Une approche statistique pour l'optimisation du MPEG-2/4 AAC (Advanced Audio Coder) en mode stéréophonique matricé (MS stéréo)

Olivier DERRIEN, Pierre DUHAMEL

CNRS / Laboratoire des Signaux et Systèmes
Supélec, Plateau du Moulon, 91192 GIF-SUR-YVETTE Cedex, France
olivier.derrien@lss.supelec.fr, pierre.duhamel@lss.supelec.fr

Résumé – Le mode stéréophonique matricé, ou stéréo MS, permet d'améliorer l'efficacité d'un codeur MPEG AAC lorsque les canaux gauche et droite sont fortement corrélés. Toutefois, la procédure d'optimisation est compliquée par le fait que le critère perceptuel de qualité ne porte plus sur les signaux quantifiés. Nous proposons une nouvelle méthode d'optimisation faisant appel à un modèle statistique de la quantification, qui s'avère plus efficace que la méthode standard.

Abstract – The MS stereo mode improves the efficiency of a MPEG Advanced Audio Coder when Left and Right channels are highly correlated. Yet, the optimization procedure is more complex, as since the perceptual criteria does not depend on the quantized channels. We propose a new optimization method based on a statistical model for the quantization. Compared to the standard algorithm, this new method improves the audio quality.

1 Introduction

Le MPEG AAC (Advanced Audio Coder) est aujourd'hui la norme de compression des signaux audio la plus efficace. Spécifié dans MPEG-2 [1] puis intégré à la norme MPEG-4, ce codeur fréquentiel est construit autour d'une transformée en cosinus discrète modifiée (MDCT). Son mode de fonctionnement classique consiste à minimiser la distorsion perçue par l'auditeur sous contrainte de débit.

Pour le codage de signaux stéréophoniques, le mode LR, ou double-mono, consiste à coder les canaux L (voie gauche) et R (voie droite) comme des signaux monophoniques indépendants, le débit étant identiquement distribué entre les deux. Lorsque les signaux des canaux L et R sont fortement corrélés, ce qui est relativement fréquent en stéréophonie, cette méthode est peu efficace. On lui préfère alors le mode MS, ou stéréophonie matricée, qui consiste à augmenter la décorrélation inter-canaux par matriçage. Les canaux codés puis transmis sont dénommés M (voie milieu) et S (voie de côtés). Après décodage, le matriçage inverse est appliqué avant la restitution des signaux. Cette méthode permet souvent d'améliorer l'efficacité du codage, mais complique la procédure d'optimisation.

Dans cet article, nous commençons par décrire l'algorithme d'optimisation standard pour le mode MS dont nous montrons qu'il ne respecte pas les principes de psychoacoustique utilisés en codage audio. Nous proposons alors une nouvelle méthode d'optimisation utilisant un modèle statistique de la quantification. Le nouvel algorithme est enfin comparé à l'algorithme standard.

2 Méthode standard en mode MS

2.1 Quantification

Dans cette section, nous décrivons la quantification du signal porté par un seul canal. En sortie de la transformée, les coefficients spectraux correspondant à la fenêtre d'analyse courante sont regroupés en sous-bandes fréquentielles de longueur variable, puis sous-quantifiés au moyen d'un quantificateur scalaire non-uniforme. La résolution peut être fixée indépendamment dans chaque sous-bande par le choix d'un paramètre entier, appelé *facteur d'échelle*. Les formules de quantification directe et inverse pour les coefficients spectraux $X(k)$ dans la sous-bande s, s'écrivent sous la forme suivante :

$$i(k) = R\left(\left[X(k)\,2^{-\frac{\varphi(s)}{4}}\right]^{\frac{3}{4}}\right) \quad (1)$$

$$\hat{X}(k) = [i(k)]^{\frac{4}{3}}\;2^{\frac{\varphi(s)}{4}} \quad (2)$$

où $\varphi(s)$ désigne le facteur d'échelle et R une fonction d'arrondi. Pour simplifier l'écriture, nous utiliserons la variable $A(s) = 2^{\frac{\varphi(s)}{4}}$, que nous appelons *paramètre d'échelle*. Le choix de $A(s)$ détermine la puissance d'erreur de quantification, c'est-à-dire la distorsion, ainsi que le nombre de bits nécessaire à la représentation binaire des indices $i(k)$ et du facteur d'échelle $\varphi(s)$, c'est-à-dire le débit.

Lorsque la quantification est appliquée à des canaux directement restitués à l'auditeur, par exemple L et R en mode LR, la psychoacoustique donne un critère de transparence sonore : il suffit que la puissance de l'erreur de quantification dans chaque sous-bande reste inférieure au seuil de masquage, calculé par le modèle psychoacoustique. A débit fixe, il arrive fréquemment que la transparence ne soit pas atteignable. Un algorithme d'optimisation détermine alors les facteurs d'échelle de telle

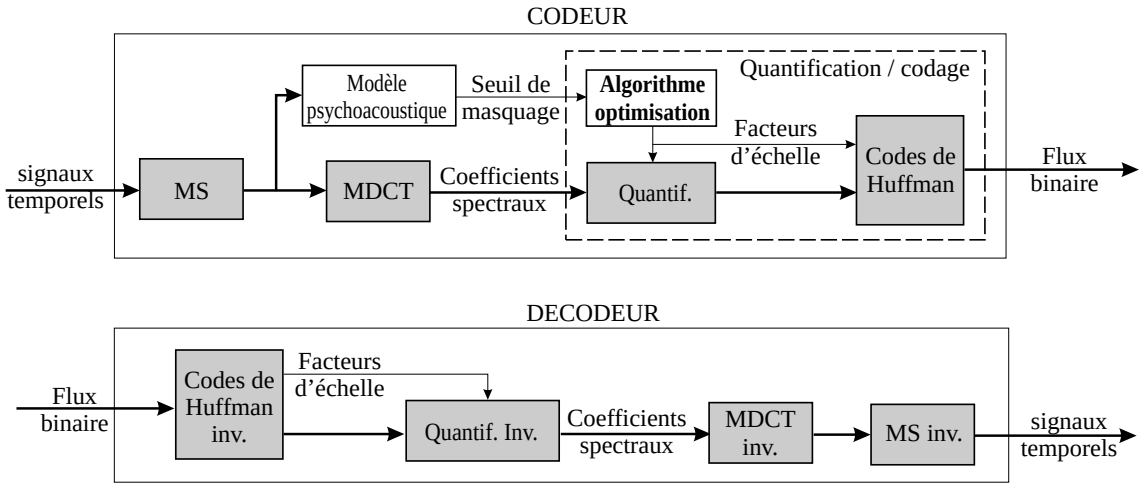

Figure 1: Schéma d'un codeur-décodeur MPEG-AAC en mode MS.

sorte que l'erreur soit la moins gênante possible.

2.2 Algorithme d'optimisation

En mode stéréo MS, le matriçage est appliqué dans le domaine temporel, avant la MDCT (voir figure 1). On note $l(n)$, $r(n)$, $m(n)$ et $s(n)$ respectivement les signaux sur les canaux L, R, M et S. La transformation directe s'écrit :

$$\begin{bmatrix} m(n) \\ s(n) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} l(n) \\ r(n) \end{bmatrix} \quad (3)$$

Après la MDCT, la quantification et le codage binaire sont appliqués indépendamment aux coefficients spectraux sur les canaux M et S, notés $M(k)$ et $S(k)$. Comme ces canaux ne seront pas directement restitués à l'auditeur, il n'est pas possible d'appliquer tel quel le modèle psyachoacoustique. La méthode classique consiste alors à ajouter une extension au modèle psychoacoustique, appelée *Imaging Control Process* [1], afin qu'il calcule des *pseudo-seuils de masquage* pour les canaux M et S. Le respect de ces pseudo-seuils est supposé équivalent au respect des seuils réels, si le mode LR était utilisé au lieu du mode MS.

La répartition du débit entre les canaux est aussi un point délicat du mode MS. La répartition optimale n'est en général pas égale, car les puissances des signaux $m(n)$ et $s(n)$ sont d'autant plus différentes que les signaux $l(n)$ et $r(n)$ sont corrélés. Une solution classique consiste à utiliser un critère d'entropie perceptuelle, qui donne la borne inférieure de débit pour une distorsion donnée (voir Johnston [2, 3]).

2.3 Discussion

La méthode standard décrite précédemment est théoriquement efficace, pourvu que les pseudo-seuils de masquage soient significatifs. Afin de le vérifier, nous réglons un codeur AAC en mode MS de telle sorte que le débit soit minimum sous contrainte de masquage. Nous utilisons une version *souple* de la contrainte qui consiste à imposer une valeur minimale, notée α, à la probabilité de respect du seuil de masquage dans chaque sous-bande et pour chaque canal :

$$\text{Proba}\left(P_q(s) \leq \tilde{T}_\psi(s)\right) \geq \alpha \quad (4)$$

où $P_q(s)$ est la puissance de l'erreur de quantification et $\tilde{T}_\psi(s)$ le pseudo-seuil de masquage. On choisit typiquement une valeur de α proche de 1. Nous prenons $\alpha = 0.95$. En d'autres termes, nous réglons le codeur de manière à ce que les seuils de masquage soint respectés 95 % du temps. Cette formulation du problème permet d'approcher de manière simple et fiable la solution optimale au moyen d'un modèle statistique de la quantification (voir [4]).

Après décodage, nous calculons l'erreur de quantification sur les canaux L et R, et comparons sa puissance dans chaque sous-bande aux seuils de masquage réels, calculés sur ces mêmes canaux. Les mesures de probabilité de respect du seuil sont présentées dans le tableau 1. On constate que la contrainte est vérifiée lors de la quantification, mais plus après dématriçage. Il n'y a donc pas d'équivalence entre seuils réels et pseudo-seuils.

signal	canal M	canal S	canal L	canal R
Cocaine	0.96	0.96	0.48	0.52
Revolution	0.96	0.96	0.58	0.41

Table 1: Probabilité de respect du seuil de masquage en mode MS avec la méthode standard, à la sortie du quantificateur (M et S) et après dématriçage (L et R).

3 Nouvelle méthode

Afin de proposer une méthode d'optimisation plus efficace, nous abandonnons le principe des pseudo-seuils de masquage et proposons de régler la quantification sur les canaux M et S en évaluant le critère de distorsion sur les canaux L et R après dématriçage, ce qui nécessite uniquement de connaître les seuils de masquage réels.

3.1 Enoncé du problème d'optimisation

Nous avons montré (voir [4]) que la solution du problème classique d'optimisation, consistant à minimiser la distorsion perçue sous contrainte de débit, peut être atteinte en résolvant plusieurs fois un second problème, plus simple, qui consiste à minimiser le débit sous une contrainte de distorsion semblable

à l'inéquation (4), à ceci près que le seuil de puissance n'est pas nécessairement un seuil de masquage. Nous qualifions ce second problème de *primal*, alors que le problème classique est appelé *dual*.

Dans le cadre de l'optimisation en mode MS, le problème primal peut s'exprimer sous la forme suivante : étant donnés des seuils $T^{(L)}(s)$ et $T^{(R)}(s)$ portant sur la puissance d'erreur de quantification sur les canaux L et R après dématriçage, on cherche les paramètres d'échelle A_M et A_S, réglant la quantification des canaux M et S, de telle sorte que le débit total soit minimum sous la contrainte de distorsion suivante :

$$\begin{cases} \text{Proba}\left(P_q^{(L)}(s) \leq T^{(L)}(s)\right) & \geq \quad \alpha \\ \text{Proba}\left(P_q^{(R)}(s) \leq T^{(R)}(s)\right) & \geq \quad \alpha \end{cases} \tag{5}$$

En supposant que le débit est une grandeur additive suivant les sous-bandes, ce qui est approximativement exact dans un codeur AAC, ce problème peut être résolu indépendamment par sous-bande.

3.2 Modèle statistique de l'erreur de quantification en sous-bande

Dans chaque sous-bande, nous cherchons à caractériser la loi des variables aléatoires $P_q^{(L)}$ et $P_q^{(R)}$ en fonction de A_M et A_S, afin d'obtenir une forme plus explicite des contraintes (5). Nous avons présenté précédemment (voir [4]) un modèle statistique de la puissance d'erreur en sous-bande s'appliquant aux canaux quantifiés. Considérons une sous-bande s particulière et notons $m^{(X)}$ et $\sigma^{2\,(X)}$ la moyenne et la variance de la variable aléatoire $P_q^{(X)}(s)$ représentant la puissance d'erreur sur le canal X dans cette sous-bande. Nous avons montré que, lorsque la résolution du quantificateur est suffisante, $P_q^{(X)}$ suit une loi gaussienne de paramètres :

$$m^{(X)} \quad = \quad \Delta_k \, c_X \, A_X^{\frac{3}{2}} \tag{6}$$

$$\sigma^{2\,(X)} \quad = \quad \Delta_k \, d_X \, A_X^3 \tag{7}$$

où Δ_k représente la largeur de la sous-bande, en nombre de composantes spectrales. c_X et d_X sont des paramètres dépendant uniquement de la fonction d'arrondi R et des moments des composantes spectrales dans la sous-bande. Dans notre cas, X peut désigner les canaux M et S, mais pas L et R qui ne sont pas directement quantifiés.

Cette modélisation suppose que les composantes spectrales de l'erreur sur M et S sont des variables aléatoires indépendantes et identiquement distribuées dans la sous-bande. Nous pouvons aussi considérer qu'il y a indépendance inter-canaux, car l'erreur de quantification peut être considérée comme indépendante du signal en haute résolution. Alors, il est possible de relier les moments de $P_q^{(L)}$ et $P_q^{(R)}$ à ceux de $P_q^{(M)}$ et $P_q^{(S)}$:

$$m^{(L)} = m^{(R)} = m^{(M)} + m^{(S)} \tag{8}$$

$$\sigma^{2\,(L)} = \sigma^{2\,(R)} = \sigma^{2\,(M)} + \sigma^{2\,(S)} + \frac{4}{\Delta_k} \, m^{(M)} \, m^{(S)} \tag{9}$$

$P_q^{(L)}$ et $P_q^{(R)}$, qui suivent la même loi selon ce modèle, s'écrivent comme la somme d'un grand nombre de variables indépendantes qui sont les erreurs par par raie spectrale. Il est donc raisonnable se supposer que cette loi est aussi gaussienne. Les contraintes (5) sont alors équivalentes à l'inéquation suivante :

$$m^{(L,R)} + \beta \, \sigma^{(L,R)} \leq T_{\min} = \min\left(T^{(L)}, T^{(R)}\right) \tag{10}$$

avec :

$$\beta = \sqrt{2}\,\text{Erf}^{-1}\,(2\alpha - 1) \tag{11}$$

En linéarisant l'écart-type par un développement au premier ordre, valable lorsque $\frac{A_M}{A_S}$ est proche de 1, c'est-à-dire lorsque les facteurs d'échelle des canaux M et S ne sont pas très différents, on peut écrire cette contrainte sous la forme :

$$\gamma_M \, A_M^{\frac{3}{2}} + \gamma_S \, A_S^{\frac{3}{2}} \leq T_{\min} \tag{12}$$

avec :

$$\gamma_X = \Delta_k \, c_X + \beta \sqrt{\frac{\Delta_k}{d_M + d_S + 4 c_M c_S}} \, (d_X + 2 c_M c_S) \tag{13}$$

3.3 Algorithme d'optimisation

Dans la section précédente, nous avons obtenu, moyennant certaines approximations, une expression simplifiée (12) de la contrainte de distorsion (5). La recherche de la solution du problème d'optimisation primal suppose que l'on connaisse la relation liant A_M et A_S au nombre de bits consommé. Nous avons montré qu'un modèle log-linéaire est une approximation valide dans le cas du codeur AAC. Nous cherchons donc à minimiser la fonction de coût suivante :

$$f(A_M,\, A_S) = -\log A_M - \log A_S \tag{14}$$

On montre que la solution optimale est alors :

$$A_X = \left(\frac{T_{\min}}{2\,\gamma_X}\right)^{\frac{2}{3}} \tag{15}$$

avec X = M,S. La valeur correspondante du facteur d'échelle n'étant pas toujours entière, nous arrondissons le facteur d'échelle à l'entier le plus proche.

Nous avons donc résolu le problème d'optimisation primal. Pour résoudre le problème dual qui nous intéresse directement, nous mettons en oeuvre une méthode similaire à une allocation de bits itérative : on initialise le seuil $T_{\min}(s)$ avec les seuil de masquage calculés sur les canaux L et R puis on résout le problème primal. Tant que le nombre de bits de codage dépasse la limite imposée par la contrainte de débit, on relève $T_{\min}(s)$ et on résout à nouveau le problème primal. On remarque qu'on n'applique l'algorithme qu'une seule fois pour les deux canaux. La répartition du débit entre les canaux est donc implicite avec cette méthode, et ne nécessite pas de critère spécifique.

4 Résultats

D'une part, de même que dans la section 2.3, nous mesurons la probabilité de respect du seuil de masquage après dématriçage avec le nouvel algorithme. Comme le montrent les résultats du tableau 2, le taux de respect du seuil est maintenant très proche de la valeur $\alpha = 0.95$ visée. Notre modèle permet donc une prise en compte satisfaisante du seuil de masquage.

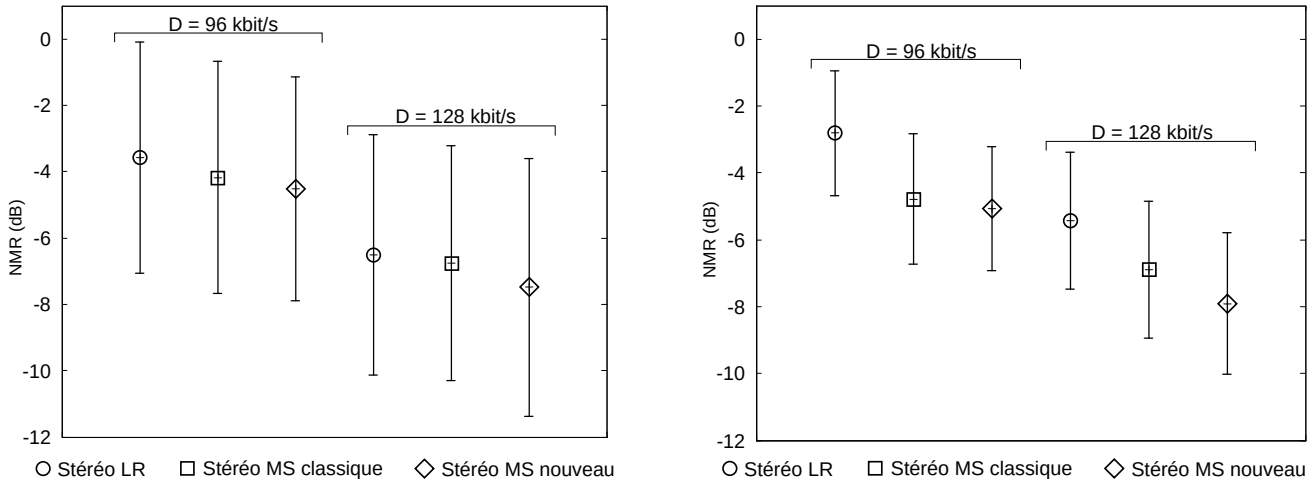

Figure 2: Mesures de NMR (valeur moyenne et intervalle à $+/-1\sigma$) sur les fichiers *Cocaine* (à gauche) et *Revolution* (à droite) pour trois modes stéréophoniques. Le débit est donné pour les deux canaux.

signal	canal L	canal R
Cocaine	0.94	0.95
Revolution	0.97	0.93

Table 2: Probabilité de respect du seuil de masquage après dématriçage, en mode MS, avec la nouvelle méthode.

D'autre part, nous réalisons le codage et le décodage de signaux audio réels, pour deux valeurs de débit classiques : 96 kbit/s et 128 kbit/s (pour les deux canaux). Des campagnes de tests d'écoute normalisés n'ayant pas pu être organisées, nous utilisons un critère d'évaluation objectif de qualité dit "NMR", proposé par K. Brandenburg *et al.* (voir [5]). La qualité est d'autant meilleure que le NMR est fortement négatif. Les résultats sont présentés en figure 2 pour deux fichiers audio de référence, échantillonnés à 48 kHz et d'une durée d'environ 6.5 s. Sur *Cocaine* (J.J. Cale, "Cocaine"), où la corrélation des signaux L et R est plutôt faible (coefficient de corrélation normalisé $c_{L,R} = 0.68$), le mode MS améliore légèrement la qualité par rapport au mode LR. Sur *Revolution* (T. Chapman, "Talkin' about revolution"), où la corrélation est forte (coefficient de corrélation normalisé $c_{L,R} = 0.93$), l'amélioration apportée par le mode MS est plus importante. Dans les deux cas, notre méthode produit un signal de meilleur qualité que la méthode standard. Le fait que l'amélioration soit plus importante à 128 kbit/s vient vraisemblablement de ce que l'hypothèse haute-résolution est plus souvent vérifiée à débit élevé.

5 Conclusion

Dans cet article, nous considérons le mode stéréo MS (stéréo matricée) du codeur MPEG AAC qui permet d'améliorer l'efficacité du codage par rapport au mode LR (double-mono), mais qui complique aussi la procédure d'optimisation. En effet, les signaux quantifiés ne sont pas directement écoutés par l'auditeur, d'où une inadéquation entre les résultats du modèle d'audition et le critère de distorsion utilisé lors de l'optimisation. La méthode classique n'est pas satisfaisante car le critère perceptuel du mode MS n'est pas cohérent avec celui du mode LR, qui est plus fiable. En outre, la répartition optimale du débit entre les canaux n'est pas assurée. Nous proposons alors de conserver le crière perceptuel du mode LR, et d'utiliser un algorithme d'optimisation différent incluant un modèle statistique de la quantification. Cette méthode permet d'améliorer la qualité sonore, au sens du critère de NMR moyen, ainsi que de résoudre naturellement le problème de la répartition de débit entre les canaux.

References

[1] International Organization for Standardization, *ISO/IEC 13818-7 (MPEG-2 Advanced Audio Coding, AAC)*, 1997.

[2] J. D. Johnston, "Transform Coding of Audio Signals Using Percetual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314–323, 1988.

[3] J. D. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria," *ICASSP 88*, May 1988.

[4] O. Derrien, P. Duhamel, and M.Charbit, "Statistical model for the quantization noise in the MPEG Advanced Audio Coder. Application to the bit allocation algorithm." *ICASSP 02*, May 2002.

[5] K. Brandenburg and T. Sporer, "NMR and Masking Flag : Evaluation of Quality Using Perceptual Criteria," in *Proceedings of the 11th International Conference of the AES*, pp. 169–179, 1992.