Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Allocation binaire et déconvolution psychoacoustique de complexité réduite dans un codeur audio de haute qualité

Marcos Perreau-Guimaraes, Madeleine Bonnet, Nicolas Moreau*
UFR Math-Info, Université René Descartes-Paris 5
45 rue des Saints-Pères, 75270 Paris cedex 06, France
*ENST, Dpt TSI, 46 rue Barrault, 75634 Paris cedex 13, France
perm,bonnet@math-info.univ-paris5.fr, moreau@tsi.enst.fr

Résumé – Les codeurs de musique actuels atteignent de taux de compression supérieurs à 8 sans perte de qualité subjective en suivant le principe : ne pas coder ce que l'oreille n'entend pas. La mise en forme du bruit de codage se fait en deux étapes distinctes : le calcul d'un seuil de masquage à partir de la théorie psychoacoustique, puis l'allocation des ressources binaires en fonction du seuil de masquage. Le calcul du seuil de masquage est un problème difficile qui n'est qu'approché dans les codeurs actuels. Nous montrons que le calcul explicite du seuil de masquage n'est pas nécessaire et nous proposons un algorithme direct à faible complexité réalisant une meilleure approximation de la théorie psychoacoustique.

Abstract – Modern music coders reach high compression rate using the principle : do not code what the ear can not listen. Coding noise shaping is performed in two steps : the masked threshold calculus and then, the bit allocation according to the masked threshold. The evaluation of the masked threshold is a difficult problem which is strongly approximated in actual coders. We show that the masked threshold calculus is useless and we propose a fast direct algorithm performing a better approximation of psychoacoustical theory.

1 Introduction

On distingue historiquement deux catégories d'applications du codage audio. Le codage de la parole concerne essentiellement des signaux en bande téléphonique et en bande élargie. On s'intéressera ici particulièrement au codage haute qualité de la musique. Suivant le type d'application visé, la bande passante va de la bande élargie ($f_e = 16$ kHz) à la bande Hi-Fi ($f_e = 44.1$ kHz). Dans un CD-audio le signal n'est pas comprimé et le débit est de 705 kbit/s par canal (il y en a 2 pour la stéréo). C'est un débit bien trop élevé pour beaucoup d'applications d'où la nécessité de comprimer ces signaux. Un effort de recherche très important ces dernières années a permis d'atteindre des taux de compression de l'ordre de 8 sans perte de qualité. Un taux de compression encore supérieur, sans perte de qualité subjective, reste un enjeu industriel important. Les codeurs développés par le groupe MPEG (Moving Picture Expert Group) codent des signaux vidéo et audio [1]. La partie audio du codeur MPEG-1 [2], normalisée en 1993, définit une famille de codeurs divisée en trois couches pour diverses fréquences d'échantillonnage (32, 44.1 et 48 kHz) et pour des débits compris entre 64 et 192 kbit/s par canal. La couche 1, la plus simple, garantit la transparence du codage pour un débit de 192 kbit/s, la deuxième pour un débit de 128 kbit/s et la troisième n'est pas tout à fait transparente à 64 kbit/s. Les deux premières couches sont utilisées pour le codage de la vidéo sur CD-ROM et la troisième est le format audio le plus utilisé sur internet (MP3). En 1997 la norme MPEG-2 a défini le codeur MPEG-2/AAC [3] (Advanced Audio Coder). Ce codeur qui peut coder 5 canaux pour obtenir un effet de spatialisation permet un taux de compression de l'ordre de 10 en étant presque transparent.

Les codeurs modernes utilisent deux types de compression. La compression sans perte élimine les redondances du signal, comme par exemple la compression de Huffman. Ces techniques sont couramment utilisées pour comprimer les fichiers informatiques. Dans le cas des signaux numériques d'origines naturelles (sons, images, video,...), ces méthodes, qui conduisent à des taux de compression inférieurs à 2 en moyenne, ne sont pas suffisantes. Pour atteindre des taux dépassant 8 dans la plupart des applications, il faut se résoudre à accepter une perte d'information.

Le travail consiste alors à classer l'information contenue dans le signal numérique en fonction de son importance suivant un critère subjectif. Idéalement l'information sacrifiée est celle qui n'a aucune importance pour le critère choisi. Pour un signal audio, le critère est la perception par l'oreille de la distorsion (bruit de codage) introduite dans le signal. Cela revient à ne pas coder ce que l'oreille n'entend pas. Cette compression avec perte, souvent combinée avec une compression sans perte, permet d'atteindre des taux très supérieurs à 2 sans perte subjective de qualité. Les premières sections décrivent succinctement les méthodes utilisées par les codeurs actuels pour mettre en forme le bruit de codage et pour estimer le critère subjectif, appelé seuil de masquage. La troisième section expose le travail réalisé à partir de la question centrale : le calcul effectif du seuil de masquage est-il nécessaire dans le cadre du codage audio?

2 Mise en forme du bruit de codage

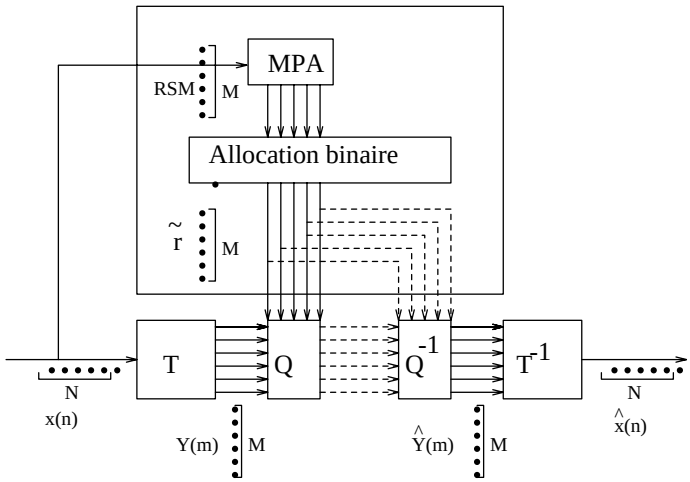

FIG. 1: *Schéma d'un codeur de musique*

La figure 1 donne le schéma général d'un codeur de musique. Un vecteur d'échantillons $x(n)$ du signal d'entrée subit une transformation temps/fréquence T qui fournit M coefficients $Y(m)$ qui sont quantifiés et transmis. Dans un codeur en sous bandes, M prend une valeur relativement faible ($M = 32$ dans le codeur MPEG1 couches 1 et 2). Dans un codeur par transformée M est en général compris entre 256 et 1024. En première approximation on peut considérer que chaque coefficient $Y(m)$ est représentatif du signal dans une sous bande de fréquence $[m\frac{fe}{2M}, (m+1)\frac{fe}{2M}]$, que l'on appelle dans la suite sous bande m. La mise en forme du bruit de codage est réalisée par le bloc d'allocation binaire qui répartit les bits disponibles en fonction des rapports signal à masque $RSM(m)$ dans chaque sous bande m. Les $RSM(m)$ sont fournis par le modèle psychoacoustique (MPA), qui est décrit dans la section suivante. Les coefficients qui correspondent à des sous bandes où le rapport signal à masque est important reçoivent le plus de bits.

L'algorithme d'allocation utilisé dans la plupart des codeurs est itératif. Le principe est simple : à chaque itération un bit est alloué au coefficient où le rapport bruit à masque (RBM) est le plus grand. Les rapports bruit à masque $RBM(m) = RSM(m)/RSB(m)$ sont obtenus à partir des rapports signaux sur bruits donnés dans des tables en fonction de la fréquence et du nombre de bits. Si à l'arrêt de l'algorithme le rapport bruit à masque $RBM(m)$ est inférieur à 1 dans toutes les sous bandes, les ressources binaires sont suffisantes pour garantir la transparence du codage. Dans le cas contraire le bruit de codage est audible, mais l'allocation binaire minimise ce bruit, en minimisant le dépassement du seuil de masquage par le bruit.

3 Calcul du seuil de masquage

Un point essentiel dans la compression des signaux audio, et plus généralement des signaux naturels, est de définir un critère objectif qui rende compte le plus fidèlement possible du critère subjectif. La psychoacoustique fournit des outils et des données expérimentales [4] qui servent à construire un modèle psychoacoustique de la réaction de l'oreille humaine au bruit de codage.
Le spectre fréquentiel du signal étant découpé en K sous bandes, notons $\sigma_X^2(k)$ et $\sigma_Q^2(k)$ des estimations de la puissance du signal original et du bruit de codage dans chaque sous bande k. Les valeurs de M et de K sont souvent différentes puisque l'estimation spectrale utilisée par le modèle psychoacoustique est en général indépendante de la transformée T. Le modèle estime, à partir de $\sigma_X^2(k)$, une courbe $\sigma_{THR}^2(k)$, appelée seuil de masquage, telle que le bruit de codage est inaudible si $\sigma_Q^2(k) < \sigma_{THR}^2(k)$ dans toutes les sous bandes k. Dans les codeurs dédiés à la musique, K est compris entre 256 et 1024.
Le calcul du seuil de masquage est assez complexe. En schématisant, on peut modéliser le passage du signal dans le système auditif par une opération non linéaire. L'organe principal de l'oreille interne est la membrane basilaire qui réalise une analyse fréquentielle des vibrations mécaniques propagées par le conduit auditif. La puissance du signal est mesurée sur des bandes de fréquences non uniformes, appelées bandes critiques. On définit une nouvelle unité de fréquence, le Bark qui correspond à la largeur d'une bande critique. La bande de fréquences audibles chez l'être humain va de 20 Hz à 20 kHz, ce qui correspond à 25 bandes critiques. Dans la pratique on utilise souvent une subdivision uniforme des bandes critiques en B sous bandes, dites sous bandes basilaires. La puissance $\sigma_X^2(b)$ du signal dans chaque sous bande basilaire b définit le spectre basilaire. La sortie de la membrane basilaire est modélisée par l'excitation

$$E_X(b) = \sum_{b'} f_{etal}(b', b)\sigma_X^2(b')$$

où $f_{etal}(b', b)$, fonction d'étalement, donne l'influence de la sous bande basilaire b' sur la sous bande b. En effet, la mesure dans l'oreille de la puissance du signal dans les sous bandes n'est pas bien localisée en fréquence.

Lorsque l'on fait écouter à un sujet successivement le son original x et le son dégradé $\hat{x} = x + q$, le système nerveux compare les excitations $E_X(b)$ et $E_{X+Q}(b)$ de ces deux signaux, et décide qu'il y a une dégradation si dans une sous bande b le rapport de ces excitations est inférieur à une valeur limite dépendant de b. Le bruit de codage ne sera pas entendu si la condition d'inaudibilité

$$E_Q(b)/E_X(b) < av(b) \qquad (1)$$

est satisfaite dans toutes les sous bandes basilaires b. Le terme $av(b)$, appelé taux ou indice de masquage, dépend de la fréquence, de la puissance du son original et de la nature tonale du son original et du bruit. La nature tonale correspond au timbre du son : un son pur (sinus) est parfaitement tonal, alors qu'un bruit blanc n'est pas du tout tonal.

Le seuil de masquage σ^2_{THR} est la fonction σ^2_Q qui vérifie la condition d'inaudibilité (1) tout en optimisant un certain critère $J(\sigma^2_Q)$ dépendant de l'application. En codage ce critère est le débit estimé par la formule classique

$$J(\sigma^2_Q) = 0.5 \sum_b w(b) \log_2 \frac{\sigma^2_X(b)}{\sigma^2_Q(b)}$$

où $w(b)$ est le nombre de coefficients $Y(m)$, qui sont effectivement codés et transmis, appartenant à la sous bande basilaire b. Le seuil de masquage σ^2_{THR} est la solution du problème d'optimisation sous contraintes :

$$\begin{cases} \min\limits_{\sigma^2_Q} J(\sigma^2_Q) \\ \sum\limits_{b'} f_{etal}(b',b)\ \sigma^2_Q(b') \le av(b) E_X(b), \quad \forall b \\ \sigma^2_Q(b) \ge 0, \quad \forall b \end{cases}$$

La résolution de ce problème par des méthodes classiques [5] de calcul numérique n'est pas envisageable dans une application de codage à cause de la complexité requise. Les codeurs actuels font deux approximations :
1- Ils considèrent que la solution du problème d'optimisation est la solution de l'équation $E_{THR}(b) = av(b)E_X(b)$, ce qui réduit le problème à inverser le calcul de l'excitation pour retrouver σ^2_{THR}. On a alors un problème de déconvolution sous contrainte, puisque la condition
$\sigma^2_{THR}(b) \ge 0 \ \ \forall b$ doit être vérifiée.
2- La résolution de ce problème est encore trop complexe et les codeurs actuels se contentent d'une multiplication par une fonction de pondération pour obtenir un résultat homogène à une puissance.

Le seuil de masquage est en pratique recalculé dans les M sous bandes de la transformée par une interpolation linéaire et les sorties du bloc MPA sont les rapports signaux à masque dans les sous bandes m.

4 Calcul direct de l'allocation binaire

L'approche classique consiste à d'abord calculer un seuil de masquage sans tenir compte du nombre de bits limité, puis à distribuer les bits disponibles de façon à s'approcher le plus possible de ce seuil. Or, si l'on met à plat le problème de la mise en forme du bruit de codage, l'objectif est de distribuer le plus efficacement possible les bits disponibles de façon à respecter au mieux la condition d'inaudibilité (1). La conclusion immédiate de cette remarque est qu'il n'est pas nécessaire de calculer explicitement un seuil de masquage. Notre proposition est alors d'utiliser directement un critère dérivé de la condition d'inaudibilité pour déterminer à chaque itération de l'algorithme d'allocation binaire le coefficient où l'allocation d'un bit est la plus efficace.
La sous bande b_0 dans laquelle l'excitation du bruit contribue le plus à la sensation de dégradation est celle où le rapport $E_Q(b)/av(b)E_X(b)$ est le plus grand

$$b_0 = \arg\max_b \left(\frac{E_Q(b)}{av(b)E_X(b)}\right) \qquad (2)$$

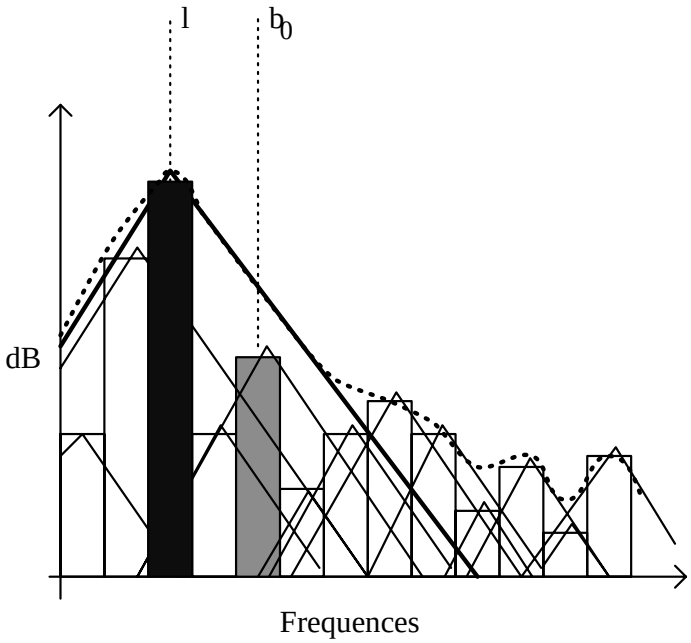

FIG. 2: *L'excitation E_Q (en pointillés) du bruit de codage dans une sous bande est obtenue en sommant les contributions de la puissance du bruit sur les sous bandes voisines. Dans cet exemple l'excitation en b_0 ne dépend pratiquement que de la puissance du bruit en l.*

Il faut donc mettre le bit dans la sous bande l de façon à diminuer le plus $E_Q(b_0)$. La figure 2 illustre le calcul de l'excitation E_Q du bruit de codage, qui est la somme des contributions de la puissance du bruit dans chaque sous bande pondérée par une fonction d'étalement, avec dans cet exemple une contribution presque exclusive de la sous bande l sur l'excitation dans la sous bande b_0. Quand on dit que l'on ajoute un bit dans une sous bande basilaire b, en fait on ajoute un bit pour coder chaque coefficient $Y(m)$ de la transformée appartenant à la sous bande basilaire b. C'est à dire que l'on ajoute $w(b)$ bits.
Si, au début d'une itération donnée, le nombre total de bits alloués aux coefficients inclus dans une sous bande basilaire b est $r(b)$, alors, la puissance du bruit de codage dans cette sous bande basilaire est donnée par

$$\sigma^2_Q(b) = \sigma^2_X(b)\ g(r(b),b) \qquad (3)$$

où $g(r(b),b)$ est le rapport signal sur bruit dans la sous bande basilaire b. Si le nombre de bits alloués à la sous bande basilaire b est incrémenté de $w(b)$, la puissance du bruit dans la sous bande basilaire b devient

$$\sigma^2_Q(b) = \sigma^2_X(b) g(r(b)+1,b) \qquad (4)$$

La diminution de la valeur de $E_Q(b_0)$ quand on incrémente le nombre de bits dans la sous bande b vaut alors

$$f_{etal}(b,b_0)\sigma^2_X(b)[g(r(b),b) - g(r(b)+1,b)] \qquad (5)$$

La sous bande basilaire où incrémenter le nombre de bits a le plus d'effet sur l'excitation $E_Q(b_0)$ est donnée par

$$l = \arg\max_b (f_{etal}(b,b_0)\sigma^2_X(b)[g(r(b),b) - g(r(b)+1,b)]) \qquad (6)$$

La sous bande basilaire l est donc celle où on va attribuer $w(l)$ bits à une itération donnée. Dans la pratique, pour un signal audio naturel, on remarque que l est toujours proche de b_0, ce qui s'explique par la décroissance très rapide de la fonction d'étalement $f_{etal}(b', b)$ lorsque $|b' - b|$ augmente.

On obtient finalement l'algorithme

$E_X(b) \leftarrow \sum_{b'} f_{etal}(b', b)\ \sigma_X^2(b')\ \ \forall b$
Calcul de $av(b)\ \ \forall b$
$r(b) \leftarrow 0\ \ \forall b$
Tant que $\sum_b r(b) <$ nombre de bits disponibles
  Pour tout b
    $\sigma_Q^2(b) \leftarrow \sigma_X^2(b)/RSB(b, r(b))\ \ \forall b$
    $E_Q(b) \leftarrow \sum_{b'} f_{etal}(b', b)\ \sigma_Q^2(b')$
  Fin pour
  $b_0 \leftarrow \arg\max_b(\frac{E_Q(b)}{av(b)E_X(b)})$
  $l \leftarrow \arg\max_b(f_{etal}(b, b_0)\sigma_X^2(b)$
    $[g(r(b), b) - g(r(b) + 1, b)])$
  $r(l) \leftarrow r(l) + w(l)$
Fin tant que

A la fin de cet algorithme, les $r(b)$ bits qui sont alloués à la sous bande basilaire b doivent être redistribués aux $w(b)$ coefficients de la transformée qui sont inclus dans la sous bande basilaire b. On utilise dans chaque sous bande basilaire l'algorithme classique d'allocation binaire : à chaque itération un bit est alloué à la sous bande m incluse dans la sous bande basilaire b où le rapport signal sur bruit est le plus faible, jusqu'à épuisement des bits alloués à la sous bande b. Finalement $\tilde{r}(m)$ bits sont alloués pour coder chaque coefficient de la transformée.

Pour tout b
  $\tilde{r}(m) \leftarrow 0\ \ \forall m \in b$
  Tant que $\sum_{m \in b} \tilde{r}(m) < r(b)$
    $m_0 \leftarrow \arg\min_{m \in b}(RSB(m, \tilde{r}(m)))$
    $\tilde{r}(m_0) \leftarrow \tilde{r}(m_0) + 1$
  Fin tant que
Fin pour

Cet algorithme ne présente pas de difficulté et est peu coûteux en terme de complexité.

5 Implémentation et résultats

Ayant reposé le problème de la mise en forme du bruit de codage d'un point de vue théorique, notre algorithme a été implémenté dans un codeur développé à l'ENST [6] pour une bande intermédiaire (f_e = 32 kHz) et un débit de 64 kbit/s. Ce codeur est destiné au codage de haute qualité parole/musique pour des applications multimédia. Dans ce codeur $M = 256$ (fenêtres de 16 ms), $K = 512$ et il y a 49 sous bandes basilaires.

L'augmentation effective de complexité du bloc de mise en forme du bruit de codage par rapport à la méthode classique ne dépasse pas 30%, ce qui fait une augmentation de l'ordre de 10% de la complexité globale du codeur. Ces valeurs sont à comparer avec une complexité de 100 à 1000 fois plus importante en utilisant des outils classiques de calcul numérique pour résoudre le problème d'optimisation psychoacoustique.

Des tests d'écoute informels réalisés sur un corpus varié parole et musique ont montré, par rapport à l'algorithme standard, une amélioration sensible.

6 Conclusion

Nous avons proposé une nouvelle méthode de mise en forme du bruit de codage qui ne calcule pas explicitement un seuil de masquage. Cette méthode permet une meilleure approximation de l'optimisation psychoacoustique que dans les codeurs actuels, tout en maintenant une complexité faible, compatible avec des applications de codage audio.

Références

[1] P. Noll *MPEG digital audio coding* IEEE Signal Processing, vol. 14, nø 5, pages 59-81, Sept. 1997.

[2] Norme internationale ISO/CEI 11172. *Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1,5 Mbit/s*, 1993.

[3] Norme internationale ISO/CEI 13818-7. *MPEG-2 Advanced Audio Coding, AAC*, 1997.

[4] E. Zwicker and E. Feldtkeller. *Psychoacoustique, l'oreille récepteur d'information.* Masson, 1981.

[5] R. Veldhuis. Bit rates in audio source coding. *IEEE J. on Selected Areas in Com.*, 10, no. 1:86–96, 1992.

[6] A. Jbira, N. Moreau, and P. Dymarski. Low delay coding of wideband audio (20 Hz - 15 kHz) at 64 kbps. *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1998.