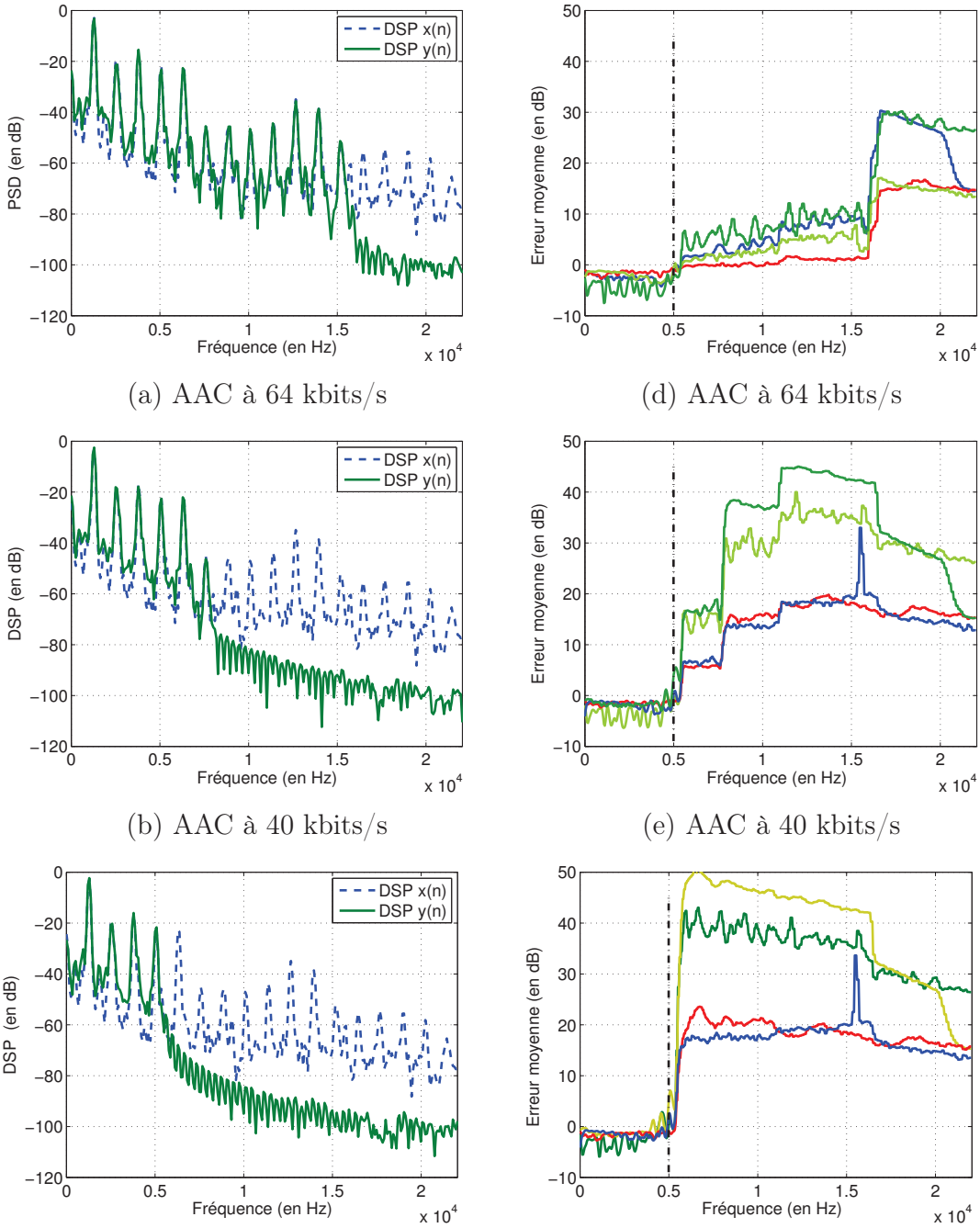


(c) MP3 à 40 kbits/s



(f) MP3 à 40 kbits/s

FIGURE 4.2 – Impact de la compression MP3 sur la DSP des signaux audio et sur les seuils de masquage pour trois débits de compression (96, 64 et 40 kbits/s) : (a), (b) et (c) Comparaison de la DSP d'un signal audio de violon et celle de sa version codée y(t) ; (d), (e) et (f) Erreur d'estimation moyenne EM(f) des seuils de masquage (entre le signal audio et sa version codée) pour quatre signaux tests (pop, violon, castagnettes et tabla).



(a) AAC à 64 kbits/s (d) AAC à 64 kbits/s

(b) AAC à 40 kbits/s (e) AAC à 40 kbits/s

(c) AAC à 32 kbits/s (f) AAC à 32 kbits/s

FIGURE 4.3 – Impact de la compression AAC sur la DSP des signaux audio et sur les seuils de masquage pour trois débits de compression (64, 40 et 32 kbits/s) : (a), (b) et (c) Comparaison de la DSP d'un signal audio de violon et celle de sa version codée y(t) ; (d), (e) et (f) Erreur d'estimation moyenne EM(f) des seuils de masquage (entre le signal audio et sa version codée) pour quatre signaux tests (pop, violon, castagnettes et tabla).

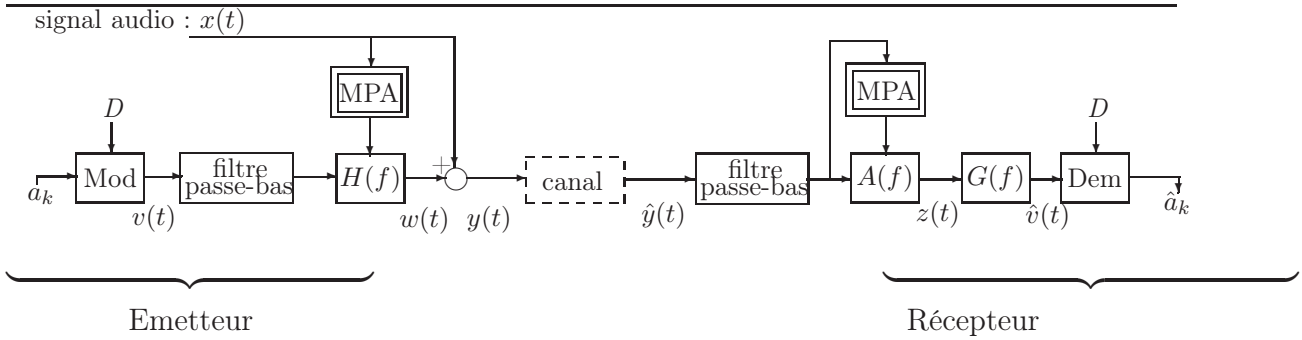
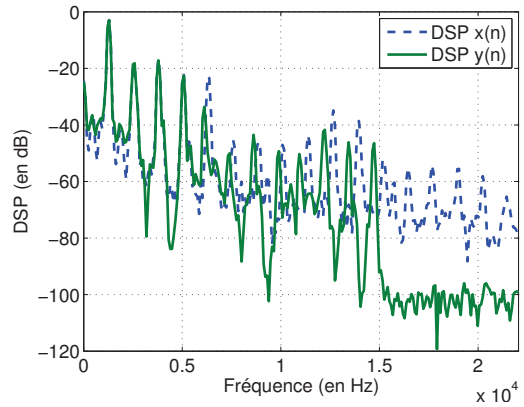
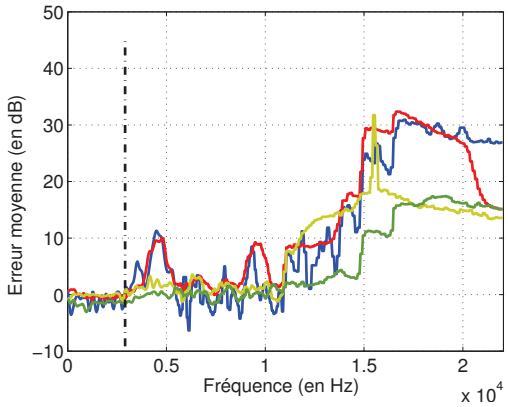


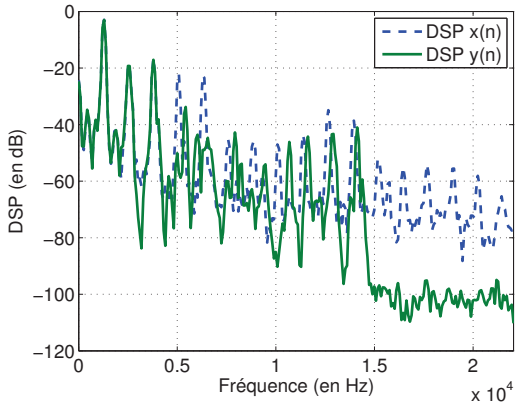
FIGURE 4.4 – Système de tatouage de référence avec filtrage passe-bas.



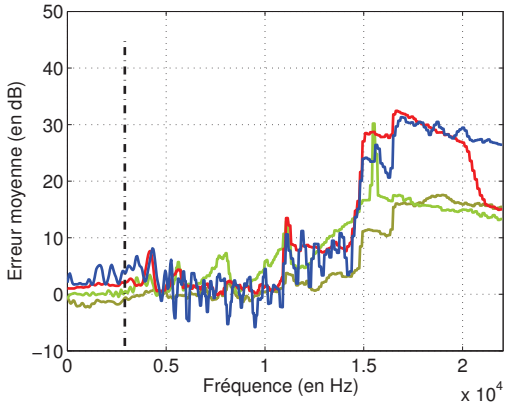
(a) aacPlus à 24 kbits/s



(c) aacPlus à 24 kbits/s



(b) aacPlus à 20 kbits/s



(d) aacPlus à 20 kbits/s

FIGURE 4.5 – Impact de la compression aacPlus sur la DSP des signaux audio et sur les seuils de masquage pour deux débits de compression (24 et 20 kbits/s) : (a) et (b) Comparaison de la DSP d'un signal audio de violon et celle de sa version codée y(t) ; (c) et (d) Erreur d'estimation moyenne EM(f) des seuils de masquage (entre le signal audio et sa version codée) pour quatre signaux tests (pop, violon, castagnettes et tabla).

L'impact de la largeur de bande laisse prévoir une perturbation sur les performances du système de tatouage. Nous nous sommes donc proposé d'évaluer l'influence de cette fréquence de coupure f_c sur les TEB du système de tatouage. Cette évaluation a été réalisée sur 3 signaux tests (violon, pop et castagnette) en utilisant un dictionnaire d'émission constitué de 4 vecteurs (2bits/symbole) étalés dans la bande de fréquence $[0\,;\,f_c]$ et un débit de tatouage fixé à 78 bits/s. Les résultats ont été obtenus dans le cas d'un canal sans perturbation et dans le cas d'une compression MPEG (MP3, AAC) pour différentes valeurs de débit de compression. Ils sont reportés dans les tableaux 4.1 et 4.2. Les résultats confirment que diminuer la fréquence de coupure f_c conduit à réduire la valeur du TEB : cette valeur est inférieur à 10^{-3} pour $f_c = 5$ kHz même en présence de compression. Par la suite, nous travaillerons donc avec une bande limitée à $[0\,;\,5\text{kHz}]$.

Le codeur aacPlus [5] étend d'une manière efficace la partie haute fréquence du signal original même pour des débits de compression très faibles. Ceci est illustré par la figure 4.5 où la largeur de bande du spectre synthétisé s'étale jusqu'à 14.8 kHz pour des débits de compression de 24 et 20 kbits/s [Technologies 2007]. Cependant, la génération de la structure fine de la partie des hautes fréquences par les opérations de translation spectrale, bien que perceptivement acceptable, conduit à une estimation erronée de $H(f)$, ce qui réduit considérablement la robustesse du système de tatouage. Il est donc nécessaire d'utiliser un filtrage passe-bas pour ces codeurs comme pour MP3 et AAC. Nous présentons dans le tableau 4.3 quelques mesures de performance du système de tatouage pour deux débits de compression (24 et 20 kbits/s) en fonction de la fréquence de coupure f_c. Nous constatons que la fréquence de coupure optimale avoisine 3.5 kHz.

Problème de détection en cas des signaux percussifs

La figure 4.6 présente les TEB des deux codeurs considérés en fonction du débit de compression. Le message émis est constitué par une séquence binaire aléatoire de 10^6 bits. Il a été modulé à raison de 2 bits/symbole et inséré dans la bande basse du signal audio [0-5 kHz], avec un débit de 50 bps.

En l'absence de codage/décodage, un TEB de l'ordre de 10^{-5} et 10^{-4} est atteint respectivement pour les signaux non percussifs et les signaux percussifs. Les performances sont très fortement dégradées dès lors que le signal subit une compression MPEG à faible débit. En particulier, pour le cas des signaux percussifs, les valeurs des TEB sont de l'ordre de 10^{-3}. En plus de la dégradation générale de la détection du tatouage déjà mentionnée par Larbi [Larbi 2005a] dans le cas d'une compression MPEG, les signaux percussifs subissent une dégradation supplémentaire, liée au pré-écho induit par la compression lors d'une attaque.

5. Le aacPlus a une qualité transparente à 24 kbps en mono.

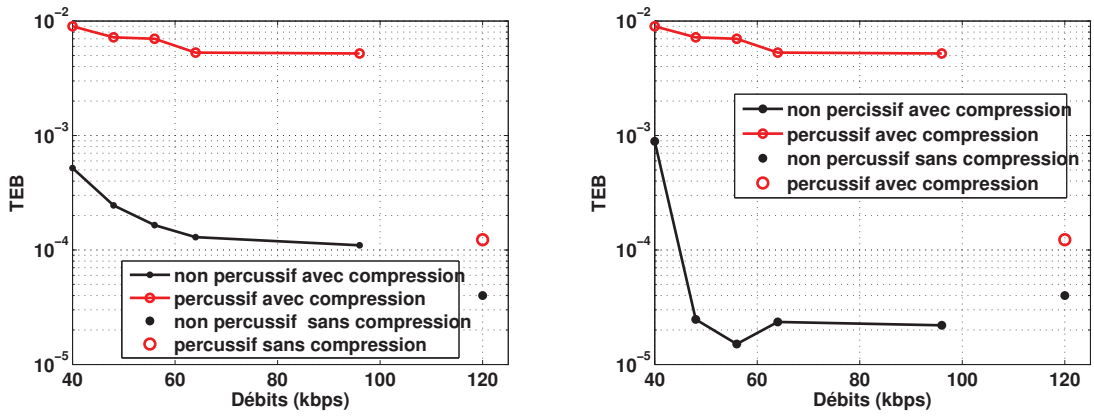


FIGURE 4.6 – Comportement des TEB en fonction du débit de compression : MP3 (à gauche) et AAC (à droite).

4.4.2 Structure améliorée du système de tatouage pour le cas des signaux percussif

Dans le système de tatouage adopté, le signal de tatouage $w(t)$ est étalé sur tout le signal hôte $x(t)$. Pour remédier au problème lié au pré-écho sur les signaux percussifs, nous proposons de ne pas tatouer les trames à attaque lors de l'insertion du signal de tatouage, comme indiqué sur la figure 4.7.

Dans certains cas de figure, et en particulier pour des débits de compression relativement faibles, le phénomène de pré-écho peut non seulement affecter la trame d'analyse en cours, mais aussi celle qui la précède. En tenant compte de la longueur du pré-écho, nous avons a été amenés à éliminer non seulement les trames à attaque mais aussi celles qui les précèdent.

En retranchant les trames à risque d'affectation par le pré-cho, on obtient le signal à tatouer $x(t)^{sa}$. A la sortie du module de tatouage, les trames éliminées sont par la suite remises en place afin d'obtenir le signal tatoué $y(t)$.

Avant le filtre de remise en forme $A(f)$ et le filtre égaliseur $G(f)$, on retranche du signal $\hat{y}(t)$ toutes les trames d'attaque et celles qui les précèdent comme l'indique la figure 4.8.

4.4.3 Analyse des performances du système de tatouage amélioré

Pour avoir des stuctures comparables, les performances du système de tatouage amélioré seront comparées à celle du système de tatouage de référence présenté dans la figure 4.1. Notons que les deux systèmes sont comparés à un même débit "moyen", avec un débit sur les trames tatouées supérieur dans le cas du système amélioré, pour compenser le non-tatouage des trames à attaque et pré-attaque.

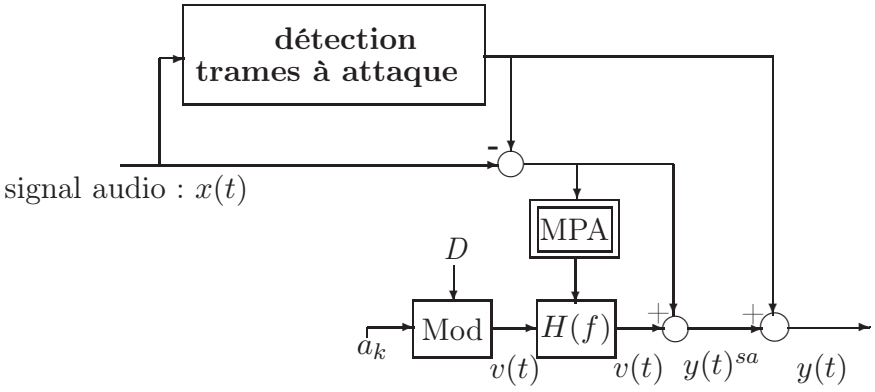


FIGURE 4.7 – Schéma de tatouage proposé côté émetteur.

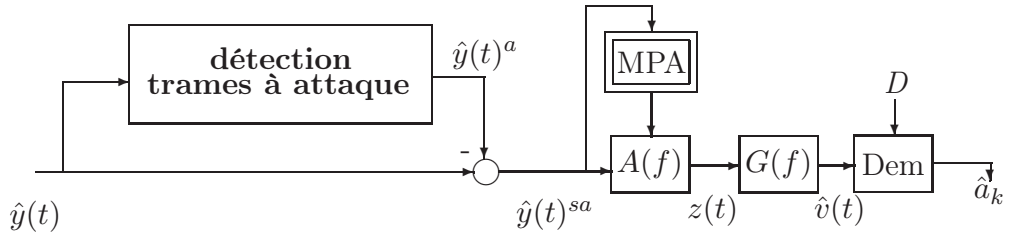


FIGURE 4.8 – Schéma de tatouage proposé côté récepteur.

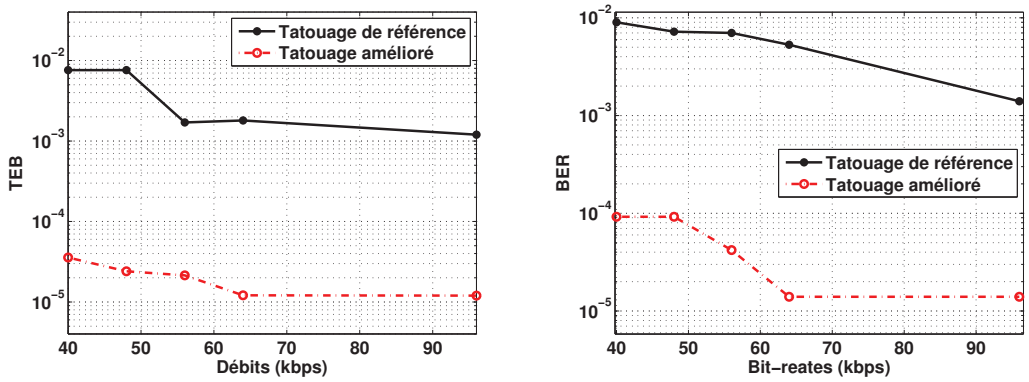
Robustesse à la compression MPEG

La figure 4.9 complète donc cette analyse de la robustesse du système aux perturbations introduites par la compression MPEG pour les signaux percussifs. Les courbes sont réalisées pour les deux types de codeurs (AAC et MP3). La méthode de sélection des trames à tatouer améliore nettement les performances de détection, avec une division par 100 du TEB.

Robustesse à la compression multiple

La figure 4.10 représente le schéma de la détection de tatouage dans le contexte du codage multiple. L'information de tatouage d est insérée uniquement dans le signal audio $x(t)$. A chaque cycle de codage/décodage (C/D), une détection de tatouage, notée WM^{-1} est appliquée sur le signal décodé.

Nous présentons dans la figure 4.11 les performances de la détection en fonction du nombre de codages. Comme illustré par la figure 4.11, les valeurs du TEB augmentent très lentement avec le nombre de codage. En outre, jusqu'à 4 cycles de codage/décodage successifs, le TEB obtenu ne dépasse pas la valeur de 10^{-4}.



(a) (b)

FIGURE 4.9 – TEB moyen en fonction du débit de compression pour un signal de castagnettes et un débit de tatouage moyen de 56 bits/s : (a) compression MP3, (b) compression AAC.

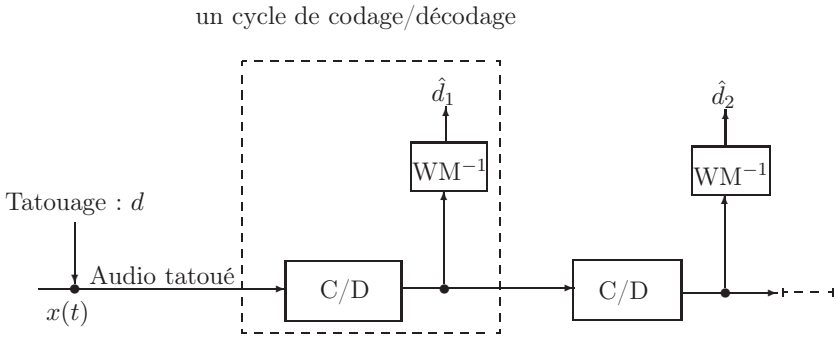


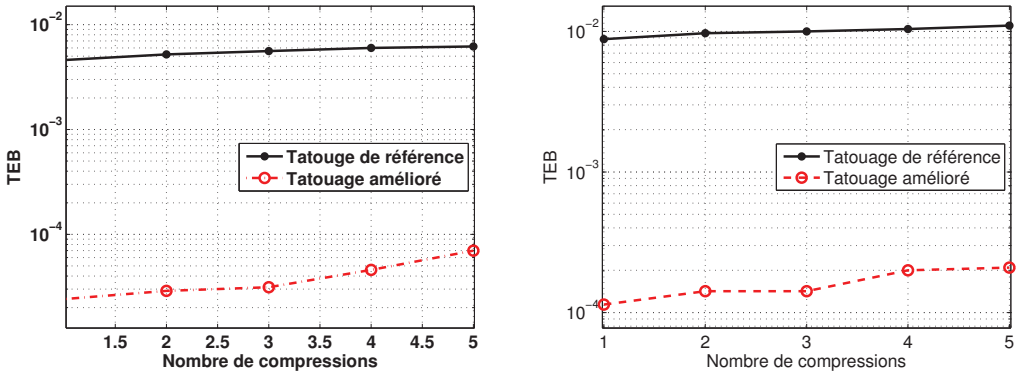
FIGURE 4.10 – Schéma de détection de tatouage dans le contexte de la compression multiple.

4.5 Conclusion

Nous avons montré que le système de tatouage audio proposé par [Larbi] peut être adapté au contexte étudié en tenant compte de deux contraintes liées au codage-décodage MPEG :

- La coupure de la bande haute (MP3 et AAC) ou la regénération approximative des hautes fréquences (aacPlus) nécessite de limiter la largeur de bande des symboles de tatouage.
- Le manque de robustesse du système de référence aux signaux percussifs nous a conduit à ne pas tatouer les trames à attaques ni celles qui les précèdent.

Les modifications proposées améliorent nettement la robustesse du système de tatouage au caractère percussif des signaux. Les performances restent stables en cas de multicodage.



(a) (b)

FIGURE 4.11 – TEB moyen en fonction du nombre de compression pour 3 signaux tests (pop, violon et castagnette) pour un débit de tatouage moyen de 51 bits/s : (a) MP3 à 64 kbits/s, (b) AAC à 48 kbits/s

Dans les chapitres suivants, nous étudierons l'intégration de ce tatouage comme canal auxiliaire dans les systèmes précédents de correction des attaques et de correction des tonales haute-fréquence.

Chapitre 5

Réduction de pré-écho assistée par tatouage audio

Sommaire

Après avoir présenté, au chapitre 2, la solution proposée pour la réduction du pré-écho et de l'amolissement des attaques, nous développons dans ce chapitre la solution complète intégrant l'outil de tatouage audio présenté au capitre 3 comme un support de transmission véhiculant les descripteurs de l'enveloppe temporelle.

Dans la première partie, nous présentons l'architecture complète du système proposé. Cette architecture repose sur un pré-traitement au niveau du codeur amenant à l'estimation des coefficients descripteurs de l'enveloppe temporelle, une transmission par tatouage audio des paramètres calculés et enfin un traitement au niveau du décodeur pour l'extraction des descripteurs et la restauration de l'enveloppe du signal décodé.

Dans la deuxième partie, nous présentons les tests comparatifs entre les solutions existantes implémentées dans les codeurs MPEG MP3 et AAC et les mêmes solutions intégrant la technique proposée.

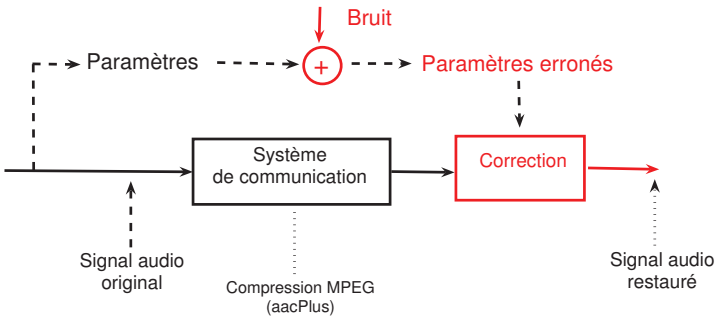


FIGURE 5.1 – Transmission en présence d'un canal binaire à taux d'erreur arbitraire.

5.1 Influence de la détection de tatouage sur le système de réduction de pré-écho

Parmi l'ensemble des perturbations auxquelles le système de réduction de pré-écho proposé se doit d'être robuste, l'erreur de détection de tatouage est sans doute la plus dégradante pour les performances du système. Comme nous l'avons présenté dans le chapitre précédent, le tatouage audio est sensible à la compression MPEG aussi bien simple que multiple.

Cette partie vise à étudier l'influence de l'utilisation du tatouage audio comme un support de transmission dans le contexte de la réduction de pré-écho et de l'amolissment d'attaque. Ainsi, nous présenterons, en premier lieu, l'effet de l'erreur de détection du tatouage sur les performances du système de réduction de pré-écho présenté dans le chapitre 2. Cette étude nous permettera, dans un deuxième lieu, de définir le débit de tatouage maximal (optimal).

5.1.1 TEB minimal en présence de la compression MPEG

Nous supposons que nous disposons d'un canal de transmission auxiliaire à taux d'erreur arbitraire introduisant une dégradation variable permettant d'avoir un TEB variable (voir figure 5.1). Les performances du système proposé sont alors données par la figure 5.2 où nous présentons les variations de l'ODG en fonction du TEB. Les valeurs du TEB varient de 10^{-1} à 10^{-5} dans le cas d'une compression MPEG (codeur MP3) à différents débits de compression (de 40 à 96 kbits/s). Le signal test considéré est une séquence de "castagnettes" échantillonnée à 44.1 kHz.

Les performances du système sont affectées par l'erreur de détection. Elles s'améliorent avec la diminution du TEB et demeurent constantes pour un TEB inférieur à 10^{-2} (pas de changement de la note ODG). Ainsi, pour garantir une bonne restauration des signaux

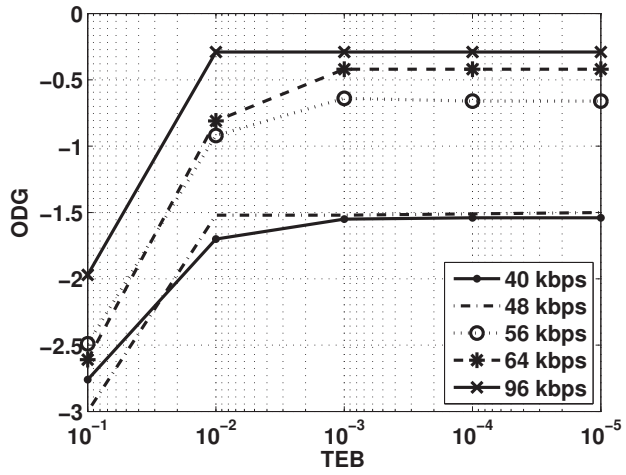
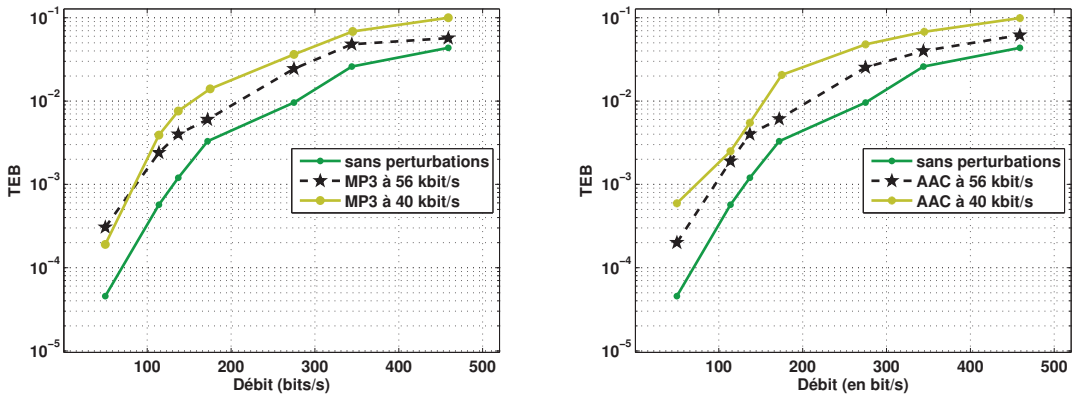


FIGURE 5.2 – Variation de l'ODG en fonction de TEB en présence d'un canal de transmission auxiliaire à taux d'erreur arbitraire.



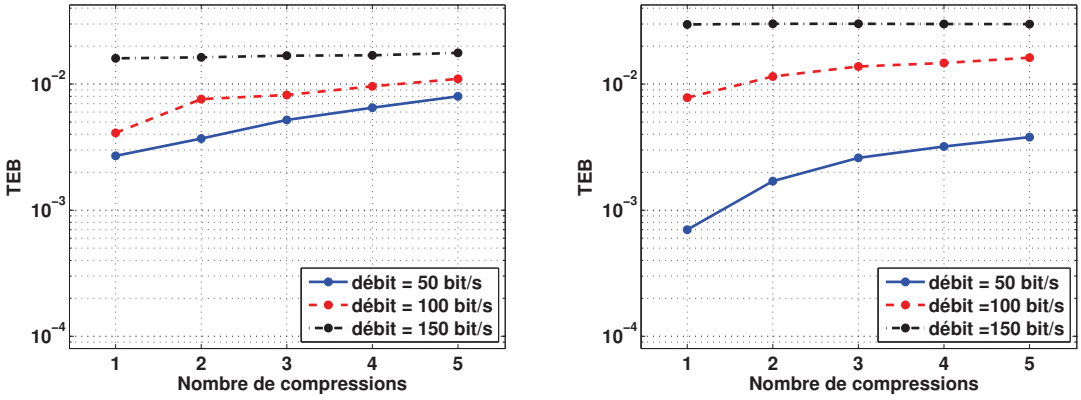
(a) (b)

FIGURE 5.3 – TEB du système de tatouage amélioré en fonction du débit de transmission dans le cas d'un canal sans perturbation et avec compression MPEG pour deux débits de compression 40 et 56 kbits/s : (a) compression MP3, (b) compression AAC.

audio, il est nécessaire d'utiliser un tatouage audio ayant une capacité d'insertion assurant une détection avec un TEB inférieur à 10^{-2}.

5.1.2 Conditions de transmission

Nous nous intéressons ici à l'évaluation de la capacité maximale du tatouage amélioré permettant de garantir le TEB $< 10^{-2}$. Les courbes présentées par la figure 5.3 montrent l'évolution moyenne du TEB en fonction de débit de tatouage. Les résultats reportés correspondent à trois signaux tests tatoués à différents débits et sousmis à une compression



(a) (b)

FIGURE 5.4 – TEB du système de tatouage de référence en fonction du nombre de compression pour 3 débit de tatouage : 50, 100 et 150 bits/s : (a) compression MP3 à 40 kbits/s, (b) compression AAC à 40 kbits/s.

MP3 et AAC.

D'après ces résultats, nous notons qu'un TEB de l'ordre de 10^{-3} est obtenu pour un débit d'insertion de 125 bits/s dans le cas d'un canal de transmission sans perturbations externes, et de 100 bits/s dans le cas d'une transmission en présence d'une compression MPEG à faible débit.

D'autres simulations ont été réalisées dans le but de choisir le débit de tatouage garantissant une robustesse maximale aux perturbations liées à la compression multiple. Les résultats de simulations obtenus sont illustrés par la figure 5.4 qui trace l'évolution du TEB en fonction du nombre de compression-décompression dans le cas des codeurs MP3 et AAC pour les débits de tatouage : 50, 100 et 150 bits/s. L'analyse de ces résultats montre que le système de tatouage est très robuste à la compression MPEG multiple pour des débits d'insertion faibles allant jusqu'à 100 bits/s). Cependant, les performances se dégradent fortement à débit de transmission élevé (supérieur à 100 bits/s).

Les performances du système de réduction de pré-écho dépendent de la précision des paramètres transmis, donc du débit de tatouage, mais aussi du taux d'erreur de détection du tatouage, qui croît avec le débit. Cette contrainte nous conduit à choisir un débit de tatouage de l'ordre de 50 bits/s. Or la correction proposée au chapitre 2 reposant sur un canal auxiliaire avec un débit proche de 500 bits/s. Nous proposons donc d'adapter la correction d'attaques à ce compromis, en ciblant uniquement la correction des trames à attaques affectées par le pré-écho.

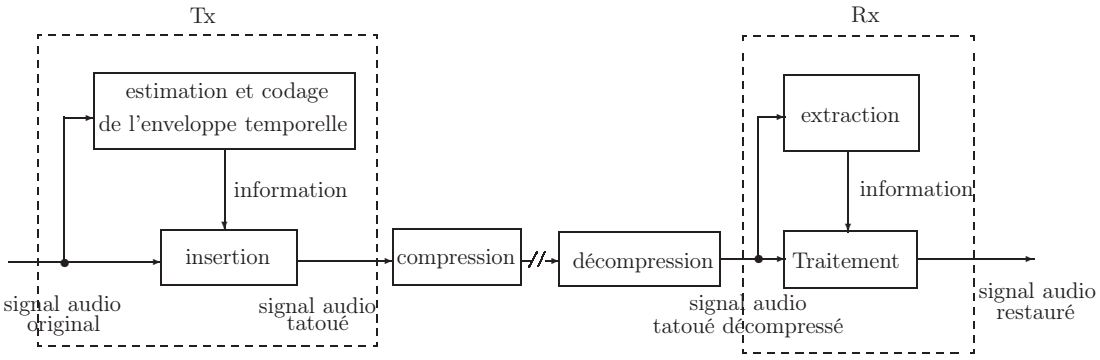


FIGURE 5.5 – Structure générale du système proposé.

5.2 Structure complète du système de réduction de pré-écho assisté par tatouage

La technique complète de réduction de pré-echo et de remise en forme d'attaque repose sur l'architecture présentée par la figure 5.5. Le signal original est injecté dans le module d'estimation d'enveloppe temporelle. Les descripteurs d'enveloppe sont ensuite codés avant d'être transmis, par tatouage audio, à un débit avoisinant les 50 bps. Ce débit nous est imposé par la contrainte liée à l'application du tatouage audio comme support de transmission. Il résulte en effet d'un compromis entre la capacité d'insertion et la robustesse du tatouage à la compression MPEG (MP3 et AAC).

A la réception, le décodeur synthétise le signal tatoué décompressé. Celui-ci est injecté dans le module d'extraction de l'information tatouée. L'information insérée à l'émission sert, en réception, à restaurer le signal décodé.

5.2.1 Traitement au niveau du codeur

Le pré-traitement côté codeur consiste à calculer et coder l'enveloppe temporelle du signal original. Cette étape est réalisée selon le diagramme de fonctionnement illustré par la figure 2.18 du chapitre 2. Pour cela, le signal original, $x(t)$, est segmenté en trames d'analyse de 2048 échantillons. Nous réalisons une détection d'attaque et une localisation de la transition sur la trame d'analyse conformément aux méthodes développées aux paragraphes 2.4.1 et 2.4.2 du chapitre 2. Vu les contraintes de débit imposées par le tatouage audio, le calcul de l'enveloppe temporelle s'effectuera uniquement sur les trames d'attaque. Ainsi,

- pour une trame à attaque, une subdivision de la trame d'analyse en deux sous-trames selon la position de la transition est réalisée. L'instant de la transition est estimé par la méthode algébrique présentée dans le chapitre 2.
- pour la trame qui précéde la trame à attaque, un calcul de son enveloppe temporelle

est également réalisé. En effet, dans certains cas de figure, le bruit généré par le pré-écho peut non seulement affecter la trame à attaque mais aussi celle qui la précède. La trame d'analyse est subdivisée en deux sous-trames de 1024 échantillons. Sur chaque sous-trame, un calcul de l'enveloppe temporelle est réalisé.

Concernant le module de calcul de l'enveloppe temporelle, un modèle ARMA d'ordre AR=5, MA=3 a été retenu. Le choix de cet ordre est fixé selon les contraintes de débit imposées et les performances de ce modèle vues au chapitre 2.

Une fois la prédiction linéaire dans le domaine fréquentiel réalisée, les coefficients ARMA modélisant l'enveloppe temporelle de chaque sous-trame sont convertis en coefficients LSF avant d'être codés. La conversion est réalisée conformément à la méthode présentée dans le paragraphe 2.3.1 du chapitre 2. Pour un ordre ARMA (5,3), le choix s'est porté sur une quantification vectorielle avec un dictionnaire de 1024 vecteurs de 8 LSF en entrée. Le vecteur LSF est ainsi codé sur 10 bits.

Le train binaire ainsi généré est inséré, par tatouage audio (WM) dans tout le signal original $x(t)$. Ainsi, les informations liées à la corrections d'une trame donnée peuvent être avant ou après et plus ou moins éloignées dans le temps. Les principales structures du bloc de tatouage (WM) sont détaillées dans le paragraphe 4.3 de chapitre précédent. Les débits de tatouage considérés sont majoritairement inférieur à 50 bits/s.

5.2.2 Traitement au niveau du décodeur

Comme nous l'avons présenté au chapitre 2, le principe de la correction d'attaque consiste à corriger l'enveloppe temporelle du signal décodé selon le mode de fonctionnement schématisé par la figure 2.19-(b) du chapitre 2.

L'information, insérée par tatouage à l'émission, est extraite par le module WM^{-1}. Cette information représente les indices correspondant aux paramètres LSF quantifiés. Les coefficients ARMA déduits de ces coefficients permettent de remettre en forme l'enveloppe temporelle des trames à attaque du signal original et celles qui les précèdent. La détection des trames à attaque et la localisation de l'instant de la transition sont réalisées directement sur le signal codé/décodé, comme nous l'avons vu au chapitre 2.

Nous avons démontré dans le paragraphe 2.4.3 du chapitre 2 une robustesse du détecteur algébrique au pré-écho. Ainsi, la localisation de la transition se fait directement sur le signal codé/décodé. Cette position, ainsi déterminée, sert à définir la subdivision de la trame à corriger. La trame qui précède une trame à attaque est divisée en deux sous-trame de 1024 échantillons et un calcul d'enveloppe de chaque sous-trame est également réalisé. Sur une sous-trame d'analyse, l'enveloppe temporelle est estimée au décodeur de la manière suivante :

$$\hat{e}(t) = |H(e^{jt})|, \tag{5.1}$$

où $H(z)$ est la fonction de transfert en z définie par :

$$H(z) = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{H_b(z)}{H_a(z)}. \tag{5.2}$$

avec $(a_i)_{1 \leq i \leq p}$ et $(b_i)_{0 \leq i \leq q}$ les coefficients ARMA extraites du tatouage.

Une estimation de l'enveloppe temporelle des trames d'attaque du signal décodé est aussi réalisée. Cette dernière utilise la technique FDLP détaillée dans le paragraphe 2.2. Une fois les enveloppes temporelles des signaux original et décodé calculées, les trames à attaque du signal décodé subissent la correction suivante :

$$\widehat{x}(t) = x_{decodé}(t) \frac{\hat{e}(t)}{\hat{e}(t)_{dec}} \tag{5.3}$$

avec

- $\hat{e}(t)$: estimé de l'enveloppe temporelle de la trame source obtenue par décodage ;
- $\hat{e}(t)_{dec}$: estimé de l'enveloppe temporelle de la trame décodée obtenue par modélisation FDLP afin de garantir une correction de l'enveloppe temprelle du signal décodé au même niveau de détail que celui de l'enveloppe temporelle fournie par le tatouage.

5.3 Evaluation des performances du système complet de réduction de pré-écho assisté par tatouage audio

Cette section s'intéresse aux performances expérimentales du système complet de réduction de pré-écho asssité par tatouage audio sur 6 signaux tests selon le protocole expérimental établi précédemment. Plusieurs objectifs sont visés dans cette étude :

- le premier est de confirmer les performances du système proposé en termes d'amélioration de la qualité audio et de la réduction de pré-écho introduit par la compression MPEG. Le système sera donc testé sur les codeurs MP3 et AAC ;
- le deuxième est d'évaluer l'efficacité du système dans le contexte d'une multiple compression MPEG. Dans ce cas, uniquement le codeur MP3 est considéré. Le choix de ce codeur sera discuté ultérieurement.

L'évaluation du système est établie en utilisant des mesures objectives fournies par le logiciel PEMO-Q. Les signaux tests considérés sont des séquences audio à caractère percussif échantillonnées à 44100 Hz. Ils sont présentés dans le tableau 5.1.

Nous présentons dans la figure 5.6 une illustration de la réduction de pré-écho pour deux signaux percussifs (castagnettes et cymbales). Le codeur MP3 à 64 kbits/s introduit un remarquable pré-écho affectant la transition franche de l'attaque. Après restauration des signaux, le pré-écho est considérablement réduit et l'attaque est restaurée.

Signal	*Durée en s*	*Signal*	*Durée en s*
Castagnettes	3.52	Tabla	2
Cymbales	3.92	Darbouka	1.15
Hihat	1.51	Deff	1.09

TABLE 5.1 – Séquences testées

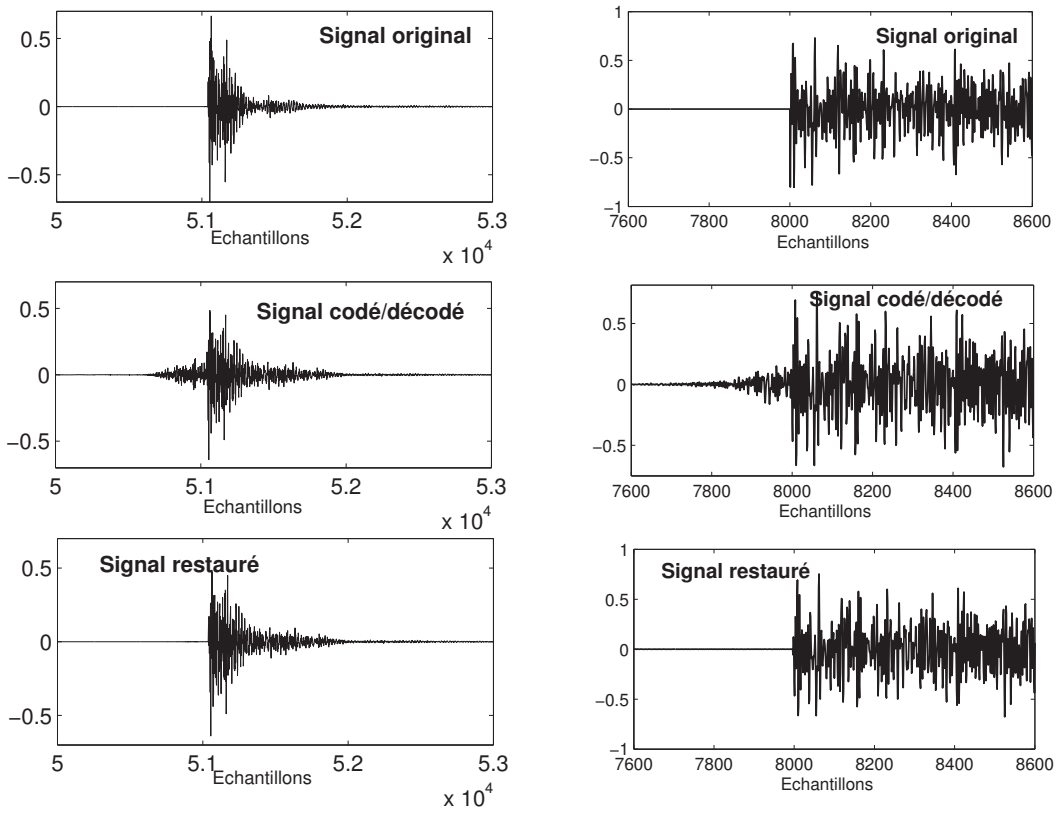


FIGURE 5.6 – Illustration de la réduction de pré-écho par l'approche proposée pour un signal de castagnettes (à gauche) et un signal de cymbales (à droite). Le codeur audio considéré est le codeur MP3 à 64 kbits/s.

5.3.1 Etude des performances dans le cas d'une compression simple

Les performances du système de réduction de pré-écho et l'efficacité de la stratégie d'insertion proposée sont étudiées sur des signaux audio réels et comparées à des systèmes de la littérature : MP3+TM et AAC+TNS à différents débits de compression variant de 40 à 96 kbits/s. Le débit d'insertion devra être inférieur à 100 bits/s pour garantir la robustesse aux perturbations dues à la compression (TEB $< 10^{-2}$). L'ensemble des tests a ainsi été effectué sur les six signaux de musique du tableau 5.1 avec un débit de tatouage

de 50 bits/s.

Nous présentons, sur la figure 5.7, l'évolution de l'ODG en fonction de débit de compression pour le cas du codeur MP3. L'application de l'option de masquage temporel (TM) intégré dans le codeur améliore légèrement la qualité audio. Toutefois, la correction proposée apporte une amélioration significative des valeurs de l'ODG, à l'exception du hihat pour lequel la qualité est comparable à celle donnée par le codeur MP3+TM seul. Les figures mettent en évidence l'efficacité du système proposé pour les deux signaux darbouka et tabla. L'ODG obtenu pour ces deux signaux est nettement supérieur à celle du codeur MP3 seul et du codeur MP3+TM. En effet, un gain de 2 points du score ODG est atteint pour la darbouka, en particulier pour une compression à faible débit. Nottons aussi que pour ces deux types de signaux, la correction proposée à 40 kbits/s offre une qualité équivalente à celle obtenue par le MP3 seul à 96 kbits/s, ce qui correspond à une réduction de débit de l'ordre de 50% .

Dans le cas du codeur AAC, les performances du système proposé sont moindres par rapport à celles obtenues avec le codeur MP3. En effet, comme illustré par la figure 5.8, le système proposé offre une qualité comparable, voire légèrement meilleure que celle obtenu par AAC+TNS seul, à l'exception du signal de Darbouka où la qualité du signal corrigé est transparente pour les débits de compression considérés (ODG $<$ -1). Ces résultats peuvent s'expliquer par le fait que les fenêtres d'analyse dans le cas du AAC sont en mode long (2048 échantillons) alors que celles utilisées dans le cas du MP3 sont en mode court (64 échantillons), et par conséquent, la durée du pré-écho introduit par le codeur AAC est inférieure à celui généré par le codeur MP3 (voir la figure 5.9). Dans ces cas de figure, la perception du pré-écho est fortement réduite par le phénomène de masquage temporel.

5.3.2 Etude des performances dans le cas d'une compression multiple

Dans cette partie, nous évaluons les performances du système proposé dans le cas de multiples compressions MP3 successives. En effet, en se référant aux expériences présentées dans le chapitre 2, une perte de la qualité audio des signaux percussifs est notée après chaque cycle de compression-décompression. Le bruit de quantification s'accumule après chaque cycle de compression de telle manière que le pré-écho, post-masqué par l'attaque lors des premiers codages, n'est plus masqué dans les codages qui suivent et devient de plus en plus dérangeant.

Afin de réduire l'effet des distorsions introduites par la compression multiple, une correction d'attaque est alors proposée après chaque cycle de codage/décodage. Les paramètres de l'enveloppe temporelle sont transmis par tatouage audio à un débit inférieur

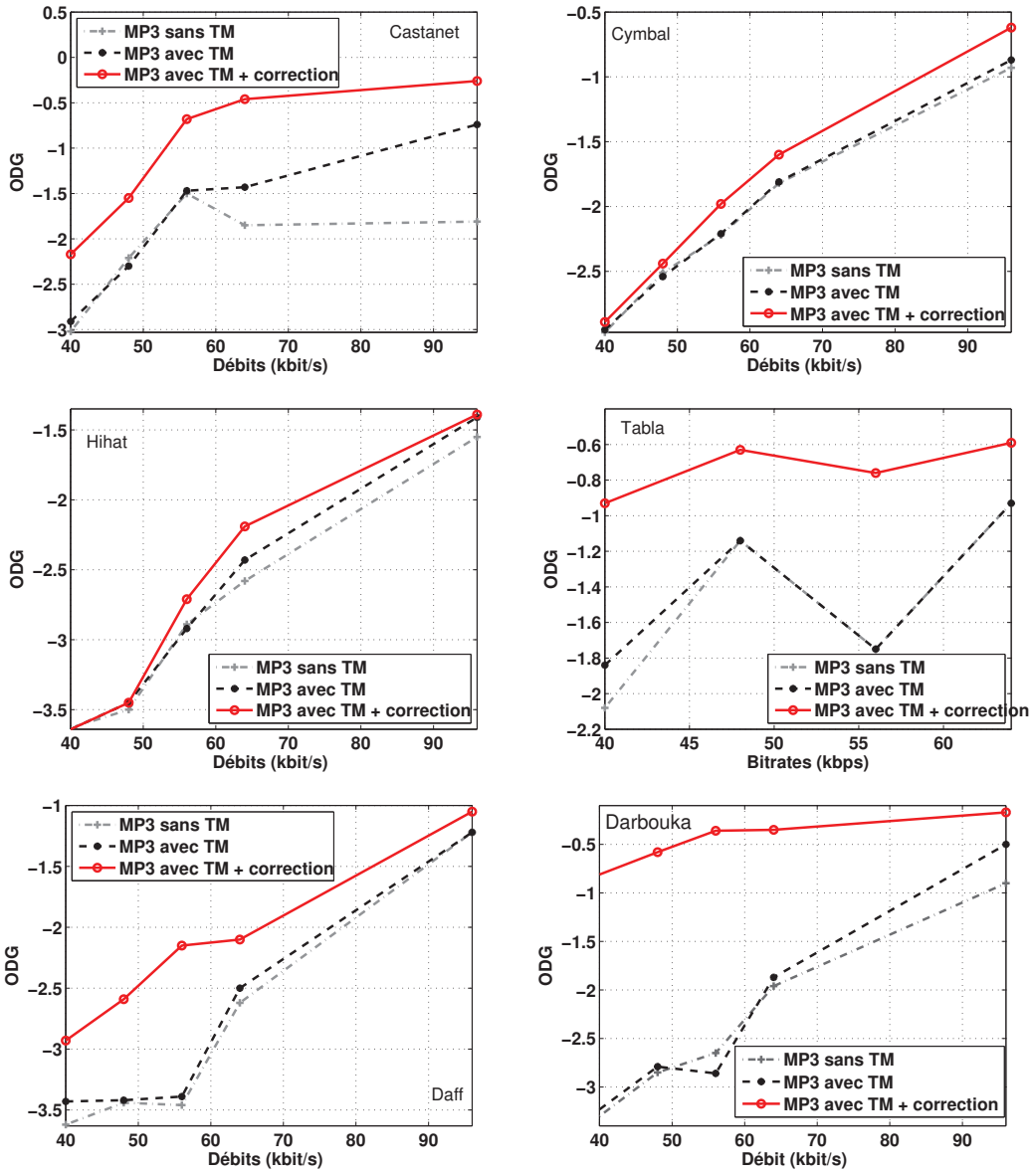


FIGURE 5.7 – Evaluation des performances du système proposé dans le cas du codeur MP3 à débit variant de 40 à 96 kbits/s pour les six signaux tests.

à 100 bits/s assurant une résistance aux multiples opérations de compression.

La figure 5.10 schématise le fonctionnement du système de correction proposé dans le cas de compressions multiples successives. Le signal de tatouage est inséré une seule fois dans le signal original et ceci avant le premier processus de codage. Après chaque cycle de codage/décodage, une extraction et une correction de l'enveloppe temporelle sont effectuées.

Nous représentons dans la figure 5.11 les valeurs de l'ODG en fonction du nombre de compressions MP3 pour un débit = 64 kbits/s. Le débit de transmission est maintenu à 50

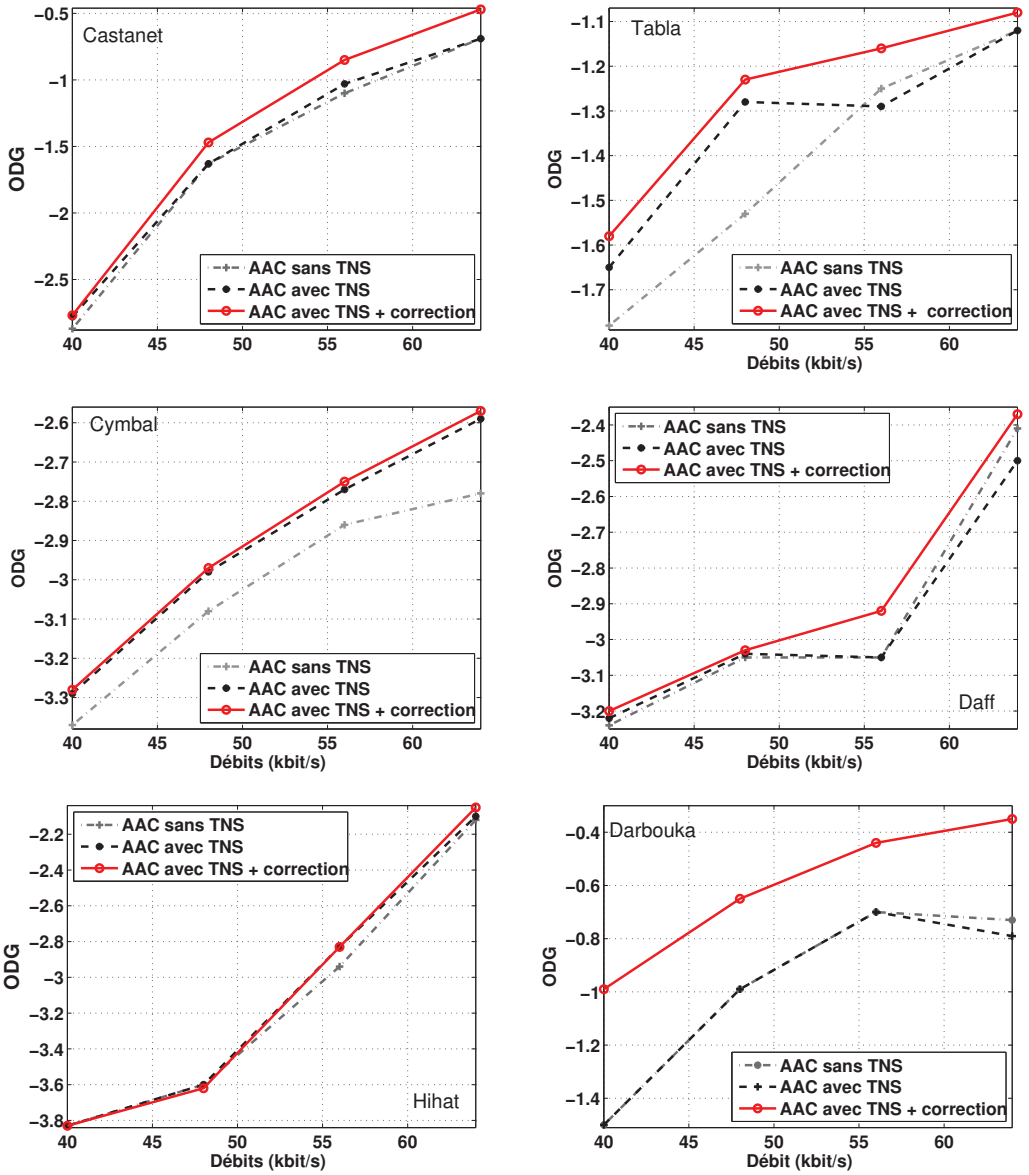


FIGURE 5.8 – Evaluation des performances du système proposé dan le cas du codeur AAC à débit variant de 40 à 96 kbits/s pour les six signaux tests.

bits/s. Les résultats de cette figure montrent que la qualité audio diminue en fonction de l'augmentation du nombre de compressions successives. Notons également que l'option Masquage Temporel (TM) n'améliore presque pas la qualité audio comparé au codeur MP3 seul. Au contraire, la technique de correction proposée associée au codeur MP3+TM améliore considérablement la qualité audio.

L'analyse de ces résultats nous permet de conclure que :

– le MP3 et le MP3+TM offrent une même qualité audio pour les signaux tests ;

– la qualité du MP3+TM est légèrement meilleure à à celle du MP3 seul sur la

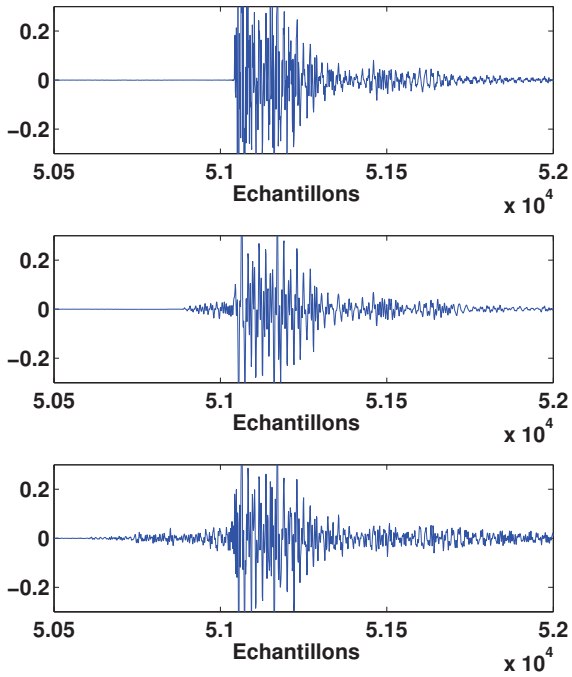
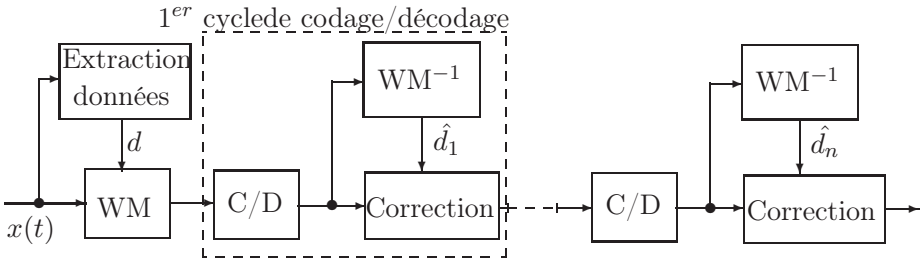


FIGURE 5.9 – Illustration du pré-écho sur un signal de castagnette codé à 40 kbits/s : signal original (en haut), signal codé/décodé par le codeur AAC (au milieu) et signal codé/décodé par le codeur MP3 (en bas).



WM : Watermarking, C : Codage, D : Décodage

FIGURE 5.10 – Diagramme de fonctionnement du système de réduction de pré-écho dans le cas d'une compression multiple.

séquence castagnettes ;

– la configuration MP3+TM+système proposé maintient une bonne qualité jusqu'à 3 codage-décodage successifs pour les séquences castagnettes et darbouka.

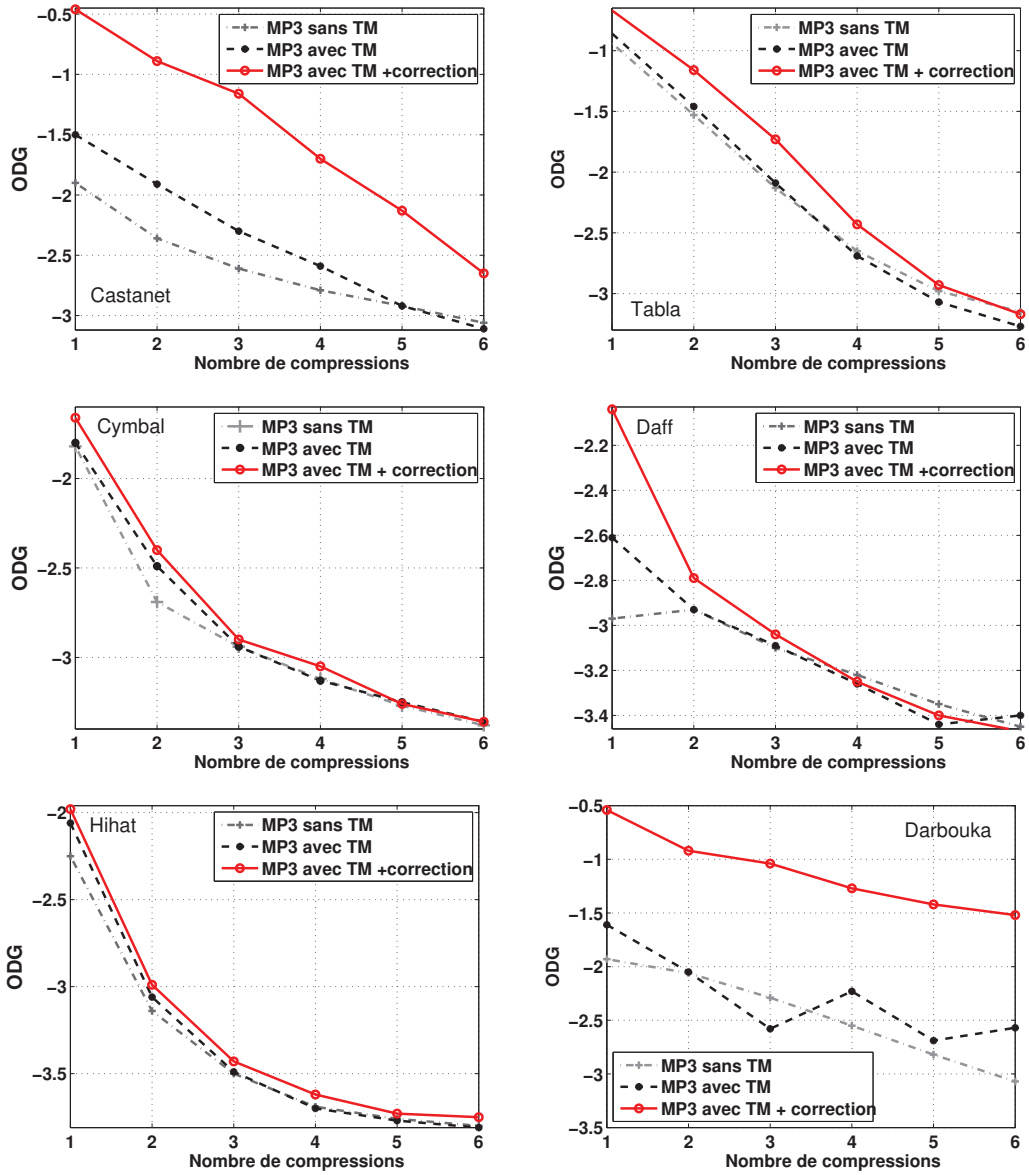


FIGURE 5.11 – Evaluation des performances du système proposé dans le cas de multiples compressions MP3 pour les six signaux tests (débit de compression = 64 kbits/s).

5.4 Conclusion

Nous avons présenté dans ce chapitre le déploiement de la technique complète de réduction de pré-écho et d'amolissement d'attaque des signaux percussifs assistée par tataouage audio. La technique proposée repose sur la correction de l'enveloppe temporelle du signal décodé.

Le tatouage audio constitue un canal "virtuel". En effet, il permet de véhiculer les descripteurs de l'enveloppe moyennant un débit de transmission de l'ordre de 100 bits/s.

Les tests présentés dans ce chapitre ont montré l'intérêt de la technique de correction d'attaque en contexte de compression MPEG bas débit des signaux percussifs. Pour certains types de signaux audio percussifs, la solution proposée associée au codeur MP3 à 64 kbits/s offre une qualité comparable à, voire légèrement meilleure que celle obtenue par le MP3 seul à 96 kbits/s, ce qui correspond à un gain en débit de 50%. Cette qualité reste toujours meilleure dans le cas d'une compression multiple. Pour une qualité équivalente à 2 codages/décodages à 64 kbits/s, notre système permet 3 codages/décodages pour les séquences Castagnettes et Darbouka.

Associé au codeur AAC+TNS, la qualité est proche de voire légèrement supérieure à celle du codeur AAC+TNS seul, à l'exception du signal "Darbouka" où le gain de qualité est plus significatif. De ce fait, une application de cette solution dans le contexte du codage AAC multiple ne laisse pas présager une amélioration de la qualité significative.

Chapitre 6

Correction de tonalité/harmonicité assistée par tatouage audio

Sommaire

6.1 Introduction

Nous avons proposé au chapitre 3 une méthode correction de tonalité et de préservation d'harmonicité dédiée principalement aux codeurs audio à extension de bande, en particulier le codeur HE-AAC. Dans ce chapitre, nous décrivons le système complet intégrant l'outil de tatouage comme mémoire porteuse d'information relative aux positions des tonales. L'architecture du système repose sur une estimation, au niveau du codeur, de l'erreur de position des tonales synthétisées par le codeur SBR, une transmission par tatouage du vecteur des décalages de tonales estimé correspondant et un traitement au niveau du décodeur qui consiste à extraire les informations relatives aux positions des tonales et à corriger la tonalité du signal décodé par translation spectrale.

Dans la première partie de ce chapitre, une étude de la robustesse du système de tatouage en présence de codeur aacPlus sera présentée. Cette étude permettra de définir

les paramètres à adopter garantissant la meilleur performance du système de correction de tonalité proposé. Dans la deuxième partie, nous présentons la structure complète du système proposé intégrant le tatouage audio. L'évaluation des performances du système sera également discutée.

6.2 Contraintes liées à l'utilisation du tatouage comme canal auxiliaire

L'erreur de détection de tatouage constitue l'une des perturbations principales auxquelles le système de correction de tonalité proposé doit être robuste. Dans le chapitre 4, nous avons démontré que le tatouage est très sensible à la compression MPEG (MP3 et AAC) aussi bien simple que multiple. Dans ce chapitre, nous nous intéressons à l'étude des performances de la détection de tatouage dans le contexte du codage à extension de bande (aacPlus) à bas débits (16 et 20 kbits/s).

Dans un premier lieu, nous étudions l'influence de la détection du tatouage sur les performances du système de correction de tonalité. Cette étude permettera en deuxième lieu de définir le débit maximal de tatouage toléré par le système proposé.

6.2.1 TEB minimal en présence d'une compression aacPlus

Dans cette partie, nous supposons qu'on dispose d'un canal de transmission auxiliaire générant des dégradations à taux d'erreur (TEB) variable. Les performances du système proposé sont évaluées par la mesure de la rugosité, présentée dans le paragraphe 3.4.2 du chapitre 3, sur une note pure d'un signal de cornemuse et deux notes d'un signal de trompette échantillonnées à 44.1 kHz. La figure 6.1 présente les variations de la rugosité obtenue par le système proposé en fonction du TEB. Les valeurs du TEB considérées varient de 10^{-1} à 10^{-4} en présence du codeur aacPlus à 16 kbits/s.

Les performances du système proposé sont naturellement dépendantes de l'erreur de détection. Elle s'améliorent avec la diminution du TEB et demeurent constantes à partir d'un TEB égal à 10^{-2}. Le tatouage peut donc être utilisé comme canal de transmission, à condition de choisir une capacité d'insertion assurant une détection avec un TEB inférieur ou égale à 10^{-2}.

6.2.2 Robustesse du tatouage à la compression aacPlus

L'objectif ici est de définir la capacité maximale du tatouage permettant de garantir le TEB $\leq 10^{-2}$ précédement fixé en présence du codeur aacPlus.

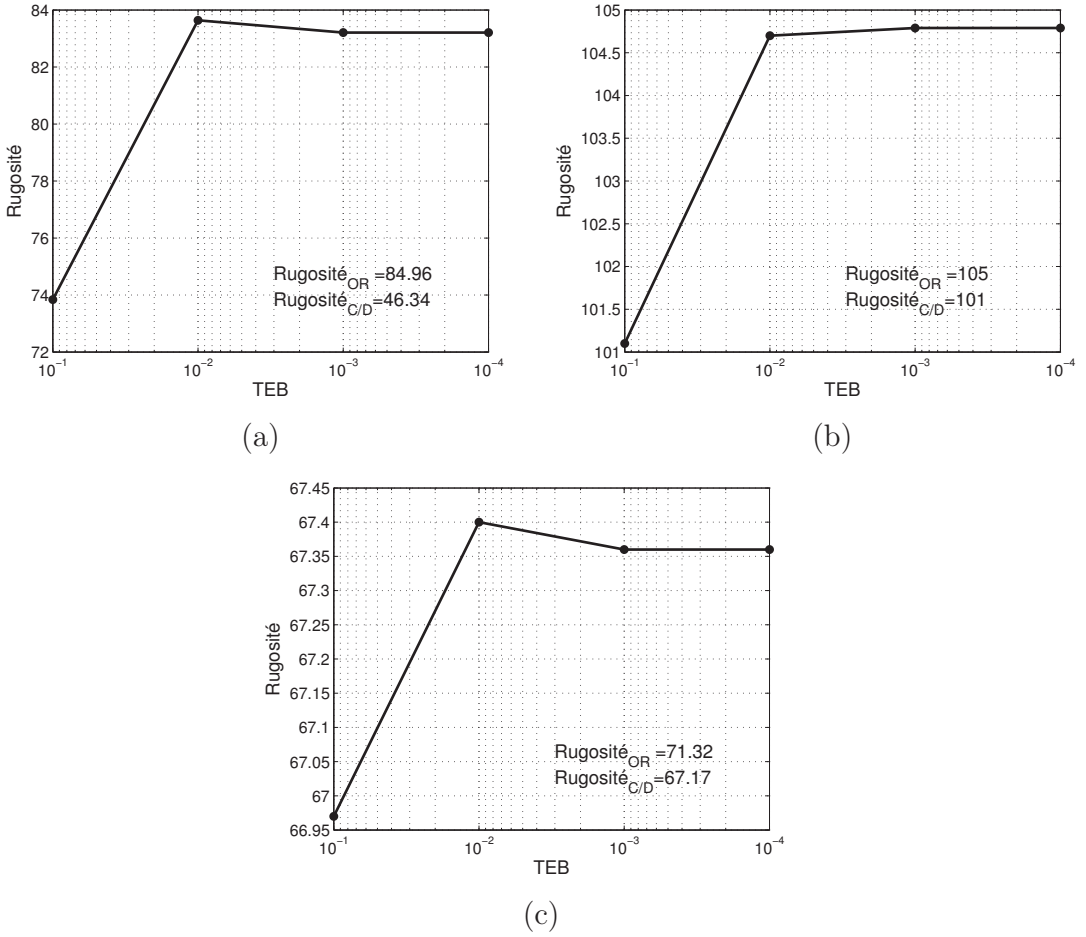
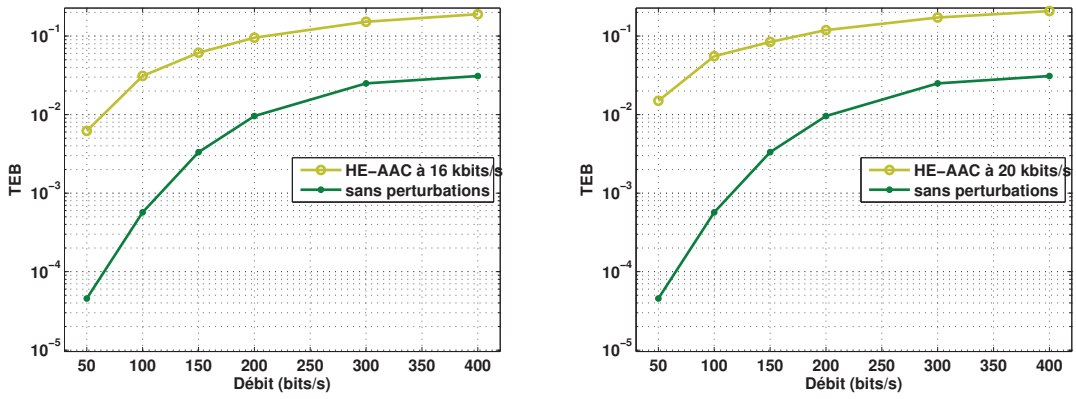


FIGURE 6.1 – Variation de la rugosité en fonction du TEB d'un canal de transmission auxiliaire pour trois signaux tests échantillonnés à 44.1 kHz : (a) cornemuse, (b) trompette$_1$, (c) trompette$_2$.

Comme nous l'avons mentionné au chapitre 4, en présence du codeur aacPlus et sous la contrainte de la robustesse du tatouage, le système de tatouage est assimilé à un support de transmission à bande limitée. Nous avons montré également que la largeur de bande f_c adéquate des vecteurs de dictionnaire $v(t)$ est fixée à 3.5 kHz pour les débits 20 et 24 kbits/s. Pour des débits inférieurs à 20 kbits/s, outre les perturbations générées par le codeur (modification du contenu haute fréquence), le signal décodé est à une fréquence d'échantillonnage différente du signal original (voir tableau 3.1 du chapitre 3). Ce ré-échantillonnage peut perturber les performances du système de tatouage

Nous présentons dans la figure 6.2 les performances du système de tatouage pour les deux cas de perturbations : (a) compression + ré-échantillonnage (16 kbps) et (b) compression seule (20 kbps). Ces tracés présentent l'évolution moyenne du TEB en fonction de débit de tatouage. Les résultats reportés correspondent à deux signaux tests (violon



(a) (b)

FIGURE 6.2 – TEB du système de tatouage de référence en fonction du débit de transmission dans le cas d'un canal sans perturbation et avec compression aacPlus pour deux débits de compression : (a) compression à 16 kbits/s avec ré-échantillonnage, (b) compression à 20 kbits/s sans ré-échantillonnage.

et pop) échantillonnés à 44.1 kHz et tatoués à différents débits et soumis à une compression aacPlus à 16 et 20 kbits/s. Le message émis est constitué par une séquence binaire aléatoire, modulée à raison de 2 bits/symbole et insérée dans la bande basse du signal audio [0, 3.5 kHz]. Notons que pour une compression à 16 kbits/s, les signaux issus du décodeur sont échantillonnés à 32 kHz. Une étape de ré-échantillonnage à la fréquence originale est réalisée avant la détection du tatouage.

D'après ces résultats, le tatouage est robuste au ré-échantillonnage. En effet, pour une débit de 50 bits/s, un TEB de l'ordre de 10^{-2} est obtenu avec une compression à 16 kbits/s+rééchantillonnage ou avec une compression simple à 20 kbit/s. Pour les deux types de perturbations (compression avec et sans ré-échantillonnage), les performances du tatouage se dégradent fortement à débits de transmission élevés (supérieur à 100 bits/s).

Tenant compte des performances limitées du tatouage en présence d'une compression à extension de bande à bas débits, il est nécessaire de réduire au maximum le débit d'information à transmettre par tatouage tout en garantissant une bonne qualité de restauration des signaux audio. Pour cela, nous proposons une nouvelle stratégie de correction qui consiste à effectuer un traitement correctif par blocs de n trames d'harmonicité/tonalité similaire. La classification des trames par bloc est réalisée par une détection de transition.

6.2.3 Détection de transition

Une transition correspond à une variation brusque d'énergie de tout, ou d'une partie du spectre du signal. Dans la majorité des cas, cette variation résulte d'un changement de

note ou apparition d'un instrument. Ceci est illustré par la figure 6.3 (en haut) où nous présentons le spectrogramme d'un signal tonal de glockenspiel. Les changements de notes aux instants 0.29, 0.56 et 0.83 s se traduisent par une variation notable de l'énergie du signal aux mêmes instants (figure 6.3 (en bas)). Nous remarquons également que entre les instants de transitions, le signal est stationnaire. Cette stationnarité est remarquée aussi bien sur le contenu fréquentiel (figure 6.3 (en haut)) que sur les variations temporelles du signal (figure 6.3 (en bas)).

La détection de transition est requise dans la technique de correction de tonalité proposée afin de segmenter le signal en blocs de trames stationnaires comme illustré sur la figure 6.3 où nous distinguons 4 blocs noté de 1 à 4. Le même traitement correctif des tonales sera appliqué au sein d'un bloc ce qui réduit le débit d'informations à transmettre. Ceci suppose que le système de codage/décodage entraîne la même erreur de position des tonales sur toutes les trames du bloc. Dans certain cas, cette hypothèse peut ne pas être vérifiée mais, une amélioration de la qualité audio des signaux restaurés est toujours notée.

La détection des transitions est réalisé conformémement à la technique développée dans le paragraphe 2.4.1 du chapitre 2. Cette technique consiste à suivre l'évolution de l'énergie du signal au cours du temps, par un bloc de sous-trames. La trame d'analyse est segmentée en sous-trames de 128 échantillons. Une trame est jugée transitoire lorsqu'on détecte une différence d'énergie supérieure à un seuil entre une sous-trame de la trame d'analyse et les 8 sous-trames précédentes.

6.3 Architecture complète du système de correction de tonalité/harmonicité

L'architecture complète du système de correction de tonales asssité par tatouage est fournie par la figure 6.4. Le signal audio subit un codage/décodage avant d'être injécté dans le module d'estimation d'erreurs relatives aux positions des tonales synthétisées par le décodeur SBR. Le décalage, correspondant à l'erreur entre les positions de tonales détectées sur le signal original et celles détectées sur le signal codé/décodé, est ensuite codé avant d'être transmis par tatouage audio. Il est évident que le module de codage/décodage utilisé au niveau du codeur augmente la complexité du traitement. Cependant, une réduction notable du débit de tatouage en résulte. Ce compromis entre complexité et capacité est imposé par la contrainte liée à l'utilisation du tatouage comme support de transmission.

A la réception, une détection du tatouage est appliquée sur le signal tatoué distordu, synthétisé par le décodeur aacPlus. Les informations ainsi déterminées servent à restaurer le contenu haute fréquence du signal décodé.

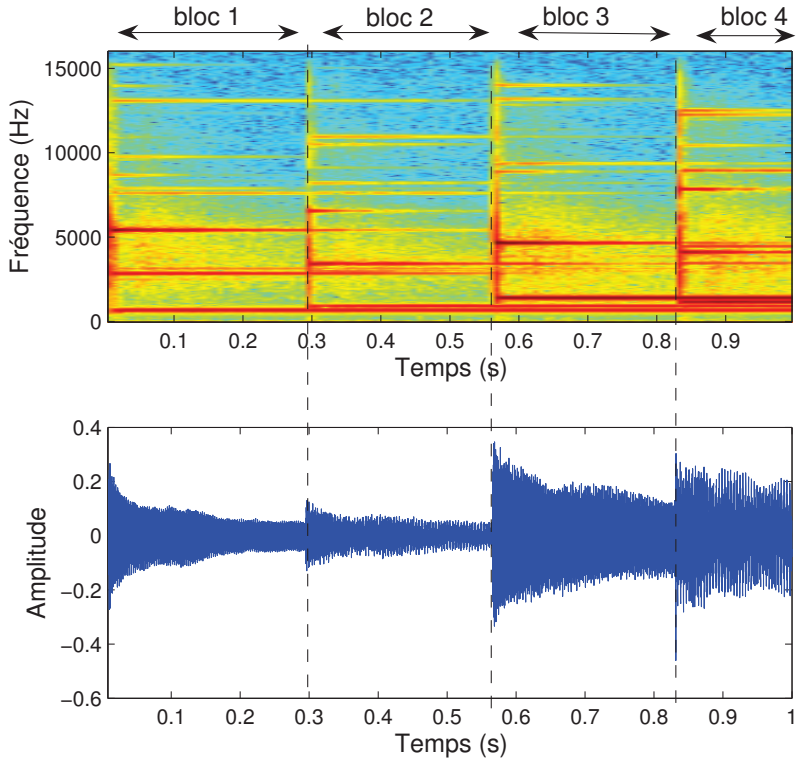


FIGURE 6.3 – Extrait d'un signal de glockenspiel (en bas), Spectrogramme correspondant.

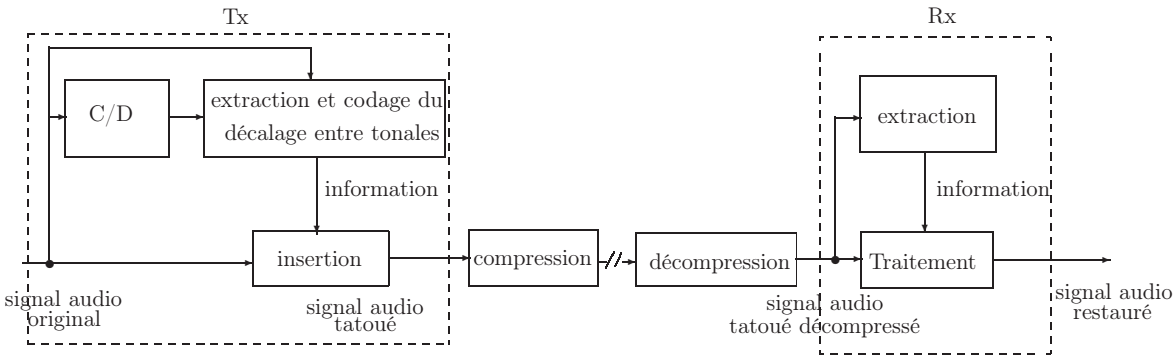


FIGURE 6.4 – Structure générale du système proposé.

Les principaux traitement réalisés dans les deux extrémités du système de transmission seront détaillés dans ce qui suit.

6.3.1 Traitement au niveau du codeur

Le traitement au niveau du codeur consiste à estimer et à coder l'erreur des positions de tonales hautes fréquences générées par le codeur SBR. Les étapes d'estimation d'erreur sont réalisées conformément au diagramme de la figure 3.2 du chapitre 3. Tout d'abord,

le signal original, $x(t)$, subit une opération de compression/décompression avant d'être segmenté en trames d'analyse de 2048 échantillons recouvrantes de 50 % et une décision tonale/non tonale est prise pour chacune de ces trames conformément à la méthode développée dans 3.2.1.2 du chapitre 3. Le signal codé/décodé subit la même segmentation.

Vu les contraintes de débit imposées par l'application du tatouage audio, l'erreur de position des tonales synthétisées est calculée uniquement sur les trames tonales de la manière suivante :

- les trames d'analyse sont regroupées en blocs stationnaires délimités par les trames transitoires (voir figure 6.3). Vu que la stationnarité du signal s'étend sur tout le bloc, l'extraction des paramètres relatifs aux positions des tonales s'effectuera uniquement sur la première trame d'analyse. L'ensemble d'erreurs entre les positions des tonales originales et celles synthétisées est regroupé dans un vecteur décalage noté Δf.
- Les trames transitoires sont des zones particulières qui délimitent les frontières entre deux blocs successifs. Ces trames sont à caractère tonal mais non stationnaire. Par conséquent, aucun traitement n'est associé.

Pout un bloc de n trames, chaque élement du vecteur décalage Δf est quantifié puis codé sur 5 bits. Le message binaire ainsi généré est inséré, par tatouage audio (WM), dans tout le signal original. L'insertion du tatouage est réalisée par la méthode détaillée dans le paragraphe 4.3 du chapitre 4 avec un débit d'insertion variable inférieur à 100 bits/s.

6.3.2 Traitement au niveau du décodeur

Le principe de la restauration de la tonalité du signal repose sur une correction des positions des tonales de la bande haute fréquence synthétisée par le codeur SBR selon le mode de fonctionnement schématisé par la figure 3.9 du chapitre 3.

Le message binaire, inséré par tatouage, est extrait au niveau du décodeur par le module WM^{-1}. L'information extraite représente l'ensemble d'erreurs relatives aux positions des composantes tonales hautes fréquences synthétisées. Le traitement correctif du signal se fait par trames d'analyse de 2048 échantillons recouvrantes de 50 %. Les trames sont regroupées en blocs stationnaires auxquels on applique le même traitement correctif. Les trames transitoires, détectées sur le signal décodé, définissent les frontières entre les blocs.

Rappelons que la correction consiste à :

- subdiviser la trame d'analyse, à l'aide d'un banc de filtres non régulier, en sous-bandes selon les positions de tonales haute fréquence identifiées sur le signal décodé ;
- translater spectralement les sous-bandes ainsi générées selon le décalage Δf transmis par tatouage. Les translations sont réalisées dans le domaine temporel conformément à la technique développée dans le paragaraphe 3.3.2 du chapitre 2 ;

– sommer les signaux des sous-bandes au reste du signal.

6.4 Evaluation des performance du système complet de correction d'harmonicité/tonalité proposé

Dans cette section, on s'intéresse aux performances expérimentales du système complet de correction de tonalité assisté par tatouage audio. Les signaux considérés sont des séquences audio fortement harmoniques échantillonnées à 44.1 kHz et une séquence audio tonale non harmonique échantillonnée à 44.1 kHz. Ils sont présentés dans le tableau 6.1. Le débit de transmission considéré est inférieur à 100 bits/s.

Signal	*Durée en s*	Nature
Trompette_1	6.06	harmonique
Trompette_2	6.34	harmonique
Violon	1.50	harmonique
Cornemuse	1.00	harmonique
Glockenspiel	3.69	Tonale

TABLE 6.1 – Séquences testées

Nous présentons dans la figure 6.5 (a,b et c) une illustration de la correction d'harmonicité sur un extrait fortement harmonique d'un signal de trompette. La compression aacPlus à 16 kbits/s introduit deux ruptures d'harmonicité : la première est à partir de la sixième tonale autour de 4 kHz et qui s'étend jusqu'à 8 kHz et la deuxième est à partir de la quinzième tonale autour de 11 kHz. Une correction notable de l'harmonicité du signal codé/décodé par l'approche proposée est observée sur la version du signal restauré (figure 6.5 (c)). En effet, les composantes tonales repositionnées sont harmoniquement liées.

Sur la figure 6.5(d,e), nous présentons un signal tonal non harmonique de glockenspiel et sa version codée/décodée par le aacPlus à 16 kbits/s. Les rectangles en pointillé indiquent la présence des tonales isolées synthétisées autre que les composantes tonales du signal original. La correction par bloc de trames stationnaires permet d'ajuster les positions des composantes tonales. En effet, on note une correction de la tonalité du signal restauré -6.5 (f)).

L'évaluation objective des signaux harmoniques par la mesure de la rugosité est fournie par le tableau 6.2. On note une amélioration de la qualité audio. En effet, la rugosité mesurée sur les signaux restaurés s'approche plus de la rugosité originelle que celle mesurée sur les signaux codés/décodés (sauf pour le violon), en particulier pour la trompette et la cornemuse.

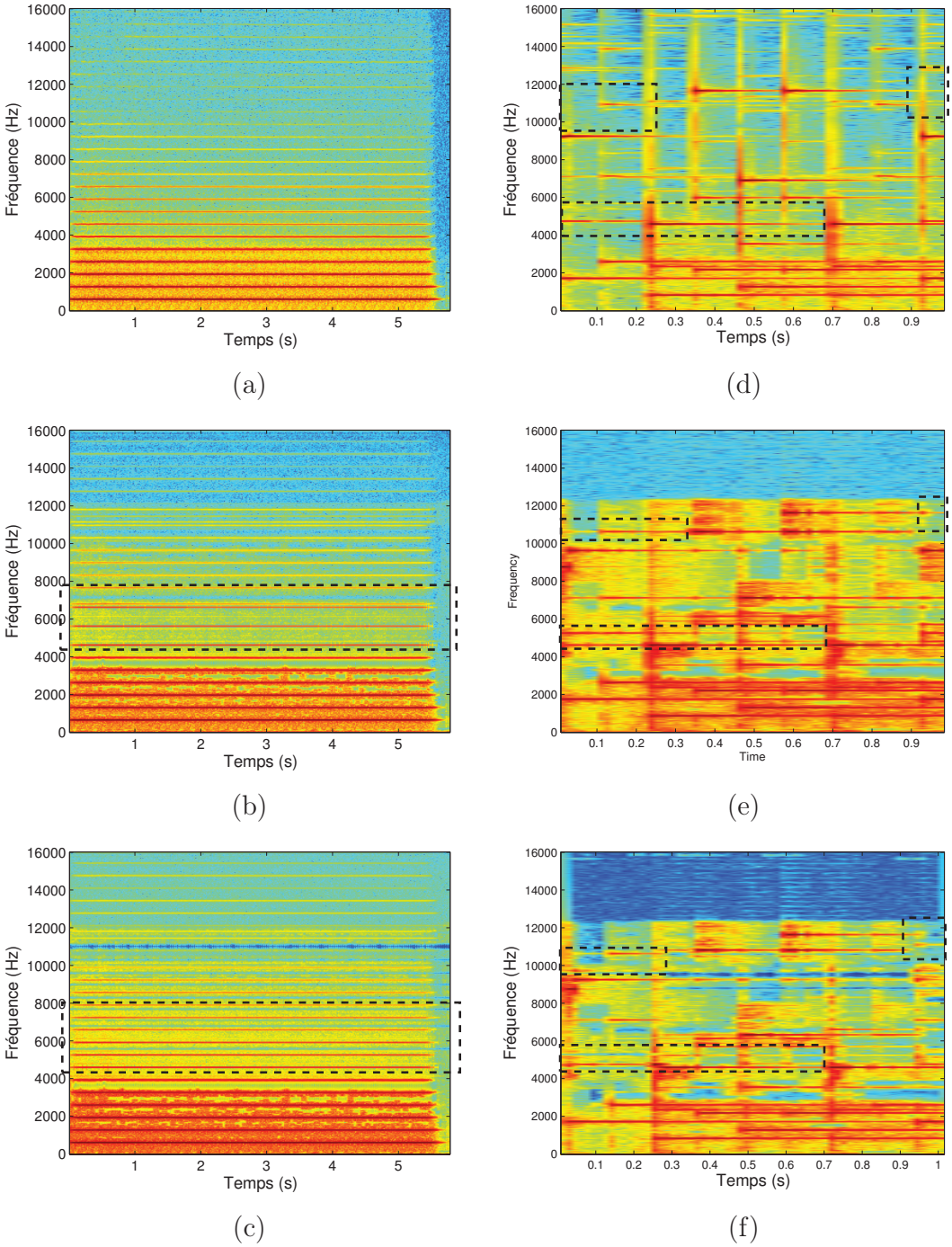


FIGURE 6.5 – Correction de l'harmonicité/tonalité par le système proposé respectivement pour une séquence de trompette et une séquence de glockenspiel : (a, d) spectrogrammes des signaux originaux, (b, e) spectrogrammes des signaux codés/décodés à 16 kbits/s, (c, f) spectrogrammes des signaux restaurés.

Signal audio	Original	codé/décodé	Restauré
Trompette_1	162.19	177.68	174.79
Trompette_2	144	166.7	162
Violon	99.95	96.83	104.64
Cornemuse	90	106.75	103.96

TABLE 6.2 – Evaluation objective de la performance du système de correction d'harmonicité par la mesure de la rugosité. Les signaux tests sont codés/déodés à 16 kbits/s.

6.5 Conclusion

Nous avons développé dans ce chapitre l'intégration de l'outil de tatouage audio dans la technique de correction d'harmonicité/tonalité proposée. La solution présentée consiste à ajuster les positions des tonales en se basant sur l'erreur de position transmise par tatouage moyennant un débit de transmission inférieur ou égal à 100 bits/s.

Les tests présentés dans ce chapitre montrent l'intérêt de la technique de correction des tonales dans le contexte de la compression aacPlus. En effet, pour les signaux fortement harmoniques, La solution proposée associée au codeur aacPlus offre une qualité meilleure que celle obtenue par le aacPlus seul. Pour les signaux tonals, comme le glockenspiel, l'évaluation du système proposé se limite à des observations de spectrogrammes. Des critères d'évaluation appropriés aux signaux audio à caractère tonal non harmonique sont en cours d'études.

Conclusion et perspectives

Nous avons proposé un système d'amélioration de la qualité des signaux audio codés/décodés à bas débits (MP3, AAC, MP3Pro, aacPlus), assisté par tatouage. Le principe est d'exploiter une information inhérente au signal audio, véhiculée par tatouage, pour réduire les distorsions introduites par le codage.

L'étude présentée dans la première partie de ce manuscrit a montré que, sur des signaux audio percussifs, la compression MPEG à bas débit, AAC et MP3, affecte principalement les zones d'attaque. Il en résulte un phénomène de pré-écho, perceptivement gênant, qui s'accentue dans le cas du codage multiple. Nous avons proposé d'y remédier en corrigeant l'enveloppe temporelle du signal après réception. Cette correction exploite la connaissance *a priori* de l'enveloppe temporelle du signal original, supposée transmise par un canal auxiliaire. L'enveloppe est modélisée par la technique FDLP "Frequency Domain Linear Prediction", permettant de bien suivre l'évolution temporelle du signal tout en réduisant le nombre de paramètres qui la décrivent, ce qui limite le débit requis pour le canal auxiliaire.

Associé aux techniques de réduction de pré-écho développées dans la littérature, le système proposé implémenté en amont et en aval des codecs AAC et MP3, offre une remarquable amélioration de la qualité des signaux codés/décodés. Ce gain a été validé par des mesures objectives de la qualité audio perçue.

Nous avons montré que le canal auxiliaire transmettant les paramètres de l'enveloppe temporelle du signal original peut être remplacé par un tatouage audio, qui agit ainsi comme une mémoire porteuse d'informations sur le signal hôte. Il est alors nécessaire de trouver un compromis entre le débit d'informations transmises par tatouage et le taux d'erreur de détection de celui-ci, qui ne doit pas dégrader le fonctionnement du système de correction d'enveloppe. Cette limitation du débit implique de concentrer la correction d'enveloppe sur les seules trames à attaque et leurs voisines.

Les tests d'évaluation du système complet de réduction de pré-écho assistée par tatouage ont montré l'intérêt de la technique proposée en contexte de compression MPEG bas débit, aussi bien pour un codage-décodage simple que dans le cas de multiples compressions successives du même signal. Testée sur des signaux audio percussifs, la solution proposée associée au codeur MP3 offre une qualité meilleure que celle obtenue par le MP3 seul et par le MP3 associé au *temporal masking* (TM). Pour certains types de signaux, à qualité équivalente, on peut gagner 50 % de débit. Associée au codeur AAC+TNS (*Temporal Noise Shaping*), la qualité est proche de, voire légèrement supérieure à, celle obtenue par le codeur AAC+TNS seul.

Les codeurs à extension de bande, en particulier le codeur HE-AAC, introduisent à bas débit un deuxième problème, la rupture d'harmonicité et la non-préservation de tonalité.

Ce phénomène affecte principalement les signaux fortement harmoniques ou à caractère tonal. Nous avons exposé au chapitre 3 une technique de correction de tonalité après décodage. Elle consiste à ajuster une ou un ensemble de composante(s) tonale(s) par de multiples translations spectrales, en exploitant l'erreur de position transmise à bas débit ($\leq$ 100 bit/s) par un canal auxiliaire. La méthode de translation spectrale est fondée sur une modulation à bande latérale unique.

La comparaison du contenu fréquentiel des signaux avant et après restauration montrent une préservation de la tonalité et de l'harmonicité pour la plupart des signaux audio testés. Le système proposé a été évalué *via* une mesure objective de la rugosité perçue. Les résultats montrent une amélioration de la qualité des signaux fortement harmoniques.

Comme pour la réduction de pré-écho, nous avons intégré le tatouage audio dans la technique de correction de tonalité. Il s'agit de transmettre l'erreur des positions des tonales par tatouage, moyennant un débit de transmission inférieur ou égal à 100 bit/s. Les mesures objectives adoptées indiquent une légère amélioration de la qualité des signaux restaurés par rapport au codeur HE-AAC seul.

Perspectives

Au terme de ce travail, plusieurs perspectives relatives aux deux techniques proposées de restauration des signaux audio s'ouvrent.

En ce qui concerne la technique de réduction de pré-écho, les techniques actuelles et futures de tatouage à grande capacité d'insertion [Xiang 2011] ouvrent des possibilités d'amélioration de notre système. D'une part, un débit accru permettrait de corriger toutes les trames. D'autre part, à cause du faible débit de tatouage, les paramètres de l'enveloppe temporelle sont tatoués avant et après l'attaque, de sorte que le son est restitué avec un délai supplémentaire. Ce délai pourrait être réduit grâce un débit supérieur de tatouage.

En ce qui concerne la préservation d'harmonicité et la correction de tonalité, la technique d'estimation des positions des tonales s'est avérée efficiente sur des signaux audio stationnaires et fortement harmoniques (par exemple une note pure de trompette). Cet outil reste toutefois difficile à contrôler sur les signaux audio tonals plus complexes et non stationnaires (par exemple le glockenspiel). En effet, la modification proposée mettant en œuvre la fonction seuil (enveloppe spectrale) se trouve limitée sur ces signaux. D'autres techniques d'estimation de tonale doivent être étudiées.

Les deux corrections assistées par tatouage ont été proposées et validées dans l'esprit d'une preuve de concept : il s'agissait principalement de montrer comment un tatouage audio véhiculant des informations sur le signal hôte peut être utilisé pour corriger les

dégradations subies par ce signal dans un canal de communication, notamment le codage-décodage. Dans ce cadre, nous nous sommes restreints à des signaux simples, mono-instrument. La suite naturelle serait d'adapter les corrections proposées à des séquences audio quelconques, multi-instruments, de manière à concevoir un système de correction générique assisté par tatouage audio.

Bibliographie

[1387 1998] ITU-R Rec. BS. 1387. *Method for objective measurement of perceived audio quality.* 1998. (Cité en page 55.)

[3GPP 2004] 3GPP. *TS 26.403 : specification series, General audio codec audio processing functions ; Enhanced aacPlus general audio codec ; Encoder specification ; Advanced Audio Coding (AAC) part,http ://www.3gpp.org/ftp/Specs/html-info/26403.html (accessed June 2012).* 2004. (Cité en pages 42 et 43.)

[3GPP 2005] 3GPP. *3rd Generation Partnership Project ; Technical Specification Group Services and System Aspects ; General audio codec audio processing functions ; Enhanced aacPlus general audio codec ; Conformance testing (Release 6).* TS 26.406 Release 6 2 V1.0.0, 2005. (Cité en page 77.)

[Aicha 2008] A. Ben Aicha et S. Ben Jebara. *Decorrelation of unput signals for stereophonic acoustic echo cancellation using the class of perceptual equivalence.* Eusipco, 2008. (Cité en page 89.)

[Allen 1977] J. Allen. *Short term spectral analysis, synthesis, and modification by discrete Fourier transform.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 3, pages 235–238, 1977. (Cité en page 24.)

[Ambikairajah 1997] E. Ambikairajah. *Auditory masking and MPEG-1 audio compression.* Electronics and Communication Engineering Journal, 1997. (Cité en page 16.)

[Arnold 2003] M. Arnold, S. Wolthusen et M. Schmucker. *Techniques and applications of digital watermarking and content protection.* Artech House Publishers, 2003. (Cité en page 90.)

[Athineos 2007] M. Athineos et D. P. W. Ellis. *Autoregressive modeling of temporal envelopes.* IEEE transaction on signal processing, vol. 55, no. 11, 2007. (Cité en pages 31 et 32.)

[Bailly 2006] G. Bailly, V. Attina, C. Baras, P. Bas, S. Baudry, D. Beautemps, R. Brun, J. Chassery, F. Davoine, F. Elisei, G. Gibert, L. Girin, D. Grison, J. Léoni, J. Liénard, N. Moreau et P. Nguyen. *ARTUS : synthèse et tatouage audiovisuel des mouvements d'un personnage animé virtuel pour l'accessibilité d'émissions télévisuelles aux téléspectateurs sourds comprenant la Langue Française Parlée Complétée.* Handicap, 2006. (Cité en page 89.)

[Baras 2002] C. Baras. *Etude de la mise en forme de l'information binaire dans un système de tatouage audio.* Mémore de D.E.A, Institut National Polytechnique de Grenoble, 2002. (Cité en page 94.)

[Baras 2005] C. Baras. *Tatouage informé de signaux audio numériques*. PhD thesis of Telecommunication, Ecole Nationale Supérieure des Télécommunications, 2005. (Cité en pages 1 et 95.)

[Cohen 1995] L. Cohen. *Time-Frequency Analysis*. In Proceedings of the 10th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2007, Part II, Englewood Cliffs. NJ : Prentice-Hall, 1995. (Cité en page 31.)

[Cox 2002] I. Cox, M. Miller et J. Bloom. *Digital watermarking*. Morgan Kaufmann Publishers, San Francisco, USA, 2002. (Cité en page 90.)

[Dietz 2002] M. Dietz, L. Liljeryd, K. Kjörling et O. Kunz. *Spectral Band Replication, a novel approach in audio coding*. Audio Engineering Society, 112th Convention, 2002. (Cité en pages xi, 19, 20 et 22.)

[Diniz 2005] F. C. C. B. Diniz et S. L. Netto. *A package tool for general-purpose signal denoising*. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 573–576, 2005. (Cité en page 93.)

[Dolson 1986] M. Dolson. *The Phase Vocoder : A Tutorial*. Computer Music Journal, vol. 10, no. 4, pages 14–27, 1986. (Cité en page 24.)

[Edler 1989] B. Edler. *Coding of audio signal with overlapping block transform and adaptive window function*. Frequenz, vol. 43, pages 252–256, 1989. (Cité en page 16.)

[Ekstrend 2002] P. Ekstrend. *Bandwidth extension of audio signals by spectral band replication*. IEEE Benlux Workshop on Model based Processing and Coding of Audio, 2002. (Cité en page 21.)

[Erne 2002] M. Erne. *Perceptual Audio Coders : What to Listen For, CD-ROM ON CODING ARTIFACT, http ://www.aes.org/publications/technical/AudioCoding.cfm*. Proceedings of the 5th International Conference of Music Retrieval, 2002. (Cité en pages 1 et 10.)

[Fontolliet 1983] P. G. Fontolliet et J. Neirynck. *Traité d'électricité : Systèmes de télécommunications*. In Presses polytechniques romandes, 1983. (Cité en page 74.)

[Gabor 1946] D. Gabor. *Theory of communication*. Journal of IEE, vol. 93, 1946. (Cité en page 31.)

[Geiger 2006] R. Geiger, Y. Yokotani et G. Schuller. *Audio Data Hiding with High Data Rates Based on Intmdct*. ICASSP, 2006. (Cité en page 87.)

[Gilloire 1998] A. Gilloire et V. Turbin. *Using auditory properties to improve the bihaviour of stereophonic acoustic echo cancellation*. ICASSP, pages 3681–3684, 1998. (Cité en pages 1 et 89.)

[Gomes 2002a] L. C. T. Gomes. *Tatouage de signaux audio.* Thèse, Université Paris V, 2002. (Cité en page 94.)

[Gomes 2002b] L. C. T. Gomes, P. Cano, E. Gomèz, M. Bonnet et E. BattleL. *Audio Fingerprinting : Concepts and Applications.* FSKD, 2002. (Cité en page 88.)

[Gomes 2003] L. C. T. Gomes, P. Cano, E. Gomèz, M. Bonnet et E. BattleL. *Audio watermarking and fingerprinting : for which applications ?* Journal of New Music Research, vol. 31, 2003. (Cité en page 88.)

[Hayes 1996] S. Hayes. *Satistical digital signal processing and modeling.* 1996. (Cité en pages 38 et 140.)

[Helmholtz 1954] V. Helmholtz. *On the Sensations of Tone.* Acustica, vol. 30, pages 201–213, 1954. (Cité en page 23.)

[Herre 1999] J. Herre. *Temporal Noise Shaping, Quantization and Coding Methods in Perceptual Audio Coding : A Tutorial Introduction.* AES 17th Conference High Quality Audio Coding, 1999. (Cité en pages 1, 18, 30 et 36.)

[Huber 2006] R. Huber et B.Kollmeier. *PEMO-Q–A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception.* IEEE Transactions on audio, speech and language processing, vol. 14, no. 6, 2006. (Cité en page 55.)

[ISO/IEC 1993] ISO/IEC. *Information technology Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3 : Audio. ISO/IEC 11172-3 :1993.* Joint Technical Committee 1 Subcommittee 29 Working Group 11, 1993. (Cité en page 67.)

[ITU-R 2001] ITU-R. *Method for objective measurements of PErceived Audio Quality (PEAQ).* ITU-R Recommendation BS.1387-1, 2001. (Cité en page 77.)

[Jayant 1984] N. Jayant et P. Noll. *Digital Coding of Waveforms, Englewood Cliffs, NJ, Prentice- Hall.* 1984. (Cité en page 18.)

[Johnston 1988] J. D. Johnston. *Transform coding of audio signals using perceptual noise criteria.* IEEE Jour. Selected Areas Commun, vol. 6, pages 314–323, 1988. (Cité en page 66.)

[Kabal 1986] P. Kabal et R. P. Ramachandran. *The Computation of Line Spectral Frequencies Using Chebyshev Polynomials.* IEEE Transaction on acoustics, speech and signal processing, vol. 34, pages 1419–1426, 1986. (Cité en page 141.)

[Kirovski 2003] D. Kirovski et H. Malvar. *Spread-Spectrum Watermarking of Audio Signals.* IEEE Transactions on Signal Processing, vol. 5, no. 4, 2003. (Cité en page 90.)

[Koukopoulos 2001] D. K. Koukopoulos et Y.C. Stamatiou. *A compressed-domain watermarking algorithm for mpeg audio layer 3.* Proceedings of the workshop on Multimedia and security, 2001. (Cité en page 87.)

[Larbi 2005a] S. Larbi. *Structures d'égalisation en tatouage audionumérique*. Thèse, École nationale supérieure des télécommunications, 2005. (Cité en pages 90, 92, 94 et 100.)

[Larbi 2005b] S. Larbi et M. Jaidane. *Audio Watermarking : A Way To Stationarize Audio Signals*. IEEE Trans. Signal Processing, vol. 53, no. 2, pages 816–823, 2005. (Cité en pages xii, 1, 44, 45, 88 et 89.)

[Laurent 1998] H. Laurent et C. Doncarli. *Stationarity index for abrupt changes detection in the time frequency plane*. IEEE Signal Processing Letters, vol. 5, no. 2, pages 43 –45, 1998. (Cité en page 44.)

[Linde 1980] Y. Linde, A. Buzo et R. M Gray. *An algorithm for vector quantizer design*. IEEE Trans on Communication, vol. 28, pages 84–95, 1980. (Cité en page 38.)

[Liu 2007] Y.W. Liu. *Sound source separation assisted by audio watermarking*. IEEE Int. Conf. Multimedia and Expo, pages 200–203, 2007. (Cité en page 89.)

[Marple 1999] L. S. Marple. *Computing the discrete-time 'analytic' signal via FFT*. IEEE Trans. Signal Process., vol. 47, 1999. (Cité en pages 31 et 33.)

[Martucci 1994] S. Martucci. *Symmetric convolution and the discrete sine and cosine transforms*. IEEE Transactions on Signal Processing, vol. 42, 1994. (Cité en page 33.)

[Mboup 2008] M. Mboup, C. Join et M. Fliess. *A delay estimation approach to change-point detection*. ICASSP, 2008. (Cité en pages 47 et 49.)

[Moreau 1995] N. Moreau. *Techniques de compression des signaux*. Masson, 1995. (Cité en pages 13, 16 et 150.)

[Nagel 2009] F. Nagel et S. Disch. *A harmonic bandwidth extension method for audio codecs*. Eusipco, 2009. (Cité en pages xi, 2, 24, 25 et 27.)

[Nagel 2010] F. Nagel, S. Disch et S. Wilde. *A continuous modulated single sideband bandwidth extension*. ICASSP, 2010. (Cité en pages 2 et 26.)

[Noll 2000] P. Noll. *MPEG Digital Audio Coding Standards*. CRC Press LLC. <http ://www.engnetbase.com>., 2000. (Cité en page 16.)

[Oppenheim 1999] A. V. Oppenheim, R. W. Schafer et J. R. Buck. *Discrete-Time Signal Processing*. In Proceedings of the 10th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2007, Part II, 2nd ed. Englewood Cliffs. NJ : Prentice-Hall, 1999. (Cité en page 31.)

[Papoulis 1985] A. Papoulis. *Levinson's Algorithm, Wold's Decomposition and Spectral Estimation*. SIAM Review, vol. 27, no. 3, pages 405–441, 1985. (Cité en page 35.)

[Pickett 1959] J. M. Pickett. *Backward Masking*. Journal of the Acoustical Society of America, vol. 31, pages 1613–1615, 1959. (Cité en page 11.)

[Plomb 1965] A. Plomb et W. J. M. Levelt. *Tonal consonnace and citical bandwidth.* Journal of the Acoustical Society of America, 1965. (Cité en pages 23 et 77.)

[Proakis 2001] J. Proakis. *Digital communications.* McGraw-Hill, New York, USA, fourth edition, 2001. (Cité en page 94.)

[Reiss 2004] J. Reiss et M. Sandler. *Audio Issues In MIR Evaluation.* The International Society for Music Information Retrieval, 2004. (Cité en pages 1, 10 et 54.)

[Sadok 2006] M. Sadok, M. Jaidane et M. Walz. *Joint Time-frequency Domain Reflectometry and Stationarity Index For Wire Diagnotics in Aircraft.* 9th Joint FAA/DoD/NASA Aging Aircraft Conference, Atlanta, USA, 2006. (Cité en page 44.)

[Sagi 2007] A. Sagi et D. Malah. *Bandwidth extension of telephone speech aided by data embedding.* EURASIP Journal on Applied Signal Processing, vol. 2007, 2007. (Cité en page 90.)

[Technologies 2007] Coding Technologies. *aacPlus v^2 TM Evaluation Package v8.0.3, Revision : 1 :31.* In User Manual, 2007. (Cité en pages xvii, 62 et 100.)

[Terhardt 1974] E. Terhardt. *On the perception of periodic sound fluctuations (roughness).* Acustica, vol. 30, pages 201–213, 1974. (Cité en page 79.)

[Vassilakis 2001] P. N. Vassilakis. *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance.* Doctoral Dissertation, Los Angeles : University of California, 2001. (Cité en pages 77 et 79.)

[Vercellesi 2007] G. Vercellesi, A. Vitali et M. Zerbini. *MP3 audio quality for single and multiple encoding.* ICME, 2007. (Cité en page 1.)

[Vitali 2011] A. Vitali, G. Vercellesi et M. Zerbini. *A multi-level approach to direct process MP3 codes in compression domain.* ST Journal of research, Multimedia Stream Technologies, vol. 3, no. 2, 2011. (Cité en page 1.)

[Wolters 2003] M. Wolters, K. Kjörling, D. Homm et H. Purnhagen. *A close look into MPEG-4 High Efficiency AAC.* Audio Engeneering Society, 2003. (Cité en page 3.)

[Xiang 2011] Y. Xiang, D. Peng, N. Wanlei et I. Zhou. *Effective Pseudonoise Sequence and Decoding Function for Imperceptibility and Robustness Enhancement in Time-Spread Echo-Based Audio Watermarking.* IEEE Transactions on Multimedia, vol. 13, no. 6, 2011. (Cité en pages 90 et 130.)

Annexe A

Calcul et représentation des paramètres du prédicteur

A.1 Calcul des paramètres ARMA

Sur le système réprésenté sur la figure A.1, le signal de sortie x_n s'écrit comme une combinaison linéaire des échantillons du signal de sortie observés aux p instants précédents, et des échantillons du signal d'entrée u_n observés à l'instant présent et aux q instants précédents (Modèle ARMA, AutoRegressive Moving Average).

$$x_n = \sum_{l=0}^{q} b_l u_{n-l} - \sum_{k=1}^{p} a_k x_{n-k}, \tag{A.1}$$

où $\{(b_l)_{1\leq l\leq q}; (a_k)_{0\leq k\leq p}\}$ représentent les coefficients du filtre h. La fonction de transfert du système décrit par l'équation A.1 s'écrit :

$$H(z) = \frac{X(z)}{U(z)} = \frac{\sum_{l=0}^{q} b_l z^{-l}}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{A.2}$$

Les racines du numérateur et du dénominateur sont respectivement les zéros et les pôles du modèle.

Différentes méthodes ont été proposées dans la littérature pour estimer les coefficients de prédiction a_k et b_k. La méthode adoptée ici est celle de Prony. Cette méthode repose sur l'hypothèse que l'entrée u_n est une impulsion de Dirac. On aura donc :

$$X(Z) = H(Z) = \frac{B(z)}{A(z)} \tag{A.3}$$

On cherche à avoir une erreur d'estimation nulle dans l'intervalle $[0,\ q]$ tout en minimisant l'erreur d'estimation au-delà de cette fenêtre. Comme $b(n)$ est de longueur q,

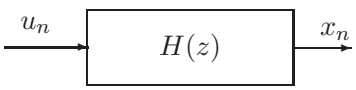


Figure A.1 – Système entrée/sortie

idéalement le produit $a(n) \times x(n)$ doit être nul pour $n > q$. Ceci se traduit par la relation suivante :

$$x(n) + \sum_{k=1}^{p} a_p(k)x(n-k) = \begin{cases} b_q(n) & n = 1, ...q \\ e(n) & n > q \end{cases} \tag{A.4}$$

Le calcul des paramètres $\{a_k\}$ repose sur la minimisation de l'energie du signal d'erreur $E_{p,q}$ définie par :

$$E_{p,q} = \sum_{k=q+1}^{+\infty} e(n)^2 = \sum_{k=q+1}^{+\infty} (x(n) + \sum_{k=1}^{p} a_p(k)x(n-k))^2, \quad n > q. \tag{A.5}$$

Ceci se traduit par l'annultation de la dérivation de $E_{p,q}$ par rapport aux coefficients $\{a_k\}$:

$$\frac{\partial E_{p,q}}{\partial a_k} = 0, \quad \forall\ k = 1, ..., p, \tag{A.6}$$

Or

$$\frac{\partial E_{p,q}}{\partial a_k} = \sum_{n=q+1}^{+\infty} 2e(n)x(n-k). \tag{A.7}$$

Ainsi, on obtient les p équations suivantes :

$$\sum_{k=1}^{p} a_p(k) \sum_{n=q+1}^{+\infty} x(n-i)x(n-k) = -\sum_{n=q+1}^{+\infty} x(n)x(n-i), \qquad 1 \leq i \leq p. \tag{A.8}$$

En définissant la fonction d'autocorrélation du signal x :

$$r_x(k,l) = \sum_{n=q+1}^{+\infty} x(n-l)x(n-k), \qquad 1 \leq k \leq p \tag{A.9}$$

et en substituant les équations A.8 et A.6, on obtient les p équations suivantes :

$$\begin{pmatrix} r_x(1,l) & r_x(1,2) & ... & r_x(1,p) \\ r_x(2,l) & r_x(2,2) & ... & r_x(2,p) \\ ... & ... & ... & ... \\ r_x(p,1) & r_x(p,2) & ... & r_x(p,p) \end{pmatrix} \begin{pmatrix} a_p(1) \\ a_p(2) \\ ... \\ a_p(p) \end{pmatrix} = - \begin{pmatrix} r_x(1,0) \\ r_x(2,0) \\ ... \\ r_x(p,0) \end{pmatrix} \tag{A.10}$$

Ce système matriciel se résout en tenant compte du fait que la matrice d'autocorrélation est une matrice de Toeplitz. Cette propriété permet de résoudre efficacement le système, c'est-à-dire sans inversion de la matrice r, par l'algorithme de Levinson et Durbin[Hayes 1996].

Une fois les coeffients a_k déterminés, une estimation des coefficients b_k est réalisée par la résolution de l'équation :

$$b_q(n) = x(n) + \sum_{k=1}^{p} a_p(k)x(n-k) \quad \forall\ n = 1, 2...q \tag{A.11}$$

A.2 Conversion AR en LSF et inversement

Le point de départ du calcul des coefficients LSF est la fonction de transfert du filtre prédicteur définie par :

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}, \tag{A.12}$$

où $\{a(k)_{0 \leq k \leq p}\}$ représente les coefficients du filtre h d'ordre p.

A partir du polynôme de l'équation , on construit un polynôme symétrique F_l et un polynôme antisymétrique (conjugué) F_2 [Kabal 1986]. Ils sont donnés par :

$$F_1(z) = A(z) + z^{-(P+1)} A(z^{-1}) \tag{A.13}$$

$$F_2(z) = A(z) - z^{-(P+1)} A(z^{-1}) \tag{A.14}$$

Ces deux polynômes, d'ordre $p+1$ ont les propriétés suivantes :

- si toutes les racines de $A(z)$ sont à l'intérieur du cercle unité alors les racines de $F_1(z)$ et de $F_2(z)$ sont sur le cercle unité ;
- les racines de $F_1(z)$ et de $F_2(z)$ apparaissent de façon alternée sur le cercle unité. Elles déterminent les coefficients LSF.

Les polynômes $F_l(z)$ et $F_2(z)$ ont respectivement, des racines en $z = +1$ et $z = -1$, qui peuvent être éliminées par division polynomiale. On définit alors les deux nouveaux polynômes suivants :

- Pour p pair :

$$G_1(z) = \frac{F_1(z)}{1 + z^{-1}} \tag{A.15}$$

$$G_2(z) = \frac{F_2(z)}{1 - z^{-1}} \tag{A.16}$$

- Pour p impair :

$$G_1(z) = F_1(z) \tag{A.17}$$

$$G_2(z) = \frac{F_2(z)}{1 - z^{-2}} \tag{A.18}$$

Les polynomes $G_1(z)$ et $G_2(z)$ sont égalememnt des polynômes symétriques d'ordre p. Chacun de ces polynômes possède donc $p/2$ racines conjuguées situées sur le cercle unité. Une factorisation des deux polynomes $G_1(z)$ et $G_2(z)$ peut être alors déduite :

- pour p pair

$$G_1(z) = (1 + z^{-1}) \prod_{i=1}^{p/2} \Big(\; 1 - 2cos(\omega_{2i-1}) z^{-1} + z^{-2} \; \Big) \tag{A.19}$$

$$G_2(z) = (1 - z^{-1}) \prod_{i=2}^{p/2} \Big(\; 1 - 2cos(\omega_{2i}) z^{-1} + z^{-2} \; \Big) \tag{A.20}$$

– pour p impair

$$G_1(z) = \prod_{i=1}^{(p+1)/2} \left(1 - 2cos(\omega_{2i})z^{-1} + z^{-2} \right) \quad (A.21)$$

$$G_2(z) = (1 - z^{-1}) \prod_{i=2}^{(p-1)/2} \left(1 - 2cos(\omega_{2i})z^{-1} + z^{-2} \right) \quad (A.22)$$

Les coefficients LSF, correspondent à la position angulaire ω_i, entre 0 et π, verifiant la relation :

$$0 = \omega_0 < \omega_1 < ... < \omega_p < \omega_{p+1} = \pi \quad (A.23)$$

La conversion AR en LSF est inversible. En effet, connaissant les coefficients ω_i, $i = 1..p$, on déduit les coefficients a_i à partir de la relation suivante :

$$A(z) = \frac{F_1(z) + F_2(z)}{2}. \quad (A.24)$$

Annexe B

Construction du dictionnaire d'un quantificateur vectoriel

Le but de la construction d'un quantificateur vectoriel est de produire un dictionnaire de M mots de codes de dimension k donnée pour lequel la distorsion de quantification definie par l'equation B.1 est minimale

$$D(Q) = E[d(x, Q(x)] = \sum_{i=1}^{M} \int_{s_i} d(x, c_i) p(x) dx \tag{B.1}$$

où $\{c_i|_{i=1:M}\}$ représentent les mots code du dictionnaire C et $\{s_i|_{i=1:M}\}$ sont les régions de l'espace définies par l'équation B.3.

Dans la plupart des cas, la distribution du signal à coder $p(x)$ n'étant pas connue, on exprime la distorsion en utilisant une séquence d'apprentissage $B\{x_1, ..., x_L\}$ de longueur L regroupant un grand nombre de réalisations de la variable aléatoire à quantifier. En pratique, nous avons utilisé une base d'apprentissage de 15926 échantillons, elle se compose de séquences de musique très variées (harmonique et transitoire). La distorsion de quantification à minimiser est donnée par :

$$D(Q) = \frac{1}{L} \sum_{i=1}^{L} d(x, Q(x)) \tag{B.2}$$

avec L la longueur de la base d'apprentissage.

La construction d'un quantificateur à M niveaux consiste à déterminer une partition S de l'espace R^P en M régions s_i et à associer un représentant c_i à chacune. Un quantificateur optimal vérifie les deux conditions d'optimalité suivantes :

– La partition S est optimale pour le dictionnaire $C = \{c_1, ..., c_M\}$ si :

$$s_i = \{x \in R^P / d(x, c_i) \leq d(x, c_j); \quad j \neq i\} \tag{B.3}$$

– Le dictionnaire C est optimal pour la partition S si :

$$\int_{s_i} d(x, c_i) p(x) dx = \arg_{u \in R^P} \int_{s_i} d(x, u) p(x) dx, \qquad i = 1, ..., M \tag{B.4}$$

où $p(x)$ est la densité de probabilité du vecteur d'entrée.

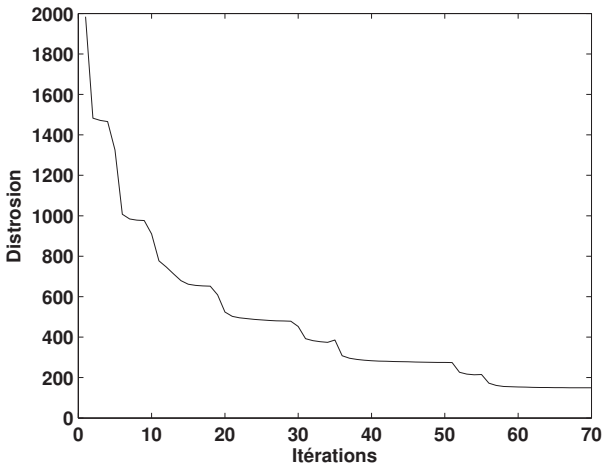


FIGURE B.1 – Distorsion en fonction du nombre d'itérations pour une base d'apprentissage de 15926 vecteurs LSF.

En choisissant la distance euclidienne quadratique, les mots code c_i du dictionnaire optimal coïncident avec le centre de gravité des classes s_i :

$$c_i = \frac{\int_{s_i} xp(x)dx}{\int_{s_i} p(x)dx} \tag{B.5}$$

Plusieurs algorithmes ont été proposés pour calculer le dictionnaire. Nous avons adopté l'algorithme de K-moyenne qui est à la base d'un grand nombre d'autres algorithmes et qui est toujours le plus souvent utilisé.

L'algorithme de la k-moyenne est un algorithme déterministe qui recalcule itérativement les mots du code du dictionnaire en satisfaisant alternativement aux deux conditions. Les opérations mathématiques de la condition B.4 sont difficiles à implémenter et, de plus, la distribution di signal n'est pas généralement connue. L'algorithme de la k-moyenne fait appel donc à une séquence d'apprentissage. Les étapes de la construction de dictionnaire de M mots de code sont les suivantes :

1. On initialise les M niveaux de reconstruction du dictionnaire $C_0 = \{c_{0,1}, ..., c_{0,M}\}$. Leur choix doit être fait de la manière à ce qu'il n'y ait pas de vecteurs égaux ;
2. On fait la classification des vecteurs d'apprentissage pour l'alphabet de reproduction actuel selon B.3. A chaque vecteur de la séquence d'apprentissage, on associe le voisin le plus proche parmi tous les centres. L'ensemble des vecteurs attribués au centre c_i constituera la classe s_i.
3. La distorsion produite par le dictionnaire actuel est calculée sur toute la séquence d'apprentissage par sommation des distances entre les vecteurs d'apprentissage x_j

et le mot code $c_{n,j}$ qui leur est le plus proche, n étant le numéro d'itération :

$$D_n = \frac{1}{L} \sum_{j=1}^{L} min_i(d(x_j, c_{n,i})) \tag{B.6}$$

4. Pour chaque classe s_i ainsi obtenue, on détermine le mot code $c_{n+1,i}$ représenant au mieux les vecteurs de la classe s_i étudiée selon le deuxième critère d'optimalité donné par B.4,

$$D(s_i) = \frac{1}{|s_i|} \sum d(x_j, c_{n+1,i}) \tag{B.7}$$

La nouvelle valeur du centre $c_{n+1,i}$ remplace alors l'ancienne $c_{n,i}$ dans le dictionnaire.

5. Tant que la condition d'arrêt n'est pas vérifiée, on reprend de façon itérative l'étape 2. L'arrêt se produit lorsque la distorsion totale sur le dictionnaire entre les deux itérations succéssives est inférieure ou égale au seuil empirique ε :

$$\frac{D_n - D_{n-1}}{D_n} \leq seuil \tag{B.8}$$

En prenant $seuil = 0$, on attend que l'algorithme trouve un minimum local. La figure B.1 montre la décroissance de l'historique de la distorsion en fonction des itérations selon une base d'apprentissage de 15926 vecteurs LSF, la taille du dictionnaire étant choisie à M=256.

Annexe C

Opérations matricielles

C.1 Produit d'Hadamard

Soit A et B deux matrices de même taille $(m \times n)$. Le produit d'Hadamard $A \odot B$,ou produit composante par composante, est une matrice C donnée par :

$$C = A \odot B = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1n}b_{1n} \\ \vdots & & \vdots \\ a_{m1}b_{m1} & \cdots & a_{mm}b_{mm} \end{pmatrix} \tag{C.1}$$

La taille de la matrice C obtenue est alors $(m \times n)$.

C.2 Factorisation de la matrice Transformée en Cosinus Discrète (TCD)

Soit $C_{N\times N}$ est la matrice de Transformé en Cosinus Discrète (TCD), définie par :

$$C = \frac{2}{\sqrt{M}} k_m k_n \cos\left(\frac{2\pi mm}{M} \right) \tag{C.2}$$

avec $m, n = 0, 1, ..., M-1$ et les coefficients k_j valent :

$$k_j = \begin{cases} 1/\sqrt{2} & j \neq 0 \\ 1 & j = 0 \end{cases} \tag{C.3}$$

On pose T, la matrice en cosinus définie par :

$$T = \cos\left(\frac{2\pi mn}{M}, \right) \tag{C.4}$$

W, une matrice diagonale, définie par :

$$W = \begin{pmatrix} 1/\sqrt{2} & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 0 \\ 0 & I_{N-1} \end{pmatrix} \tag{C.5}$$

et $Z_{r,l}$, matrice zero-padding, définie par :

$$Z_{r,l} = [0_{n\times l} I_n 0_{n\times r}]^T, \tag{C.6}$$

où $I_{n\times n}$ est la matrice identité.

La matrice C devient :

$$C = \frac{2}{\sqrt{M}} WTW \tag{C.7}$$

En substituant $T = \frac{\sqrt{M}}{2} Z^T(F + F^H)Z$, on peut avoir :

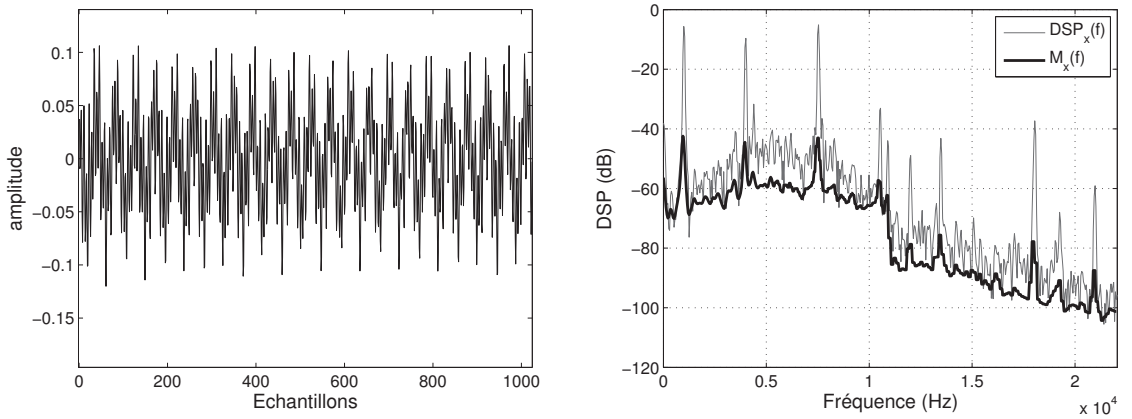
$$\begin{aligned} C &= WZ^T(F + F^H)ZW & \text{(C.8)} \\ &= WZ^T FSW^{-1} & \text{(C.9)} \end{aligned}$$

Puisque $FS = (F + F^H)Z$, la multiplication à droite par W^{-1} établit l'égalité finale.

Annexe D

Masquage fréquentiel et tatouage audio

L'effet de masquage fréquentiel se produit lorsque l'oreille humaine se trouve limiter à percevoir un son X_1, appelé masqué, en présence d'un autre son X_2 appelé masquant. Les deux sons, proches en fréquences et de puissances différentes, ne sont pas alors entendus de la même manière. On distingue principalement quatre types de masquage fréquentiel : bruit à bande étroite masque bruit à bande étroite (BMB), tonal masque tonal (TMT), bruit à bande étroite masque tonale (BMT) et tonale masque bruit à bande étroite (TMB).



(a) (b)

Figure D.1 – (a)- Extrait d'un signal d'un signal de glockenspiel échantillonné à 44.1 kHz (1024 ech.), (b)- DSP de x_n (en trait fin) et seuil de masquage $M_x(f)$ (en trait gras).

Les modèles psychoacoustique (MPA) ou d'audition fournit une estimation d'un seuil de masquage en fonction des caractéristiques du signal audio sur une durée donnée, notament celle de la fenêtre d'analyse. En effet, le seuil est une courbe dans le domaine fréquentiel propore à chaque segment du signal. Elle délimite le niveau maximal du bruit pouvant être injécté au segment considéré du signal, sans que ce bruit ne soit audible par l'oreille humaine. L figure D.1 présente le spectre d'un signal de golckespiel et le seuil de masquage correspondant, obtenu par simulation du modèle d'audtion 2 de MPEG. L'estimation de ce seuil est obtenue par modélisation du système auditif humain par un

banc de filtres de largeur de bande 1 Bark [1]

Ainsi, afin de garantir un tatouage inaudible, la DSP maximale $S_t(f)$ du signal de tatouage t_n doit doit rester en dessous du seuil de masquage $M_x(f)$ de x_n fourni par le MPA pour une fenêtre d'analyse :

$$S_x(f) \leq M_x(f), \quad |f| \leq \frac{fe}{2}. \tag{D.1}$$

1. Bark : unité de fréquence de 1 à 24, correspondant à [20 Hz - 20 kHz]. L'équation de conversion est donnée par : $f_{Bark} = 13.\arctan[0.76.\frac{f_{Herz}}{1000}] + 3.5.\arctan\left[\left(\frac{f_{Herz}}{7500}\right)^2\right]$ [Moreau 1995]

Annexe E

La Transformée de Hilbert

La transformée de Hilbert est un processus par lequel les fréquences négatives d'un signal sont en avance de phase de $\pi/2$ et les fréquences positives sont en phase retardée de $\pi/2$. La transformée de Hilbert d'un signal $g(t)$ est définie par :

$$\widehat{g}(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{g(\tau)}{t-\tau}d\tau = g(t) * \frac{1}{\pi t} \tag{E.1}$$

$$g(t) = -\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{\widehat{g}(\tau)}{t-\tau}d\tau = -\widehat{g}(t) * \frac{1}{\pi t} \tag{E.2}$$

La transformée de Fourier de $\widehat{g}(t)$ vaut :

$$\begin{aligned}\widehat{G}(\omega) &= G(\omega) \times \mathrm{TF}\left[\ \tfrac{1}{\pi t}\ \right] &\text{(E.3)}\\ &= -\mathrm{j}sign(\omega) \times G(\omega) &\text{(E.4)}\end{aligned}$$

Propriétés de la transformée de Hilbert

1. $g(t)$ et $\widehat{g}(t)$ ont le même spectre d'amplitude ;
2. Si $\widehat{g}(t)$ est la transformée de Hilbert de $g(t)$ alors la transformée de Hilbert de $\widehat{g}(t)$ est $-g(t)$;
3. $g(t)$ et $\widehat{g}(t)$ sont orthogonales sur l'intervalle $[-\infty,\ +\infty]$

$$\int_{-\infty}^{+\infty} g(t)\widehat{g}(t)dt = 0. \tag{E.5}$$

Représentation complexe de signaux

La représentation complexe d'un signal réel $g(t)$, appelée aussi signal analytique, correspond au résultat de la transformée de Hilbert multiplié par j auquel on ajoute le signal original (voir figure E.1). On obtient alors :

$$g_+(t) = g(t) + \mathrm{j}\widehat{g}(t) \tag{E.6}$$

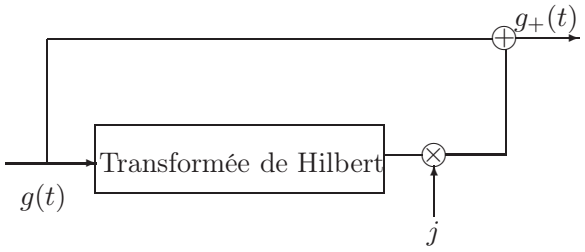


FIGURE E.1 – Génération du signal analytique.

La transformée de fourier de $g_+(t)$ est donnée par :

$$G_+(\omega) = G(\omega) + sgn(\omega)G(\omega) \tag{E.7}$$

Par conséquent

$$G_+(\omega) = \begin{cases} 2G(\omega), & \omega > 0 \\ G(0), & \omega = 0 \\ 0, & \omega < 0 \end{cases} \tag{E.8}$$

Une caractéristique importante du signal analytique est que son contenu spectral se situe dans l'intervalle de Nyquist positif. En effet, le déplacement de la partie imaginaire de $\pi/2$ ($\times$ j) et l'ajout de la partie réelle annulent les fréquences négatives. Il en résulte donc un signal sans fréquence négative.

De la même manière, on définit $g_-(t)$:

$$g_-(t) = g(t) - \mathrm{j}\widehat{g}(t) \tag{E.9}$$

$$G_-(\omega) = G(\omega) - sgn(\omega)G(\omega) \tag{E.10}$$

Par conséquent

$$G_-(\omega) = \begin{cases} 2G(\omega), & \omega < 0 \\ G(0), & \omega = 0 \\ 0, & \omega > 0 \end{cases} \tag{E.11}$$

Publications

- I. SAMAALI, G. MAHE, M. TURKI, "Watermark-aided pre-echo reduction in low bit-rate audio coding", JAES, Vol. 60, No. 6, 2012 June, pp. 431-443.

- I. SAMAALI, M. TURKI, G. MAHE, Attack restoration in low bit-rate audio coding, using an algebraic detector for attack localization, International Symposium on Image/Video Communications over fixed and mobile network (ISIVC 2010), Rabat, Maroc, september 2010.

- I. SAMAALI, M. TURKI, G. MAHÉ, Temporal envelope correction for attack restoration in low bit-rate audio coding, European Signal Processing Conference (EUSIPCO 2009), Glasgow, Royaume-Uni, août 2009, pp. 929-933.

- K. KHALDI, A. BOUDRAA, M. TURKI, T. CHONAVEL, I. SAMAALI, Audio encoding based on the Empirical Mode Decomposition, European Signal Processing Conference (EUSIPCO 2009), Glasgow, Royaume-Uni, août 2009, pp. 924-928.

- I. SAMAALI, M. TURKI, G. MAHE, Criteria to measure the quality of TVAR estimation for audio signals, European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, septembre 2007, pp. 798-802.