Insertion de données à haut débit dans des signaux de musique basée sur la transformée IntMDCT

Jonathan PINEL, Laurent GIRIN, Cléo BARAS

GIPSA-Lab
961 rue de la Houille Blanche, BP 46, 38402 GRENOBLE Cedex, France
{jonathan.pinel, laurent.girin, cleo.baras}@gipsa-lab.grenoble-inp.fr

Résumé – L'insertion de données consiste à cacher/insérer une information binaire dans un signal de manière imperceptible. Dans cette étude nous proposons un système d'insertion de données à haut débit approprié pour des signaux audio non-compressés (PCM comme pour le CD-Audio ou le format .wav). Cette technique est adaptée pour des applications non-sécuritaires, comme le contenu augmenté, qui requièrent un débit élevé mais avec peu de contraintes de robustesse face à des attaques. Le système proposé est basé sur une technique de quantification, la Quantization Index Modulation (QIM), appliquée aux coefficients de la transformée Integer Modified Discrete Cosine Transform (IntMDCT) et guidée par un modèle psychoacoustique (MPA). Cette technique permet d'obtenir des débit allant jusqu'à 300 kb/s par canal, améliorant les précédentes performances du système basé sur la MDCT classique [1].

Abstract – Data hiding consists in hiding/embedding binary information within a signal in an imperceptible way. In this study we propose a high-rate data hiding technique suitable for uncompressed audio signals (PCM as used in Audio-CD and .wav format). This technique is appropriate for non-securitary applications, such as enriched-content applications, that require a large bitrate but no particular robustness to attacks. The proposed system is based on a quantization technique, the Quantization Index Modulation (QIM), applied on the Integer Modified Discrete Cosine Transform (IntMDCT) coefficients of the signal and guided by a PsychoAcoustic Model (PAM). This technique enables embedding bitrates up to 300 kbps (per channel), outperforming a previous version based on regular MDCT [1].

1 Introduction

Le tatouage, formalisé dans les années 80 [2], consiste à insérer de l'information de manière indétectable dans un média, en conciliant débit, robustesse, sécurité et indétectabilité perceptuelle [3]. Majoritairement envisagé dans un but de protection des droits digitaux, le tatouage s'est aussi développé pour les applications où l'on cherche simplement à transmettre de l'information additionnelle (non secrète) à travers un média [3] sans caractère sécuritaire. Il s'agit alors d'un système de communication avec émetteur et décodeur dans lequel le média agit comme un canal (un support) de transmission avec la contrainte que l'information insérée soit indétectable perceptuellement. Le tatouage prend alors l'appellation générale d'insertion de données, et se concentre sur le débit d'insertion et l'indétectabilité perceptuelle, *a contrario* de la robustesse et de la sécurité. Cet article s'inscrit dans ce cadre applicatif : le média considéré est un signal de musique au format PCM standard non compressé (échantillonnage à 44.1 kHz, quantification scalaire uniforme sur 16 bits) compatible avec le CD-audio et les fichiers numériques de type *wav*. Plusieurs systèmes ont déjà été proposés pour ce type d'applications, les plus performants [1], [4] utilisant des techniques de tatouage par quantification de type *Quantization Index Modulation* (QIM) [5].

Dans [1], l'information est insérée par quantification des coefficients fréquentiels de la *Modified Discret Cosine Transform* (MDCT). L'insertion est effectuée en deux temps : dans les basses fréquences du spectre, un Modèle PsychoAcoustique (MPA) est utilisé pour calculer les paramètres des quantificateurs qui maximisent la quantité d'information tatouée (débit) sous contrainte d'inaudibilité ; parallèlement, dans les hautes fréquences, une insertion à paramètres fixes indique de quelle façon l'information dans les basses fréquences a été insérée. Cette technique atteint un débit d'environ 250 kbits/s mais entraîne des erreurs de décodage (de l'ordre de 10^{-6}) suite au bruit qu'ajoute la quantification du média tatoué sur 16 bits lors de sa reconversion au format PCM.

Dans [4], l'insertion s'effectue sur la base des coefficients Integer MDCT [6] (IntMDCT). Ce domaine fréquentiel offre le net avantage de ne plus avoir à se préoccuper de la requantification du média tatoué au format PCM et rend possible un décodage sans erreur. Le débit d'insertion obtenu (d'environ 140 kbits/s) est malheureusement plus faible que celui de [1] : en effet, un MPA est également utilisé pour maximiser le débit de tatouage, mais il est construit sur les bits prépondérants du spectre (*lead bits*) pour permettre au décodeur le recalcul des paramètres de quantification à l'identique, sans lequel le décodage ne pourrait se faire sans erreur ; la puissance du tatouage et donc le débit n'en sont que plus contraints et plus limités.

Dans cet article nous présenterons donc une amélioration du système de tatouage [1], intégrant la IntMDCT afin de conserver un débit d'insertion élevé et une inaudibilité forte tout en

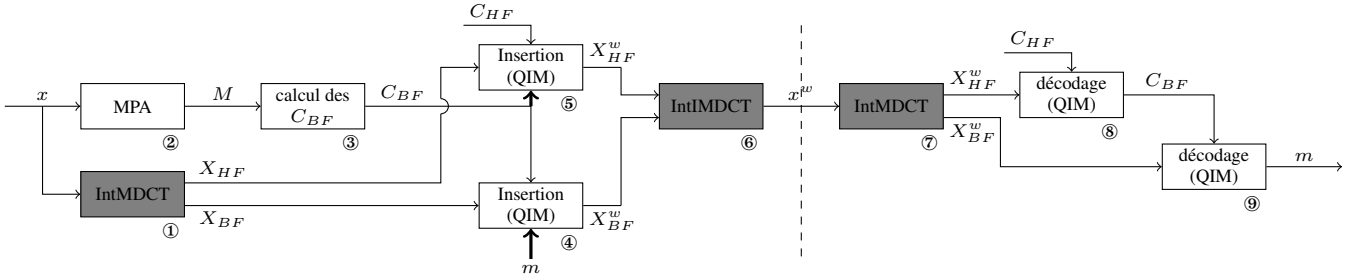

FIGURE 1 – Schéma bloc du système. A gauche de la ligne pointillée se trouve l'émetteur et à droite le décodeur.

atteignant un taux d'erreur de décodage nul (comme dans [4]).

Ce papier est organisé comme suit : la Section 2 est une vue d'ensemble du système tandis que la Section 3 présente chacun des blocs principaux en détail. La Section 4 montre les expérimentations et les résultats obtenus, et finalement des conclusions et perspectives sont discutées dans la Section 5.

2 Vue d'ensemble du système

Le système proposé, présenté figure 1, est construit sur les mêmes principes que [1]. Cette section présentera donc simultanément une vue d'ensemble du système proposé et de l'Etat de l'Art.

2.1 Emission

Pour réaliser l'insertion, le signal d'entrée x est donc d'abord transformé dans le plan Temps-Fréquence (TF) (bloc ①). Au lieu d'utiliser la MDCT (à l'instar de [1]), la transformée utilisée ici est l'IntMDCT, l'approximation entière de la MDCT [6]. Le processus d'insertion consiste à quantifier les coefficients fréquentiels $X(t, f)$ (blocs ④ et ⑤) en utilisant un ensemble spécifique de quantificateurs $\mathcal{S}_{C(t,f)}$ et en suivant la technique QIM scalaire [5]. Une fois les coefficients quantifiés, le signal est retransformé dans le domaine temporel en utilisant la transformée inverse de l'IntMDCT (IntIMDCT, bloc ⑥). Le signal est ensuite requantifié sur 16 bits (PCM). Contrairement à [1] où ce processus introduit un bruit, il est ici transparent de part la nature de la transformée (le signal conserve des valeurs entières).

Pour chaque trame t et chaque canal fréquenciel f, le MPA (bloc ②) fournit un seuil de masquage $M(t, f)$ utilisé pour calculer $C(t, f)$ le nombre maximum de bits disponibles à chaque point du plan TF sous contrainte d'inaudibilité. Cette capacité d'insertion $C(t, f)$ détermine à la fois « quelle quantité » d'information est insérée et « comment » cette information est insérée (et récupérée), puisqu'elle est l'unique paramètre des quantificateurs à la base de l'insertion. Par conséquent, ces capacités $C(t, f)$ doivent être connues au décodeur et doivent donc être soit estimées (comme dans [4]) soit transmises (comme dans [1]). Dans cette étude nous gardons le même principe général de [1] et proposons le processus suivant :

– A l'émetteur, les coefficients IntMDCT sont séparés en une partie « basses fréquences » (dénotée BF Fig. 1 et par la suite, représentant 15/16 du spectre) et une partie « hautes fréquences » (dénotée HF).
– La partie BF est utilisée pour transmettre l'information additionnelle « utile » m que l'on souhaite acheminer dans le signal audio hôte x. A cette fin, les capacités $C_{BF}(t, f)$ sont maximisées sous contrainte d'inaudibilité. Cette étape est le cœur de la méthode et est présentée plus en détail dans la section suivante.
– Ensuite, la partie HF est utilisée pour insérer les valeurs des capacités $C_{BF}(t, f)$ calculées précédemment et qui configurent totalement le processus d'insertion dans la partie BF. Pour être réalisable, les capacités $C_{HF}(t, f)$ doivent être connues à l'émetteur et au décodeur. Ainsi des valeurs fixes leurs sont attribuées, indépendamment de l'indice de trame et du contenu audio (on les note donc $C_{HF}(f)$), exploitant le fait qu'à ces fréquences élevées le système auditif humain est peu performant.

2.2 Décodage

Le décodage consiste simplement à inverser les opérations effectuées à l'émetteur : le signal hôte x^w est transformé dans le plan TF (bloc ⑦) et les coefficients IntMDCT sont séparés en deux parties BF et HF. Les capacités $C_{HF}(f)$ étant connues au décodeur, l'information insérée dans la partie HF est récupérée (bloc ⑧), donnant les valeurs $C_{BF}(t, f)$ qui sont à leur tour utilisées pour décoder l'information « utile » m contenue dans la partie BF (bloc ⑨).

3 Présentation détaillée

3.1 Transformée temps-fréquence

Dans cette étude, nous utilisons l'approximation entière de la MDCT, l'IntMDCT, afin de s'affranchir du bruit introduit sur les coefficients MDCT par la quantification temporelle PCM 16 bits [1] tout en conservant les propriétés intéressantes de la MDCT. La longueur de trame utilisée est de 2048 afin d'avoir une résolution fréquentielle suffisante tout en étant en accord avec la dynamique des signaux musicaux.

Toutes les transformations ayant une représentation sous forme de matrices orthogonales peuvent être approchées par

des transformées entières. Dans le cas de la MDCT il y a deux matrices orthogonales, la première étant la matrice de MDCT proprement dite et la seconde la matrice de fenêtrage et de repliement temporel. Le principe des transformées entières est de décomposer les matrices à approximer en rotations de Givens 2×2. Si l'on note $c = \cos(\theta)$ et $s = \sin(\theta)$ ($\theta \neq 2k\pi,\ k \in \mathbb{Z}$), ce type de matrices peut se factoriser ainsi :

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{c-1}{s} & 1 \end{pmatrix} \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{c-1}{s} & 1 \end{pmatrix}. \quad (1)$$

Dans cette décomposition, chacune des trois sous-matrices représente un opérateur du type :

$$\mathcal{L}_a \ : \begin{cases} \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \\ (x, y) \longrightarrow (x, y + ax) \end{cases}. \quad (2)$$

Une fois cette décomposition effectuée, on remplace simplement ces opérateurs $\mathcal{L}_a$ par leur approximation entière :

$$L_a \ : \begin{cases} \mathbb{Z}^2 \longrightarrow \mathbb{Z}^2 \\ (x, y) \longrightarrow (x, y + [ax]) \end{cases} \quad (3)$$

(avec $a \in \mathbb{R}$ et $[.]$ désigne l'opérateur d'arrondi), dont l'inverse est aussi un opérateur entier du même type (L_{-a}). Ce procédé est appelé « Lifting Scheme » (voir par exemple [6] pour une explication plus détaillée).

3.2 Technique d'insertion

La technique d'insertion est une QIM [5] scalaire qui constitue à tatouer chaque coefficient IntMDCT $X(t, f)$ indépendamment des autres avec $C(t, f)$ bits d'informations. Pour ce faire, un ensemble $\mathcal{S}_{C(t,f)}$ de $2^{C(t,f)}$ quantificateurs $\{\mathcal{Q}_c\}_{0 \leq c < 2^{C(t,f)}}$ est défini arbitrairement. Ceci implique que quel que soit $C(t, f)$, l'ensemble $\mathcal{S}_{C(t,f)}$ est généré de manière identique à l'émetteur et au décodeur. Les niveaux de quantification des différents quantificateurs d'un ensemble sont entrelacés (voir Figure 2), et chaque quantificateur est indexé par un mot c codé sur $C(t, f)$ bits. Du fait de l'entrelacement régulier des quantificateurs et des valeurs entières des coefficients IntMDCT, le pas de quantification de chaque quantificateur est donné par :

$$\Delta(t, f) = 2^{C(t,f)}. \quad (4)$$

L'insertion d'un mot de code c dans le coefficient IntMDCT $X(t, f)$ est simplement faite en quantifiant $X(t, f)$ avec le quantificateur $\mathcal{Q}_c$ indexé par x (voir Figure 2) pour obtenir le coefficient tatoué :

$$X^w(t, f) = \mathcal{Q}_c\left(X(t, f)\right). \quad (5)$$

Au décodeur, l'ensemble $\mathcal{S}_{C(t,f)}$ est généré en utilisant la valeur $C(t, f)$ (qui est soit décodée pour les BF, soit fixée et donc connue pour les HF). Ensuite le quantificateur $\mathcal{Q}_c$ sur lequel se situe le coefficient $X^w(t, f)$ est sélectionné, et le mot de code décodé est tout simplement l'index c du quantificateur.

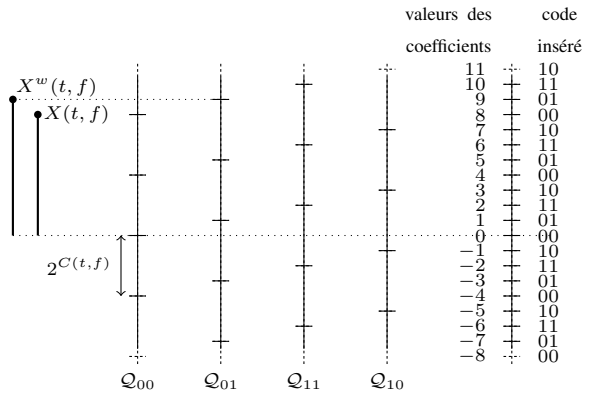

FIGURE 2 – Exemple d'insertion QIM utilisant un ensemble $\mathcal{S}_{C(t,f)}$ de quantificateurs pour $C(t, f) = 2$ avec leur code gray respectif et la grille globale. Le code binaire 01 est inséré dans le coefficient IntMDCT $X(t, f)$ par quantification en $X^w(t, f)$ en utilisant le quantificateur indexé par 01.

3.3 Modèle psychoacoustique

Le MPA utilisé dans le système (bloc ②) est directement inspiré du MPA de la norme MPEG-AAC [7]. Le MPA fournit un seuil de masquage $M(t, f)$ représentant la puissance maximale de bruit qui peut être ajouté en un point du plan TF donné tout en restant inaudible. Le MPA inclut un modèle de masquage fréquentiel ajusté suivant la tonalité du signal ainsi qu'un contrôle de pré-écho. A l'émetteur, le seuil de masquage est calculé pour chaque trame du signal et ensuite translaté dans sa globalité afin que la charge disponible corresponde exactement à la taille du message à transmettre m.

3.4 Calcul des capacités

Le calcul des capacités $C(t, f)$ est un point central de la méthode proposée. Le respect du format PCM étant déjà assuré par l'utilisation de la IntMDCT, le problème est d'optimiser le débit d'insertion sous contrainte d'inaudibilité. Dans cette étude, cette contrainte est que la puissance de l'erreur d'insertion dans le pire des cas reste sous le seuil de masquage. L'insertion étant effectuée par quantification uniforme, l'erreur de quantification dans le pire des cas est égale à la moitié du pas de quantification $\Delta(t, f)$, qui est directement relié à $C(t, f)$ par l'équation (4). Cette contrainte d'inaudibilité peut donc être écrite pour chaque point du plan TF :

$$\left(\frac{\Delta(t, f)}{2}\right)^2 < M(t, f). \quad (6)$$

En reliant, d'après (4), $\Delta(t, f)$ à la capacité $C(t, f)$ que l'on souhaite maximiser, on choisit finalement (voir [1]) :

$$C_{BF}(t, f) = \left\lfloor \frac{1}{2} \log_2\left(M(t, f)\right) + 1 \right\rfloor, \quad (7)$$

où $\lfloor . \rfloor$ dénote l'opération d'arrondi à l'entier inférieur. Expérimentalement les valeurs calculées sont toujours inférieures à 15. Ces valeurs peuvent donc être codées sur 4 bits (de 0

à 15). Cependant, insérer dans la région HF autant de mots de 4 bits qu'il y a de canaux dans la partie BF est impossible. C'est pour cette raison que des sous-bandes sont définies dans la partie BF, dans lesquelles toutes les capacités sont choisies égales à la valeur minimale des capacités de la sous-bande (dans le système présentée la partie BF est composée de 30 sous-bandes de 32 coefficients). Dans la partie HF (64 derniers coefficients), les capacités $C_{HF}(f)$ sont fixées à 1 ou 2 bits afin d'une part de respecter la contrainte d'inaudibilité et d'autre part d'avoir assez de bits pour coder les capacités BF ($30 \times 4 = 120 < 64 \times 2 = 128$).

4 Résultats

Les performances du système proposé sont évaluées en terme de qualité audio en fonction du débit d'insertion. La qualité audio est estimée à l'aide de l'algorithme de Perceptual Evaluation of Audio Quality (PEAQ) [8] et vérifiée par des tests d'écoute informels. La métrique utilisée, l'ODG (Objective Difference Grade) varie entre 0 et -4, et compare le contenu tatoué au contenu d'origine (0 indiquant une différence imperceptible et -4 une différence extrêmement gênante). Les tests ont été faits sur 12 extraits de 10 secondes de musique de différents styles musicaux (classique, jazz, rock et pop. . .) au format CD-Audio (44.1 kHz, 16 bits).

La figure 3 montre à travers quelques résultats que le débit d'insertion est fortement dépendant du contenu audio (*e.g.* inaudibilité jusqu'à 300 kb/s pour l'extrait de pop et seulement 200 kb/s pour le classique, moins énergétique), dû à l'utilisation d'un MPA. La comparaison avec le système précédent [1] montre que les résultats du système proposé sont bien meilleurs : à même ODG les débits sont plus élevés d'environ 50 kb/s en moyenne, et sont aussi beaucoup plus élevé que les 140 kb/s annoncés dans [4].

5 Conclusions et perspectives

Dans ce papier nous avons présenté un système d'insertion de données pour des signaux musicaux au format CD-Audio qui peut fournir un débit d'insertion allant jusqu'à 300 kb/s par canal (dépendant du contenu audio). Ce débit représente plus de 40% du débit original et présente un gain significatif par rapport aux résultats obtenus précédemment dans [1] et dans [4]. Cette technique est appropriée pour des applications de contenu enrichi, comme par exemple le système de séparation de source informée présentée dans [9].

A court terme nous comptons nous pencher sur les problèmes de synchronisation possible voir d'autres types d'attaques simples. A long terme, les signaux compressés étant désormais largement utilisés, nous pensons nous tourner dans de futures travaux vers le problème de compression/insertion conjointes en adaptant les principes présentés dans ce papier aux formats audio compressés, comme le MPEG4-SLS qui utilise aussi la IntMDCT.

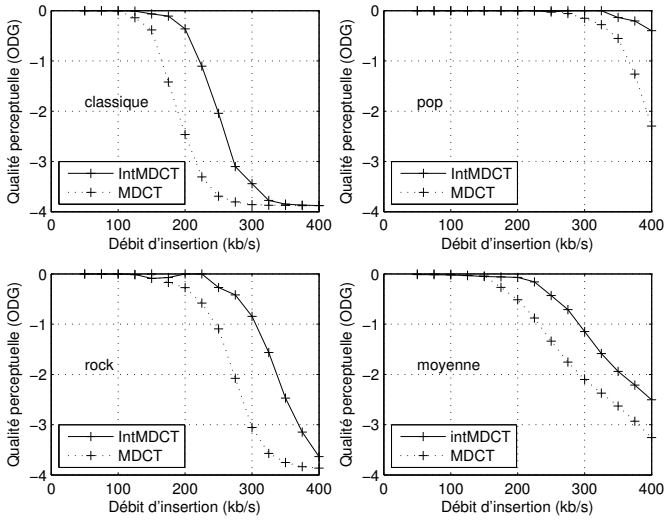

FIGURE 3 – Qualité perceptuelle en fonction du débit d'insertion pour 3 extraits et en moyenne du corpus de test pour le système avec intMDCT présenté et celui basé MDCT de [1].

Références

[1] J. Pinel, L. Girin, and C. Baras. A high-capacity watermarking technique for audio signals based on mdct-domain quantization. In *Proc. Int. Congress on Acoustics*, Sydney, Australia, 2010.

[2] M. Costa. Writing on dirty paper. *IEEE Trans. Inform. Theory*, 29(3) :439–441, 1983.

[3] C. Baras. *Tatouage informé de signaux audio numériques*. PhD thesis, Telecom ParisTech, 2005.

[4] R. Geiger, Y. Yokotani, and G. Schuller. Audio data hiding with high data rates based on intMDCT. In *Proc. IEEE Int. Conf. Acoust. and Speech, Signal Proc.*, Toulouse, France, 2006.

[5] B. Chen and G. Wornell. Quantization index modulation : A class of provably good methods for digital watermarking and information embedding. *IEEE Trans. Inform. Theory*, 47(4) :1423–1443, 2001.

[6] R. Geiger, Y. Yokotani, and G. Schuller. Improved integer transforms for lossless audio coding. In *Proc. Asilomar Conf. Signal, Systems and Computers*, Pacific Grove, California, 2003.

[7] ISO/IEC JTC1/SC29/WG11 MPEG. Information technology—Generic coding of moving pictures and associated audio information—Part 7 : Advanced Audio Coding (AAC), IS13818-7(E), 2004.

[8] ITU-R. Method for objective measurements of perceived audio quality (PEAQ), Recommendation BS.1387-1, 2001.

[9] M. Parvaix and L. Girin. Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding. In *Proc. IEEE Int. Conf. Acoust. and Speech, Signal Proc.*, Dallas, Texas, 2010.