pp. 442-461

Le codeur MPEG-2 AAC expliqué aux traiteurs de signaux

Olivier DERRIEN*,
Sonia LARBI**,
Marcos PERREAU GUIMARES***,
Nicolas MOREAU*

Résumé

Descendant direct des codeurs audio MPEG-1, dont le MP3 est la figure emblématique, le MPEG-2 "Advanced Audio Coder" rassemble les techniques de compression les plus récentes et les plus efficaces. D'une architecture classique, il est construit autour d'une transformée en cosinus discrète à résolution variable. En sortie de ce banc de filtres, l'opération de compression proprement dite consiste en une allocation dynamique de bits par sous-bandes fréquentielles, associée à un module de codage entropique. L'allocation de bits est supervisée par un modèle psychoacoustique qui détermine le seuil d'audibilité des déformations subies par le signal dans le domaine fréquentiel. Cet article se veut un complément explicatif de la norme, mais aussi une introduction aux techniques de codage perceptuel de la musique.

Mots clés : Codage parole, Codage son, Compression bande passante, Exposé didactique, Psychoacoustique, Codeur, Normalisation, Architecture système, Transformation cosinus, Transformation discrète, Audition, Quantification signal, Code longueur variable.

THE MPEG-2 AAC CODER EXPLAINED TO THE SIGNAL PROCESSING EXPERTS

Abstract

The MPEG-2 Advanced Audio Coder is the latest issue of the MPEG audio encoders/decoders family, whose most popular version is known as MP3. It gathers many of the latest highly efficient sound compression techniques in a quite classically structured coder. The main part is based on a Discrete Cosine Transform with variable resolution. The output from this filterbank is compressed by the combination of an adaptive bit allocation module, according to frequency subbands, and a set of noiseless Huffman codebooks. Bit allocation is controlled by a psychoacoustic model which determines an audibility threshold for signal distortion in the frequency domain. This article intends to explain the ISO standard without replacing it, and also to be a general introduction to perceptual audio coding.

Keywords : Speech coding, Sound coding, Passband compression, Didactic paper, Psychoacoustics, Coder, Standardization, System architecture, Cosine transformation, Discrete transformation, Hearing, Signal quantization, Variable length code.

Sommaire

I. INTRODUCTION

Le but du codage audio est de réduire le débit du son numérique tout en préservant la qualité. Si le débit est une grandeur objective mesurable, il n'en va pas de même pour la qualité ; grandeur subjective par excellence, essentiellement évaluée par des tests d'écoute. Depuis le succès du disque compact (CD) dans le grand public, la référence en terme de qualité est le son dit « CD » : codage MIC (en anglais PCM)(sans compression) sur deux canaux indépendants, 16 bits par échantillon à une fréquence de 44.1 kHz. La dynamique et le rapport signal à bruit sont de l'ordre de 90 dB. Le débit net pour deux canaux est alors de 1.41 Mbit/s. On y ajoute un codage correcteur d'erreur, pour rendre le signal plus robuste aux dégradations physiques du disque, et enfin un codage par plages de bits, pour respecter certaines contraintes de la lecture par faisceau laser ; d'où un débit brut de 4.32 Mbit/s.

Derrien@tsi.enst.fr, slarbi@excite.com, perm@math-info.univ-paris5.fr, moreau@tsi.enst.fr
* ENST, 46 rue Barrault, 75634 Paris Cedex 13.
** ENIT, Laboratoire de Systèmes de Communications, BP 37 1002 TUNIS Belvédère.
*** Université Paris V, R. Descartes, 45 rue des Saint-Pères, 75270 Paris Cedex 06

Un tel débit est trop élevé pour des supports moins généreux que le CD classique. Il faut donc envisager de comprimer le signal, tout en garantissant une qualité équivalente. Les contraintes actuelles en terme débit (typiquement 64 à 128 kbit/s en stéréo, voire moins pour des applications à l'Internet) imposent l'emploi de *codeurs perceptuels*, qui utilisent les caractéristiques *psychoacoustiques* de l'oreille afin que la distorsion due au codage reste masquée par le signal de musique.

Les premiers codeurs audio étaient monophoniques. Aujourd'hui, l'application à la transmission du son dit « en bande hi-fi » justifie la quasi-généralisation de la stéréophonie. Cependant la récente application du codage aux bandes son de cinéma a imposé l'adoption de formats multicanaux, essentiellement le 5.1. Dans ce cas, le signal comporte cinq canaux large bande totalement indépendants (Avant Gauche et Droite, Centre et Arrière Gauche et Droite), plus un canal d'extrême grave, passe-bas.

Dans la famille des codeurs perceptuels, les MPEG-Audio occupent une place prépondérante. La description de ces codeurs est issue de l'organisation internationale de normalisation (ISO). Ils portent le nom du groupe de travail impliqué, le *Moving Pictures Expert Group*. Tous les codeurs MPEG-Audio sont donc des normes internationales, et leur mise en œuvre est libre de droits. Afin de laisser une porte ouverte aux améliorations futures, MPEG n'a pas normalisé entièrement ses codeurs. Seul le format du flux de bits est fixé. Cela impose la structure des décodeurs, et naturellement une partie de celle des codeurs. Pour ces derniers, la majorité de la réalisation reste libre.

La première génération de normes MPEG Audio date de 1992 [1, 2]. Ces codeurs, MPEG-1, se déclinent en trois versions : couche I, II et III, par ordre de complexité et d'efficacité croissante. Il est admis que la couche I permet un codage transparent à un débit de 192 kbit/s par canal, la couche II à 128 kbit/s par canal, et la couche III en dessous de 128 kbit/s par canal. Ce dernier codeur connaît aujourd'hui une très grande popularité sous le nom de MP3. Les codeurs MPEG-1 fonctionnent aux fréquences d'échantillonnage 32, 44,1 et 48 kHz. Ils prévoient des modes de traitement des configurations stéréophoniques. Des extensions non-normalisées permettent de traiter des fréquences d'échantillonnage plus basses.

En 1994, MPEG normalise des codeurs de seconde génération, dits MPEG-2, permettant cette fois de traiter les configurations multicanaux (au-delà de la stéréophonie), ou de fonctionner à des fréquences d'échantillonnage plus basses. Le premier codeur MPEG-2 multicanal est appelé MPEG-2 BC, pour *Backward Compatible*, car le flux binaire qu'il génère est décodable par un MPEG-1. Le MPEG-2 LSF, pour *Low Sampling Frequency*, admet les fréquences d'échantillonnage de 16, 22,05 et 24 kHz.

Ensuite, MPEG travaille à la normalisation d'un codeur multicanal de très haute qualité, plus efficace, mais non-compatible avec les décodeurs MPEG-1. Le MPEG-2 *Non Backward Compatible* est finalisé en 1997, et prend le nom de MPEG-2 AAC. Le document normatif correspondant est alors intitulé : IS 13818-7, *MPEG-2 Advanced Audio Coding, AAC* [3]. On note de nombreuses similitudes avec le codeur MPEG-1 couche III, dont AAC semble être le descendant direct. Ce codeur est annoncé comme transparent à un débit de 64 kbit/s, en moyenne, par canal. En effet, des tests menés au NHK, au Japon, et à la BBC à la fin de 1996 ont montré que le codeur AAC satisfait les critères de qualité de l'UIT à un débit total de 320 kbit/s sur 5 canaux. Les fréquences d'échantillonnage admises sont 8, 11,015, 12, 16, 22,05, 24, 32, 44,1, 48, 64, 88,2 et 96 kHz. Les échantillons du signal d'entrée doivent être représentés sur 16 bits, et le nombre maximal de canaux codables est de 48. Il s'agit d'un codeur modulaire proposant plusieurs options qui s'ajoutent à un noyau. Le codeur AAC a été intégré à la nouvelle norme MPEG-4, avec l'adoption de nouveaux modules optionnels permettant d'en augmenter encore l'efficacité [4, 5].

Dans cet article, nous présenterons en premier lieu les principes sur lesquels sont bâtis les codeurs perceptuels, ainsi que l'architecture générale du codeur MPEG-2 AAC. Nous présenterons ensuite les trois modules principaux, que l'on retrouve dans tous les codeurs perceptuels standards, à savoir la transformation temps-fréquence, la modélisation du fonctionnement de l'oreille et finalement l'opération de quantification et de codage proprement dit. Nous évoquerons enfin les quatre modules optionnels qui peuvent éventuellement compléter les trois principaux.

II. ARCHITECTURE DU CODEUR AAC

II.1. Principes du codage perceptuel

Pour introduire le codage perceptuel, commençons par décrire la représentation numérique la plus simple d'un signal audio : le codage PCM utilisé sur le disque compact. Le flux audio est constitué d'une succession d'échantillons à la cadence f_e, ou fréquence d'échantillonnage. L'amplitude des échantillons est discrétisée sur une échelle à L niveaux uniformément répartis. Il s'agit donc une *quantification scalaire uniforme*. Pour décrire plus précisément ce principe, on considère que le signale audio est à temps discret, mais *à amplitude réelle*. Il est noté $x(n)$. On quantifie alors l'amplitude sur $L = 2^b$ niveaux, et on obtient le signal $\hat{x}(n)$, à amplitude discrète. On peut donc représenter chaque échantillon par un mot de b bits, où b est appelé la *résolution*. L'erreur de quantification, signal à amplitude réelle, est définie ainsi :

$$q(n) = x(n) - \hat{x}(n) \quad (1)$$

Pour un signal $x(n)$ à amplitude normalisée (entre – 1 et + 1), $q(n)$ est toujours compris dans l'intervalle

$\left[-\frac{1}{L-1}, \frac{1}{L-1}\right]$. Lorsque L est élevé, cet intervalle est très étroit et on est dans l'hypothèse de *haute résolution*. Dans ces conditions, on peut considérer que la densité spectrale de $q(n)$ est uniforme (bruit blanc), et on caractérise ce signal par sa puissance σ_Q^2. Lorsque le signal $x(n)$ est stationnaire de puissance σ_X^2, la puissance de l'erreur de quantification s'écrit simplement :

$$\sigma_Q^2 = c\,\sigma_X^2\,2^{-2b} \tag{2}$$

où c est une constante valant environ 2,7 lorsque $x(n)$ est gaussien [6]. On en déduit le résultat classique suivant : Lorsqu'on augmente (resp. on diminue) la résolution b d'une unité, on divise (resp. on multiplie) la puissance de l'erreur d'un facteur 4, soit une diminution (resp. une augmentation) d'environ 6 dB par bit de codage. Pour le CD, on a fixé b à 16, ce qui garantit une valeur de σ_Q^2 suffisamment faible pour ne pas être gênante. Le rapport signal à bruit maximal est alors de 96 dB.

Cependant, un tel codage n'est optimal que lorsque x est un bruit blanc, ce qui n'est généralement pas le cas. Pour améliorer l'efficacité du codage, et donc chercher à *comprimer* le signal par rapport au codage PCM, on essaie de décorréler les échantillons du signal avant de les quantifier. Dans ce cas, le quantificateur travaille de façon optimale, et on peut baisser sa résolution sans dégradation audible. On transforme le signal x en un signal r par une opération de filtrage. Le signal r, que l'on va quantifier au lieu de x, est appelé *erreur de prédiction*. Le *gain de prédiction* est alors défini par :

$$G_p = \frac{\sigma_X^2}{\sigma_R^2} \tag{3}$$

La façon la plus simple d'obtenir r est d'appliquer à x le filtre $H(z) = 1 - A(z)$ où $A(z)$ est un filtre de prédiction linéaire. On réalise alors une *quantification scalaire prédictive*. Plus la prédiction est efficace, plus le gain de prédiction est élevé, et plus l'erreur de prédiction est blanche. Avec les hypothèses de l'équation 2, la puissance du bruit de quantification devient :

$$\sigma_Q^2 = c\,\frac{\sigma_X^2}{G_p}\,2^{-2b} \tag{4}$$

Malheureusement, la théorie du codage est très explicite quant aux limites de la prédiction : quelle que soit la méthode employée pour obtenir l'erreur de prédiction r, et quelle que soit la nature du signal x, G_p tend vers une asymptote. Pour un gain d'une cinquantaine de décibels, qui est déjà une valeur importante, l'économie sur la quantification est seulement de 8 bits par échantillons, soit un taux de compression de 2, ce qui reste médiocre.

Si on veut aller au delà, il faut faire appel à d'autres techniques. En particulier les codeurs dits *perceptuels* permettent actuellement les meilleures performances sur un signal de musique à haute qualité (proche du CD) [7, 8].

Tout d'abord, on constate que le rapport signal à bruit de quantification $\frac{\sigma_X^2}{\sigma_Q^2}$, n'est pas une bonne mesure de la qualité du signal codé. En effet, un rapport de 60 dB obtenu par quantification scalaire uniforme est généralement jugé inacceptable par l'oreille, alors que dans certaines conditions, un rapport de seulement 20 dB peut sembler imperceptible. La psychoacoustique nous apprend que c'est moins la puissance du bruit de quantification que sa *densité spectrale de puissance* (DSP) qui est pertinente à l'écoute. Dans un codeur à débit de sortie fixe, tels les MPEG Audio, le nombre *moyen* de bits de codage est fixé, ce qui impose une certaine valeur du rapport signal à bruit de quantification, typiquement de l'ordre de 20 dB. Mais le codeur peut mettre en forme spectralement ce bruit (donc le colorer) de telle sorte qu'il devienne inaudible. Pour connaître la DSP du bruit en limite d'audibilité, on s'appuie sur la propriété suivante : pour que le bruit de quantification soit *masqué* par le signal, on considère qu'il suffit que sa DSP $S_Q(f)$ reste inférieure à une courbe, appelée *seuil de masquage*, et notée $S_m(f)$, dépendant uniquement de la DSP du signal $S_X(f)$[1]. Pour calculer ce seuil, on utilise un *modèle d'audition*, aussi appelé *modèle psychoacoustique*.

Cependant, la mise en œuvre de ce principe dans un codeur de musique se heurte à des difficultés : Le signal n'est pas stationnaire, au mieux localement stationnaire, et des taux de compression élevés sont incompatibles avec l'hypothèse de haute résolution. Ces restrictions sont autant de contraintes sur la structure du codeur.

Comme les phénomènes de masquage sont essentiellement de nature spectrale, il faut exprimer le signal dans le domaine fréquentiel, à l'aide d'un banc de filtres ou d'une transformée (architectures théoriquement équivalentes). Le codeur réalise donc la partie *analyse fréquentielle* et le décodeur la *synthèse*. À la transformation fréquentielle à court terme utilisée, notée T, il doit exister une transformation réciproque, notée T^{-1}, telle que la succession de T et T^{-1} soit égale à l'identité (condition de *reconstruction parfaite*). Du choix de la longueur de T dépend le compromis entre résolution temporelle et résolution fréquentielle : Une transformée longue augmente la précision de la mise en forme spectrale du bruit de quantification, mais celui-ci sera étalé sur un horizon temporel plus long, sur lequel le signal risque de ne plus être stationnaire. Dans les codeurs MPEG-1, T est un banc de filtres PQMF, alors que dans AAC, il s'agit d'une transformée en cosinus discrète (MDCT). Parallèlement à T, le modèle psychoacoustique analyse le signal et calcule le seuil de masquage correspondant à la fenêtre temporelle courante. Enfin, le module de quantification alloue aux coefficients transformés des bits de codage, de telle sorte que la distorsion générée reste inférieure au seuil de masquage, condition couramment appelée *contrainte de masquage* [9]. Si le débit de sortie ne le permet pas, $S_Q(f)$ est rapprochée au maximum de $S_m(f)$. Mais dans ce cas, le

1. Dans l'hypothèse où on ne prend pas en compte le phénomène de masquage temporel. C'est le cas dans la majorité des codeurs audio.

codage n'est pas perpetuellement transparent, et des distorsions risquent d'être audibles.

Ensuite, les coefficients transformés quantifiés sont transmis au décodeur sous la forme d'un train binaire, selon une syntaxe très particulière constituant le cœur de la norme, mais que nous n'aborderons pas dans cet article. Le décodeur déchiffre la syntaxe du train de bits, puis restitue les niveaux de quantification initiaux des coefficients transformés, et enfin applique T^{-1} pour produire le signal reconstruit $\hat{x}(n)$.

On remarquera que des informations annexes relatives au nombre de bits de codage de chaque coefficient spectral, sont nécessaires lors du décodage. Ces informations doivent être explicitement transmises, ce qui consomme du débit. Afin de préserver un bilan de compression positif, le module de quantification alloue les bits à des groupes de coefficients spectraux adjacents, appelés *sous-bandes*. Dans la représentation en banc de filtres de T, elles sont interprétées comme les bandes passantes des filtres d'analyse. Les informations relatives au codage sont donc transmises pour chaque sous-bande.

II.2. Fonctionnement du codeur AAC

AAC est un cas particulier de codeur perceptuel tel que nous venons de le décrire. De structure modulaire, il comporte des modules indispensables au fonctionnement du système de base, formant le noyau, et d'autres optionnels, dont l'implémentation n'est pas obligatoire. En général, ils améliorent l'efficacité du codage au détriment de la charge de calcul.

Dans le schéma fonctionnel du codeur (voir figure 1), on distingue à gauche les modules d'analyse, et au centre les modules de traitement des données, commandés par les précédents. Les modules du noyau sont :

- MDCT. Il s'agit de la transformation spectrale utilisée. Ce cas particulier de banc de filtres modulés à reconstruction parfaite est ici appelé *Transformation en Cosinus Discrète Modifiée*. Cette transformation produit des coefficients réels. La particularité de AAC est d'utiliser deux tailles de transformée (2048 ou 256 points), pour ajuster le compromis entre résolution temporelle et résolution fréquentielle aux caractéristiques de stationnarité à court terme du signal. Le banc de filtres est alors commandé par un module d'analyse qui peut être incorporé au modèle psychoacoustique.
- **Modèle Psychoacoustique**. Il ne se contente pas de calculer le seuil de masquage. Outre la taille de la MDCT, il commande les premiers modules de traitement. De plus, en configuration multicanal, il analyse l'ensemble des voies en parallèle.
- **Facteurs d'échelle et Quantification**. Le codage et la quantification des coefficients spectraux dans AAC ne font pas appel à une allocation de bits classique. En premier lieu, on fait subir aux coefficients une transformation de changement d'échelle, analogue à une élévation à la puissance 4/3 suivie d'une division par une puissance entière de 2. Cette puissance entière, appelée *facteur d'échelle*, peut varier pour chaque sous-bande. Ensuite, les *coefficients spectraux quantifiés* sont obtenus par application d'une quantification scalaire uniforme de pas 1, qui prend la forme de l'opération d'arrondi à l'entier le plus proche. Cette opération a pour effet de réduire fortement la dynamique des coefficients spectraux (on passe typiquement de 3.10^4 à 30). Il est donc nécessaire de transmettre conjointement ces coefficients et les facteurs d'échelle, qui sont tous des entiers. L'algorithme de quantification consiste alors à déterminer les meilleures valeurs des facteurs d'échelles, en essayant de respecter simultanément la contrainte de débit et la contrainte psychoacoustique. Il est itératif et met en jeu les fonctions suivantes :
  - **Contrôle de débit/distorsion.** Ce module évalue la puissance du bruit de quantification par décodage local, la compare avec le seuil de masquage fourni par le modèle psychoacoustique, et modifie les facteurs d'échelle de l'itération suivante pour essayer de satisfaire la contrainte psychoacoustique selon le débit disponible. Il s'agit donc d'un processus de codage par *analyse-synthèse*.
  - **Codage entropique.** Il transforme les coefficients quantifiés au moyen d'un code de Huffman dont les dictionnaires sont définis par la norme. Ce codage est sans bruit et permet de diminuer encore le volume des données. La forme codée des échantillons quantifiés doit être prise en compte pour évaluer le nombre de bits utilisés et l'ajuster au débit disponible.
  - **Multiplex**. Pour terminer, un multiplexeur met en forme les coefficients spectraux codés, les facteurs d'échelle, et les informations de contrôle provenant des autres modules, dans des trames dont la succession forme le flux de bits normalisé.

Ensuite, le codeur peut utiliser des modules optionnels. Le principal est sans doute le **Matriçage M/S**. Cette option de traitement stéréophonique consiste en une opération de matriçage des canaux par paires, et suppose une variante du modèle psychoacoustique. Ce module, ainsi que les trois autres modules optionnels, seront succinctement évoqués dans la section 6.

Le décodeur est construit sur un modèle réciproque, sans toutefois inclure les modules d'analyse, ici superflus (voir Fig. 1).

Conformément à l'esprit des normes MPEG, seul le décodeur est entièrement normalisé. Dans le codeur, les seules réalisations imposées concernent la MDCT, le matriçage M/S, les valeurs et la représentation des facteurs d'échelle, la forme des coefficients quantifiés et les tables de Huffman. Le modèle psychoacoustique et le contrôle de débit/distorsion sont libres. La description

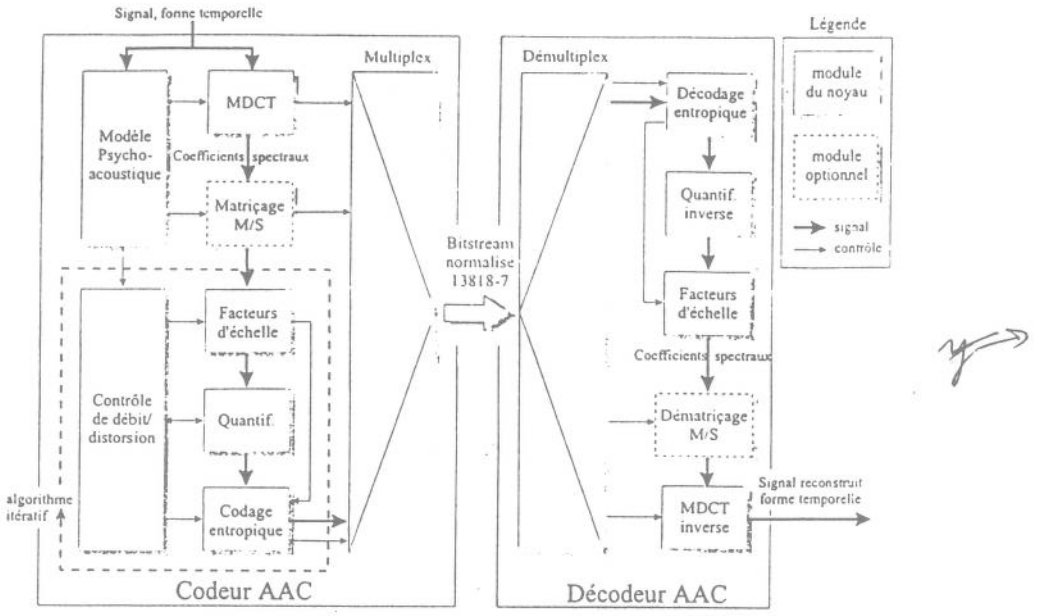

FIG. 1. — Schéma fonctionnel du codeur et du décodeur AAC.

AAC encoder/decoder block diagram.

qui va suivre est alors issue de l'annexe informative de la norme, qui propose des solutions d'implémentation.

III. LA TRANSFORMÉE EN COSINUS DISCRÈTE MODIFIÉE

Il existe un formalisme unique permettant de décrire un banc de filtre ou une transformée, bien que chacune de ces deux architectures puisse donner lieu à une implémentation différente. Nous allons commencer par introduire la description générale d'un banc de filtres, puis la transformée correspondante, avant de formuler la MDCT.

III.1. Banc de filtres

Un banc de filtres est généralement constitué de M filtres passe-bande, chacun correspondant à une *sous-bande fréquentielle* (voir Fig. 2). Ces filtres sont toujours supposés *à réponse impulsionnelle finie* (RIF), de longueur N. Les réponses impulsionnelles des filtres d'analyse et de synthèse sont notées respectivement $h_k(n)$ et $f_k(n)$, k est l'indice de sous-bande et n l'indice temporel. Le facteur de sur et sous-échantillonnage est noté M'. Il doit être inférieur ou égal à M, sinon il y a perte d'information et le banc de filtres n'est plus à reconstruction parfaite. Lorsque $M'=M$, le banc de filtres est dit *à échantillonnage critique* ou *à décimation maximale*. En codage, on se place presque exclusivement dans cette situation, car elle réalise un bon compromis entre quantité d'informations et débit total dans les sous-bandes. Par la suite, on supposera donc que $M'=M$.

Les signaux en sortie de banc de filtres d'analyse s'écrivent :

$$(5) \qquad y_k(m)=u_k(mM)=\sum_{l=0}^{N-1} h_k(l)\,x(mM-l)$$

Les signaux $y_k(m)$ sont émis, et on reçoit les $\hat{y}_k(m)$. Si on note $q_0=\lceil \frac{N}{M} \rceil$, et si on complète $f_k(n)$ par des zéros lorsque n dépasse $\{0\cdots N-1\}$, le signal reconstruit en sortie du banc de filtres s'écrit :

$$(6) \qquad \hat{x}(mM+p)=\sum_{k=0}^{M-1} \hat{v}_k(mM+p)$$

$$=\sum_{k=0}^{M-1}\sum_{q=0}^{q_0-1} f_k(qM+p)\,\hat{y}_k(m-q)$$

III.2. Notations matricielles et transformée

Les formules d'analyse (5) et de synthèse (6) sont de nature convolutive. Or l'association des sous et sur-échantillonneurs aux filtres rend ces formules équivalentes à des opérations en bloc, représentables par des multiplications matricielles.

Pour le filtrage d'analyse, on suppose $N \geq M$. Le recouvrement temporel des vecteurs de signal permet alors d'éviter les « effets de bloc » (discontinuités lors de la concaténation de blocs de signal reconstruit). On supposera aussi que $N=q_0M$, où q_0 est un entier.

Un vecteur de signal comprenant N échantillons successifs, et dont le dernier est $x(mM)$ s'écrit :

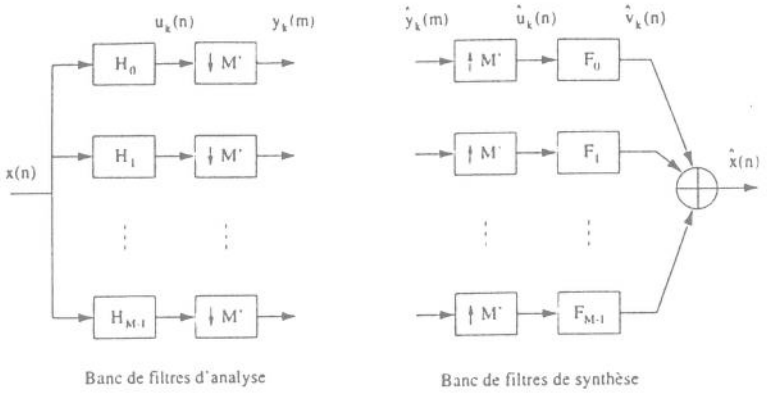

FIG. 2. — Structure générale d'un banc de filtres.

Filterbank usual structure.

$$(7)\qquad \underline{X}(m)=(\ (x((m-q_0)\,M-1)\ \ldots\ x(mM)\ )^t$$

Entre deux vecteurs correspondant à des indices m successifs, on note un recouvrement d'un facteur $\frac{q_0-1}{q_0}$.

Les échantillons en sortie des filtres d'analyse s'expriment aussi sous forme vectorielle, de telle sorte que chaque composante correspondante à une sous-bande :

$$(8)\qquad \underline{Y}(m)=(\ y_0(m)\ \ y_1(m)\ \ \ldots\ \ y_{M-1}(m)\ )^t$$

L'ensemble des réponses impulsionnelles des filtres d'analyse est regroupé dans une seule matrice de taille $M\times N$:

$$(9)\qquad \mathbf{H}=\begin{pmatrix} h_0(N-1) & \ldots & h_0(0) \\ \vdots & & \vdots \\ h_{M-1}(N-1) & \ldots & h_{M-1}(0) \end{pmatrix}$$

Alors, l'opération d'analyse s'écrit simplement sous la forme :

$$(10)\qquad \underline{Y}(m)=\mathbf{H}\,\underline{X}(m)$$

Cette opération est équivalente à l'équation (5).

Pour le filtrage de synthèse, on adopte des notations similaires. Le vecteur des coefficients transformés reçu est noté :

$$(11)\qquad \hat{\underline{Y}}(m)=(\ \hat{y}_0(m)\ \ \hat{y}_1(m)\ \ \ldots\ \ \hat{y}_{M-1}(m)\ )^t$$

On note $\mathbf{F}$ la matrice des réponses impulsionnelles des filtres de synthèse, de taille $M\times N$:

$$(12)\qquad \mathbf{F}=\begin{pmatrix} f_0(0) & \ldots & f_0(N-1) \\ \vdots & & \vdots \\ f_{M-1}(0) & \ldots & f_{M-1}(N-1) \end{pmatrix}$$

La synthèse se décompose en deux opérations. La première est la multiplication matricielle suivante :

$$(13)\qquad \underline{Z}(m)=\mathbf{F}^t\,\hat{\underline{Y}}(m).$$

La seconde opération est une addition de blocs décalés (*add-overlap*) des vecteurs $\underline{Z}(m)$. Si on note $z_0(m)$ à $z_{q_0-1}(m)$ les blocs de taille M composant le vecteur $\underline{Z}(m)$:

$$(14)\qquad \underline{Z}(m)=(\ \underline{z}_0(m)\ \ \underline{z}_1(m)\ \ \ldots\ \ \underline{z}_{M-1}(m)\ )^t$$

Alors on obtient le bloc de signal reconstruit suivant par l'opération :

$$(15)\qquad \hat{\underline{x}}(m)=\sum_{q=0}^{q_0-1}\underline{z}_q(m-q)$$

où la somme est vectorielle (addition composante par composante). Pour retrouver le signal $\hat{x}(n)$, il suffit de concaténer les blocs $\hat{\underline{x}}(m)$ dans l'ordre des m croissants. L'ensemble des opérations de synthèse et équivalente à l'équation (6).

III.3. La MDCT du codeur MPEG-2 AAC

AAC utilise une transformation particulière appartenant à la famille des *bancs de filtres modulés à reconstruction parfaite* [10]. Elle est aussi appelée TDAC, pour *Time Domain Aliasing Cancellation* [11] ou MTL pour *Modulated Lapped Transform* [12].

Dans AAC, on la nomme MDCT pour *Modified Discrete Cosine Transform*. Comme toutes les transformées en cosinus, elle a une interprétation spectrale [13]. On choisit $N=2M$, d'où $q_0=2$. La particularité des bancs de filtres modulés est d'utiliser des matrices de filtres d'analyse et de synthèse identiques, chaque réponse impulsionnelle s'écrivant de plus comme le produit terme à terme d'un *filtre prototype*, aussi appelé *fenêtre de MDCT*, par les coefficients d'une *matrice de modulation* :

$$(16)\qquad \mathbf{H}=\mathbf{F}=(h(n)\,d_k(n))_{k\in\{0\ldots M-1\},\,n\in\{0\ldots 2M-1\}}$$

où $h(n)$ est la réponse impulsionnelle du filtre prototype, et $\mathbf{D}=(d_k(n))$ la matrice de modulation, de taille $M\times 2M$.

On peut alors simplifier les équations décrivant l'analyse et la synthèse en notant $\otimes$ le produit de deux vecteurs composante et $\underline{h}$ le vecteur des coefficients du filtre prototype. L'équation (10) devient :

$$\underline{Y}(m) = \mathbf{D}\ (\underline{X}(m) \otimes \underline{h}) \tag{17}$$

L'équation (13) devient :

$$\underline{Z}(m) = \underline{h} \otimes (\mathbf{D}\,\underline{\hat{Y}}(m)) \tag{18}$$

et l'équation (15) devient :

$$\underline{\hat{x}}(m) = \underline{z}_0(m) + \underline{z}_1(m-1) \tag{19}$$

Le principe de la MDCT directe et inverse est schématisé en figures 3 et 4. Les coefficients de la matrice de modulation sont des cosinus.

$$d_k(n) = \sqrt{\frac{2}{M}} \cos\left[(2k+1)\,(2n+1+M)\,\frac{\pi}{4M}\right] \tag{20}$$

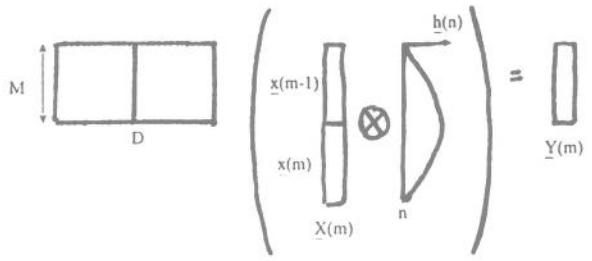

FIG. 3. — MDCT, partie analyse, pour $N=2M$. Le produit terme à terme, $\otimes$, avec le filtre prototype $\underline{h}$ est effectué avant le produit matriciel.

MDCT analysis section, $N=2M$. The element-by-element multiplication $\otimes$ with prototype filter must be computed before matrix multiplication.

On montre que les conditions de reconstruction parfaite portent uniquement sur le filtre prototype. On doit avoir :

$$\forall n \in \{0 \dots M-1\},\quad \begin{cases} h^2(n) + h^2(n+M) = 1 \\ h(n)h(M-1-n) = h(n+M)h(2M-1-n) \end{cases} \tag{21}$$

En général, on impose une propriété de symétrie temporelle au filtre prototype :

$$\forall n \in \{0 \dots 2M-1\},\quad h(n) = h(2M-1-n) \tag{22}$$

Dans ce cas, la seconde condition de l'équation (21) est vérifiée. Pour satisfaire la première condition, on choisit :

$$h(n) = \sin\left[(2n+1)\,\frac{\pi}{4M}\right] \tag{23}$$

AAC est donc construit autour d'un banc de filtres à reconstruction parfaite. Cependant, le signal après synthèse n'est identique au signal d'origine que si on ne modifie pas les coefficients transformés entre l'analyse et la synthèse. Une éventuelle modification peut être vue comme une opération de filtrage par sous-bande, variant dans le temps. Suivant la valeur du taux de recouvrement et la longueur de la réponse impulsionnelle équivalente de ces filtres, on peut observer des phénomènes de repliement temporel et fréquentiel. Ces phénomènes sont d'autant plus perceptibles que le taux de recouvrement est faible et que les modifications sont brutales d'un bloc à l'autre. C'est une caractéristique commune à tous les bancs de filtres à sous-échantillonnage, mais qui reste très complexe à formaliser [14].

En codage audio, on quantifie les coefficients transformés, ce qui génère une distorsion, elle-même potentiellement gênante. Mais si la distorsion affectant un ensemble de coefficients transformés est trop brutale d'un bloc à l'autre, des artéfacts gênants semblables à des phénomènes de battement (analogue à une modulation d'amplitude du signal) risquent d'apparaître et dégrader encore la qualité sonore perçue. Ces phénomènes sont d'autant moins maîtrisables que leurs condi-

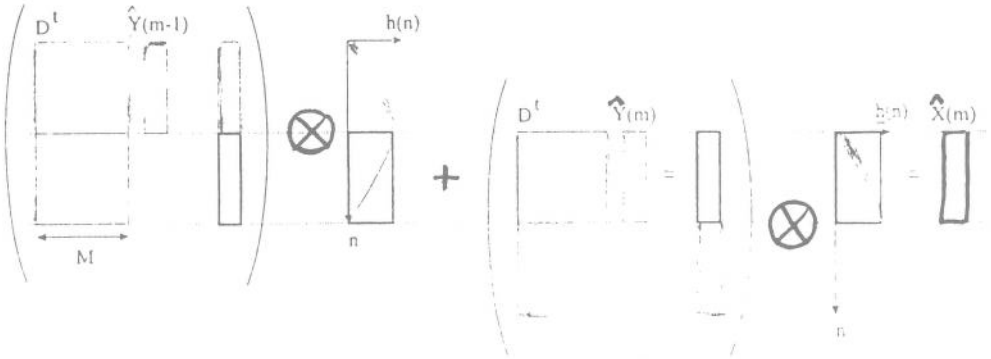

FIG. 4. — MDCT, partie synthèse, pour $N=2M$. Le produit terme, $\otimes$, sont effectués sur les résultats des produits matriciels, puis la somme vectorielle s'applique aux blocs entourés en gras.

MDCT synthesis section, $N=2M$. The element-by-element multiplication $\otimes$ must operate on the result of matrix multiplication, then vector sum can operate on bold line surrounded blocks.

tions d'apparition sont difficiles à caractériser. Malheureusement, la MDCT de AAC est à recouvrement minimal (50 %), ce qui signifie que ce codeur est potentiellement sensible à ce genre de défaut, surtout à bas débit.

III.4 Changement de la taille des fenêtres

Dans un codeur perceptuel, le choix de la longueur de la transformée (paramètre M) influe directement sur la qualité du signal reconstruit : Une transformée courte a de bonnes performances en terme de résolution temporelle, au détriment de la résolution fréquentielle. Dans ce cas, les spectres comportent un petit nombre de composantes, et le timbre des signaux à fort caractère tonal sera moins bien rendu. Inversement, une transformée longue n'aura pas ce genre de défaut, mais sera caractérisée par une moins bonne résolution temporelle. En pratique, cela se traduit par un phénomène perceptuel que l'on nomme couramment *pré-écho :* Le bruit qu'engendre la quantification d'un spectre étant temporellement réparti sur l'ensemble de la longueur de la fenêtre à la reconstruction, une attaque (augmentation brutale du niveau du signal au milieu d'une fenêtre), pourra être perçu avec de l'avance.

Pour le codeur AAC, on a le choix entre deux valeurs de M, avec la possibilité de passer dynamiquement de l'une à l'autre pour s'adapter aux caractéristiques du signal. Chacune correspond à une taille de fenêtre différente :

– 2M = 2048, fenêtre longue
– 2M = 256, fenêtre courte

Le passage d'une taille de fenêtre à l'autre doit conserver la propriété de reconstruction parfaite, malgré l'impossibilité de garantir un recouvrement de 50 % entre des fenêtres successives. Cette situation particulière nécessite l'utilisation de fenêtres (i.e. de filtres prototypes) dites de *transition.*

Lors du passage d'une fenêtre longue à une fenêtre courte, les recouvrements sont définis dans la norme, et sont illustrés par la figure 5. On utilise les fenêtres courtes $h_{256}(n)$ pour tous les vecteurs de signal de longueur 256, en association avec la matrice de modulation $\mathbf{D}_{256}$. On applique une fenêtre de transition, notée ici $h_{2048\to256}(n)$, au vecteur $\underline{X}(m)$, en association avec la matrice de modulation $\mathbf{D}_{2048}$. On admet que la condition de reconstruction parfaite reste valide lors du changement de taille de fenêtre, bien que cette affirmation n'aille pas de soi. En effet, on change de matrice de modulation.

On décompose $\underline{X}(m)$ en quatre sections, pour lesquelles la condition de reconstruction parfaite impose :

$$h_{2048\to256}(n)=h_{2048}(n) \quad \text{pour } n \in \{0 \dots 1023\}$$
$$h_{2048\to256}(n)=1 \quad \text{pour } n \in \{1024 \dots 1471\}$$
$$h_{2048\to256}(n)=h_{256}(n-704) \quad \text{pour } n \in \{1472 \dots 1599\}$$
$$h_{2048\to256}(n)=0 \quad \text{pour } n \in \{1600 \dots 2047\}$$

La forme temporelle de la fenêtre $h_{2048\to256}(n)$ ainsi définie est représentée en figure 6.

Lors du passage de $M=128$ à $M=1024$, on retrouve

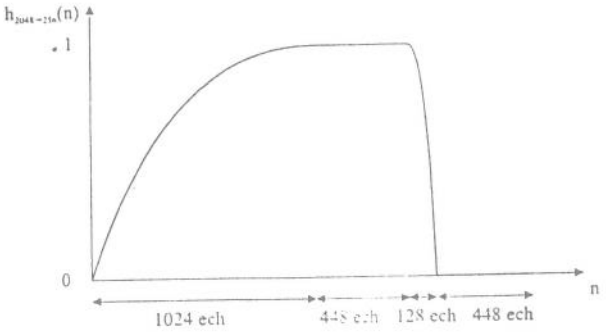

FIG. 6. — Représentation temporelle de la fenêtre de MDCT $h_{2048\to256}(n)$.

Temporal view of $h_{2048\to256}(n)$ MDCT window.

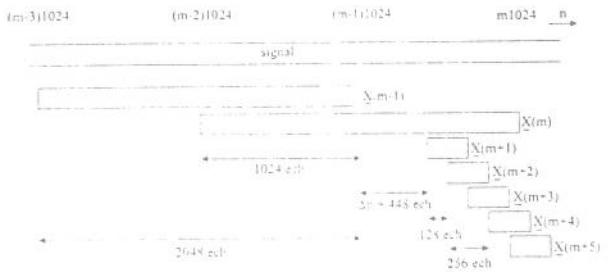

FIG. 5. — Recouvrement d'une fenêtre longue par des fenêtres courtes.

One long window overlapping short windows.

le problème symétrique. La fenêtre adéquate est alors simplement la version symétrique de celle précédemment décrite :

$$(24) \qquad h_{2048\to256}(n)=h_{2048\to256}(2047-n)$$

On montre que la condition de reconstruction parfaite reste vraie quelles que soient les fenêtres choisies, à condition d'aligner au moins 8 fenêtres courtes successives entre deux fenêtres de transition, situation illustrée à la figure 7. Pour satisfaire cette contrainte, ainsi que pour des raisons de simplicité des structures internes du codeur, AAC manipule les vecteurs de signal correspondant aux fenêtres courtes par bloc de huit vecteurs successifs. Cela donne un ensemble de 2048 échantillons, soit la même taille qu'un seul vecteur correspondant à

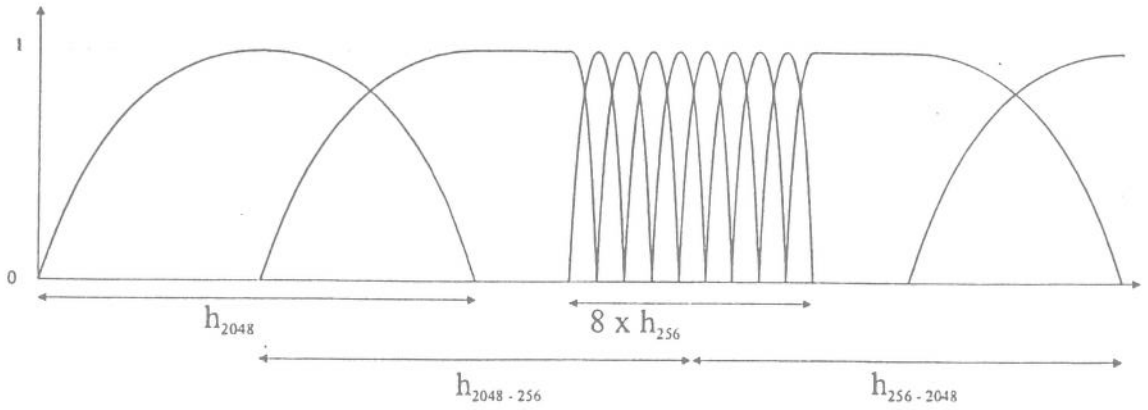

FIG. 7. — Agencement des fenêtres de MDCT lors d'un changement de taille de fenêtre.

Windows overlap during window-size variations.

une fenêtre longue. On parle alors de *bloc de signal*, qu'on qualifie de *bloc long* s'il correspond à un vecteur long composé de 2048 échantillons consécutifs, ou de *bloc court* s'il correspond à 8 vecteurs courts se recouvrant à 50 %, composés chacun de 256 échantillons consécutifs.

IV. MODÈLE PSYCHOACOUSTIQUE

Le modèle psychoacoustique a pour but de déterminer le seuil de masquage dans les sous-bandes de codage. De plus, comme il a été annoncé dans la section II. 2, le modèle psychoacoustique détermine aussi la taille de transformée (paramètre N). Dans cette section, nous ne présenterons que le modèle monophonique. L'option stéréo sera abordée en annexe.

IV.1. Le phénomène de masquage fréquentiel et sa modélisation

Nous allons maintenant introduire le *masquage fréquentiel*, phénomène de nature *psychoacoustique* sur lequel s'appuient tous les codeurs perceptuels, et AAC en particulier. Ce masquage reflète le fonctionnement mécanique et nerveux de l'oreille. On cherche à le modéliser de la façon la plus fine et la plus réaliste possible dans les codeurs pour déterminer la distorsion maximale que l'on va pouvoir introduire sans qu'elle soit perçue.

Les expériences fondamentales mettant en évidence le masquage sont depuis longtemps couramment utilisées dans le domaine du codage audio. Les modélisations qui en découlent sont fiables pour des signaux élémentaires (son pur, bruit blanc ou bruit à bande étroite). La généralisation à des signaux complexes[2] reste très difficile, tant sur le plan expérimental que sur le plan théorique.

Il existe plusieurs approches, donnant des résultats comparables, entre lesquelles les différences sont très fines, mais néanmoins souvent significatives. Nous avons choisi de présenter uniquement l'approche reprise dans le codeur AAC. Pour plus de précisions, nous renvoyons aux ouvrages de Zwicker [15] et de Moore [16] qui font autorité en la matière. Cependant, les ouvrages de Green [17] et Botte [18] proposent une approche du domaine plus globale et plus accessible. Cette section est essentiellement issue de ces références.

IV.1.1. Niveau de pression acoustique

Le signal physique perçu par l'oreille est la *pression acoustique*, petites variations de la pression instantanée autour de la pression atmosphérique. En conditions d'écoute *monaurales*, c'est-à-dire lorsque le son n'est perçu que par une seule oreille, ou lorsque les deux oreilles perçoivent des sons identiques, nous percevons essentiellement la puissance de ce signal. Exprimé en décibels par rapport à la plus petite puissance audible, on obtient le *niveau de pression acoustique*, ou SPL[3]. Cette grandeur est utilisée pour mesurer l'intensité sonore dans les expériences psychoacoustiques.

IV.1.2 Seuil d'audition absolu

Dans le cas de sons purs, la mesure du plus petit SPL audible en fonction de la fréquence donne le *seuil d'audition absolu*, ou MAF[4]. Cette courbe, représentée en bas de la figure 8, n'est pas plate : On constate que

2. Il ne s'agit pas de signaux à valeurs dans l'espace des nombre complexes, mais de signaux qui ne sont pas élémentaires.
3. Sound Pressure Level.
4. Minimum Audible Field.

l'oreille est nettement moins sensible aux limites du spectre audible. Sur cette figure, on a aussi représenté en haut le *seuil de douleur*. On notera toutefois que des dégradations irréversibles apparaissent bien avant cette limite. L'espace entre ces deux courbes est appelé *domaine d'audition*. On constate que si la dynamique maximale de la perception auditive peut dépasser les 120 dB, la dynamique utile est plus proche des 90 dB.

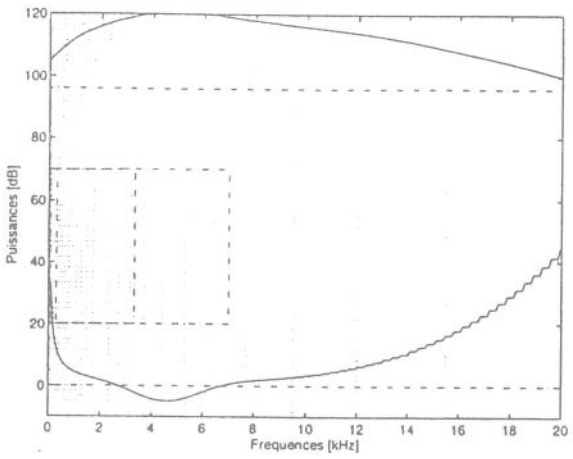

FIG. 8. — Domaine d'audition. La puissance est exprimée par le SPL en dB.

Audibility domain. Power range is scaled with SPL in dB.

IV.1.3. Bandes critiques

Des expériences attribuées en premier à Zwicker ont montré que l'oreille se comporte un peu comme un banc de filtres : Elle intègre la puissance du signal sonore sur des bandes fréquentielles. En d'autres termes, deux raies spectrales (deux sons purs par exemple) seront perçues comme une seule raie, de puissance égale à la somme des puissances des deux raies, si leurs fréquences sont espacées d'une valeur inférieure à Δf autour de leur fréquence médiane f : La variation de Δf en fonction de f, représentée sur la figure 9 (courbe continue), est appelée *largeur de bande critique*.

Bien que l'oreille puisse former une bande critique autour de n'importe quelle fréquence, on considère souvent d'une façon un peu artificielle, que l'espace des fréquences est partitionné en 24 bandes aux frontières fixes, comme sur la figure 9. L'échelle fréquentielle des *barks* est liée à ces bandes : la fréquence médiane de la b-ième bande critique vaut b Barks. Il est donc légitime d'intégrer le spectre du signal sur les bandes critiques pour modéliser la perception auditive.

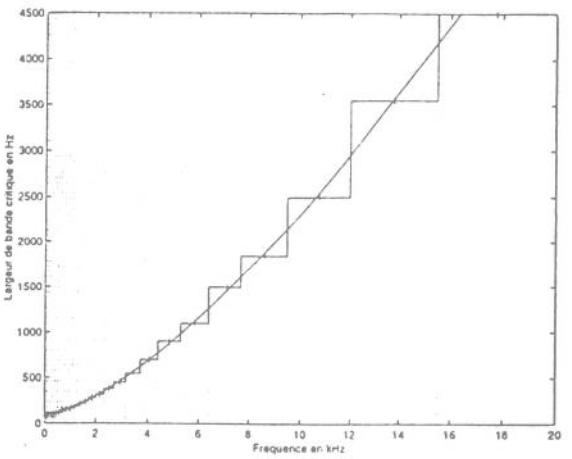

FIG. 9. — Largeur des bandes critiques en fonction de la fréquence, et partition de l'axe des fréquence par juxtaposition de 24 bandes critiques.

Critical bandwidth as a function of frequency, and frequency domain partition according to 24 critical bands.

IV.2. Courbe de masquage

Pour décrire le masquage fréquentiel, on considère souvent un son élémentaire, dit *masquant*, de fréquence f_0 (soit un son pur à la fréquence f_0, soit un bruit à bande étroite autour de f_0), et de puissance σ_0^2. On y superpose un autre son élémentaire, dit *masqué*, de caractéristiques f et σ^2. Pour tout f, il existe une valeur limite de σ^2 en dessous de laquelle la perception de la somme du son masquant et du son masqué est identique à la perception du son masquant seul. La variation de cette valeur limite de σ^2 en fonction de f, pour f_0 et σ_0^2 donnés, est appelée *courbe de masquage* du son masquant. Cette courbe est représentée en figure 10 pour f_0=1 kHz, et pour plusieurs valeurs de σ_0^2. On observe d'une part que lorsque f s'éloigne suffisamment de f_0, on retrouve le seuil d'audition absolu. D'autre part, autour de f_0, la courbe a une forme globalement triangulaire. Le rapport de puissance σ_0^2/σ^2 pour $f=f_0$ (à la pointe) est appelé *indice de masquage*.

En première approximation, on peut considérer que les pentes de la partie triangulaire sont indépendantes de σ_0^2 et de f_0, si les fréquences sont exprimées en Barks et les puissances en dB. En codage audio, on considère que l'indice de masquage dépend seulement[5] de la nature plus ou moins tonale du masquant : d'environ 20 dB pour un son pur, il passe à environ 5 dB pour un bruit à bande étroite. Le pouvoir masquant d'un bruit est donc supérieur à celui d'un son pur.

IV.2.1. Seuil de masquage

Dans le cas d'un signal complexe, on ne dispose pas d'expériences psychoacoustiques directement exploi-

5. En psychoacoustique, on montre qu'il dépend aussi de la nature tonale du son masqué, caractéristique inaccessible en codage audio.

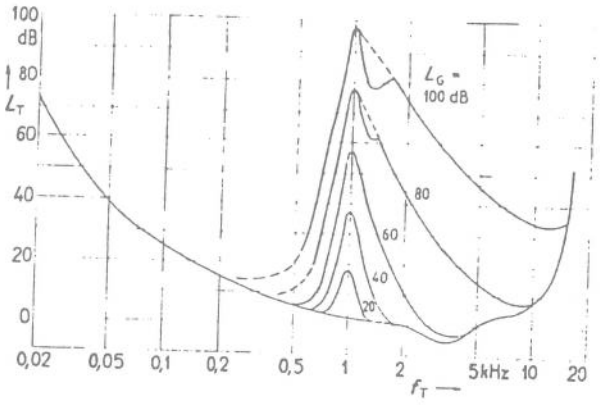

FIG. 10. — Seuil d'audition absolu et courbes de masquage d'un son pur par un bruit à bande étroite autours de 1 kHz, pour plusieurs puissances différentes du bruit masquant. D'après [15] p.57.

Minimum audible field and masking curves corresponding to a tone-masking-noise situation around 1kHz, for different values of noise power.

tables pour le codage. Néanmoins, on admet qu'en combinant toutes les courbes de masquage des raies spectrales du signal (intégrées par bandes critiques ou par sous-bandes fréquentielles), on obtient le *seuil de masquage*. Il est défini comme l'énergie par sous-bande en limite d'audibilité du signal masqué, et on l'assimile à la plus haute DSP d'un bruit large bande qu'on peut ajouter au signal sans qu'il soit perçu.

Donc pour chaque raie spectrale, on additionne les contributions des courbes de masquage de toutes les autres raies. Cependant, le choix d'une loi d'addition n'est pas simple [19]. Souvent, comme c'est le cas dans AAC, on utilise une loi d'addition classique, qu'on implémente sous la forme d'une *convolution* : Le spectre du signal est convolué par une *fonction d'étalement* représentant la partie triangulaire des courbes de masquage, normalisée en amplitude (toujours avec les fréquences exprimées en Barks). Ensuite, il reste à appliquer l'indice de masquage selon le caractère tonal ou non de chaque raie du spectre. Dans AAC, on considère une variation continue de l'indice de masquage entre le cas tonal et non-tonal. Cela suppose un estimateur de la tonalité du signal.

En conclusion, on peut dire que le phénomène de masquage fréquentiel est assez bien maîtrisé dans des cas simples. Lorsqu'on cherche à le modéliser dans des cas plus complexes, des difficultés apparaissent. Il n'existe alors pas de modèle psychoacoustique unique et irréfutable, mais un très grand nombre de modèles différents, dont la mise au point dépend de l'application. Chaque codeur audio fait un choix qui lui est propre, bien que les solutions apportées soient généralement très similaires d'un codeur à l'autre.

IV.3. Calcul du seuil de masquage dans le codeur AAC

Comme nous venons de l'évoquer dans la section précédente, tout modèle psychoacoustique doit effectuer en premier lieu une analyse spectrale du signal. Dans le codeur AAC, on travaille avec les mêmes blocs que la MDCT, mais la transformation spectrale est cette fois une DFT. On supposera pour l'instant que le paramètre N (taille des blocs de signal) vaut 2048. Les modifications pour le cas $N=256$ seront abordées dans la section 4.5.

Le signal audio sur la fenêtre courante de taille N est noté $x(n)$, où n est un indice temporel appartenant à $\{0 \cdots N-1\}$.

L'estimation spectrale consiste en un fenêtrage (fenêtre de Hann) suivi d'une DFT :

$$X(k)=DFT\left\{\frac{x(n)}{2}\left[1-\cos\left(\frac{2\pi(n+\frac{1}{2})}{N}\right)\right]\right\} \quad (25)$$

où k est un indice fréquentiel appartenant à $\{0 \cdots N-1\}$.

Le signal $x(n)$ est réel. $X(k)$ possède donc la propriété de symétrie hermitienne. Comme on ne s'intéresse qu'à la distribution de la puissance en fonction de la fréquence, il suffit de restreindre la variable k à l'intervalle $\{0 \cdots \frac{N}{2}\}$.

On obtient une estimation de la densité spectrale de puissance du signal, ou spectre, avec la formule du périodogramme :

$$S(k)=\frac{1}{N}\,|X(k)|^2 \quad (26)$$

La figure 11 donne un exemple de signal temporel $x(n)$, avec la fenêtre de Hann qui lui est appliquée et le spectre de Fourier correspondant. Cette portion de signal, échantillonné à 32 kHz, servira d'exemple pour les différentes étapes du modèle psychoacoustique.

L'indice fréquentiel k représente une fréquence en échelle linéaire. Pour calculer le seuil de masquage, il faut une échelle fréquentielle en Barks. Le modèle psychoacoustique de AAC ne travaille pas exactement suivant les bandes critiques, mais utilise une échelle plus fine que les Barks : les *partitions*, notées avec l'indice p. Une partition correspond à peu près à un tiers de bande critique.

On note $k^p_{min}(p)$ et $k^p_{max}(p)$ les indices fréquentiels représentant les limites inférieure et supérieure de la partition p. Le spectre integré selon une échelle fréquentielle perceptuelle prend le nom de *spectre basilaire*[6]. Dans AAC, on calcule l'énergie par partition, ce qui est sensiblement équivalent :

$$e^p(p)=\sum_{k=k^p_{min}(p)}^{k^p_{max}(p)} S(k) \quad (27)$$

L'exposant p signifie que ces grandeurs sont définies pour des partitions. L'énergie par partition est représentée en pointillés sur la figure 13.

6. Zwicker définit le spectre basilaire comme une puissance par sous-bande, ce n'est donc pas un spectre au sens du traitement du signal. Dans AAC, la grandeur calculée est homogène à une énergie.

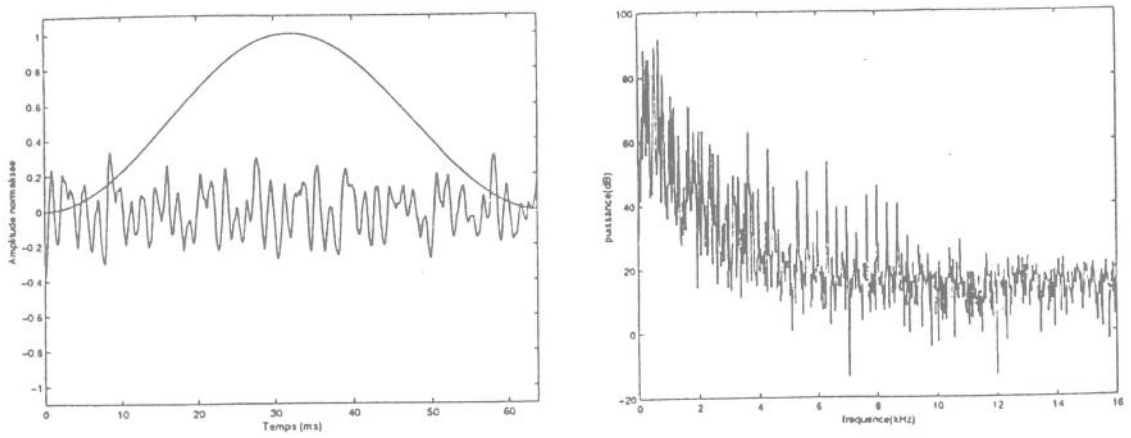

FIG. 11. — Exemple de signal (guitare classique). À gauche, forme temporelle avec fenêtre de Hann (l'amplitude est normalisée par rapport à la pleine échelle). À droite, spectre de Fourier correspondant.

A signal example (Spanish guitar). On the left : temporal view and Hann window (normalized to full scale). On the right : corresponding Fourier spectrum.

Généralement, on modélise les courbes de masquage en convoluant le spectre basilaire par une fonction d'étalement. On obtient alors l'*excitation basilaire*. Dans le codeur AAC, on travaille avec l'énergie par partition, on a donc recours à une pseudo-convolution intégrant une normalisation de l'énergie (équation 28). On obtient alors une énergie étalée par partitions : E^p, qu'on assimile à l'excitation. La fonction d'étalement est tracée en figure 12, et notée $\mathrm{sf}(f_B)$. On remarque que la fréquence est exprimée en Barks. L'indice de partition p n'étant pas homogène à des Barks, il faut faire référence à la fréquence médiane en Barks de la partition p, notée $\overline{b}(p)$.

$$E^p(p) = \frac{\sum_{p'} e^p(p')\ \mathrm{sf}\left(\overline{b}(p) - \overline{b}(p')\right)}{\sum_{p'} \mathrm{sf}\left(\overline{b}(p) - \overline{b}(p')\right)} \tag{28}$$

La figure 13 visualise cette énergie étalée. On constate que l'excitation est plus régulière que le spectre. En effet, la pseudo-convolution est analogue à un filtrage passe-bas de l'énergie.

Enfin, il reste à calculer l'indice de masquage. La variation continue entre masquant tonal et non-tonal est réalisée par une fonction de mélange dépendant d'un *indice de tonalité*, noté It^p, compris entre 0 et 1. Dans AAC cet indice de masquage est exprimé par sous-bandes, contrairement à d'autres codeurs, à partir d'un estimateur de la prédictibilité du spectre.

L'indice de masquage s'écrit :

$$\mathrm{Im}^p(p)_{dB} = 18\ \mathrm{It}^p(p) + 6\ (1 - \mathrm{It}^p(p)) \tag{29}$$

On constate donc que l'indice de masquage varie entre 6 et 18 dB.

L'application de l'indice de masquage à l'énergie étalée donne donc le seuil de masquage par partition, noté S^p_m, représenté en figure 13 :

$$S^p_m(p)_{dB} = E^p(p)_{dB} - \mathrm{Im}^p(p)_{dB} \tag{30}$$

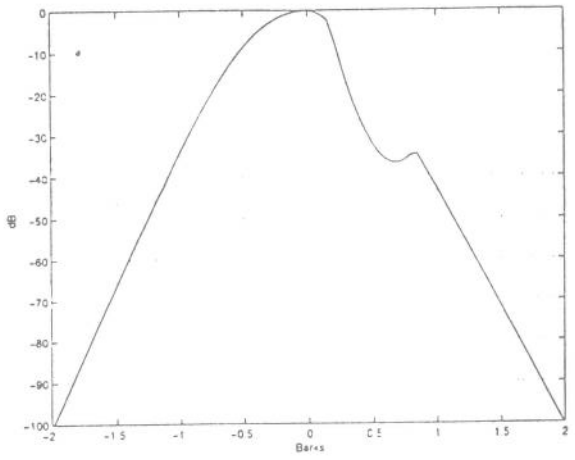

FIG. 12. — Forme de la fonction d'étalement.

Shape of the spreading function.

Ce seuil de masquage n'est toutefois pas homogène à une DSP, car lui aussi pondéré par la largeur des partitions.

À ce niveau du calcul intervient la prise en compte du seuil d'audition absolu, défini pour un niveau d'écoute standard. Une des spécificités de AAC est d'exprimer le seuil d'audition absolu comme une énergie par partitions, ce qui explique le fait qu'il dépende de la fréquence d'échantillonnage. Le codeur conserve alors le maximum entre le seuil de masquage et le seuil d'audition absolu, ce qui signifie que le seuil de masquage définitif est toujours supérieur ou égal au seuil d'audition.

Remarque : Le document normatif suggère d'intégrer ici une opération de *contrôle de pré-écho* dont l'utilité peut sembler assez anecdotique. En fait, cette opération, analogue à un lissage des variations trop brusques du seuil de masquage d'une fenêtre à l'autre, est plus importante qu'il n'y paraît, et peut être interprétée autrement : D'une part, cela permet de prendre en compte (d'une façon assez grossière) le phénomène de masquage temporel (que nous n'avons pas introduit, mais qui peut parfois jouer un rôle non-négligeable). D'autre part, cela évite l'apparition d'artéfacts dus au repliement temporet et fréquentiel à l'add-overtapp (partie synthèse de la MDCT).

Sous cette forme, le seuil de masquage ne peut pas être utilisé par le module de quantification et codage : les sous-bandes de codage (repérées par l'indice s) sont différentes des partitions. Pour exprimer le seuil de masquage sur les partitions, on a retenu la solution qui consiste à conserver le minimum du seuil sur toutes les partitions recouvrant cette sous-bande. D'autre part, le seuil de masquage est pondéré par la largeur des partitions. Pour obtenir un seuil pondéré par la largeur des sous-bandes, il faut donc effectuer l'opération suivante :

$$S_m^s(s) = \min_{p \in B(s)} \left(\frac{S_m^p(p)}{k_{\max}^p(p) - k_{\min}^p(p) + 1} \right) \quad (31)$$

$$(k_{max}^s(s) - k_{min}^s(s) + 1)$$

$k_{min}^s(s)$ et $k_{max}^s(s)$ désignent les indices correspondant aux limites fréquentielles de la sous-bande s. L'exposant s désigne une grandeur définie sur une sous-bande, et B(s) l'ensemble des partitions recouvrant la sous bande s.

Enfin, pour récupérer une grandeur réellement exploitable, on normalise le seuil de masquage par l'énergie en sous-bandes :

$$e^s(s) = \sum_{k=k_{min}^s(s)}^{k_{max}^s(s)} S(k) \quad (32)$$

afin d'en déduire le rapport signal à masque, ou SMR[7], par sous-bande :

$$SMR^s(s) = \frac{e^s(s)}{S_m^s(s)} \quad (33)$$

Le SMR est une grandeur relative indépendante de l'estimateur spectral, ce qui est utile dans le codeur AAC où le modèle psychoacoustique utilise une DFT et le codage une MDCT. Sur la figure 14, nous avons superposé aux coefficients de la MDCT le seuil de masquage déduit du SMR.

IV.4. Choix de la longueur de la MDCT

Le critère proposé pour le choix de la taille de la transformée est l'*Entropie Perceptuelle*, abrégée en PE [20]. Elle est définie comme le nombre minimal théorique de bits de codage par échantion nécessaires au respect de la contrainte psychoacoustique. La théorie du codage de source [21] est tout à fait explicite sur ce point. Elle indique que ce nombre de bits s'exprime en fonction du rapport signal à masque, (cas des fréquences continues) :

$$\mathrm{PE} = \frac{1}{2} \int_{-\frac{1}{2}}^{\frac{1}{2}} \log_2 \left[\frac{S(f)}{S_m(f)} \right] df \quad (34)$$

où $S(f)$ est la DSP du signal, et $S_m(f)$ le seuil de masquage. Ce nombre de bits est un nombre réel. Comme le spectre et le seuil de masquage sont considérés constants sur les partitions, on peut approcher l'expression théorique par :

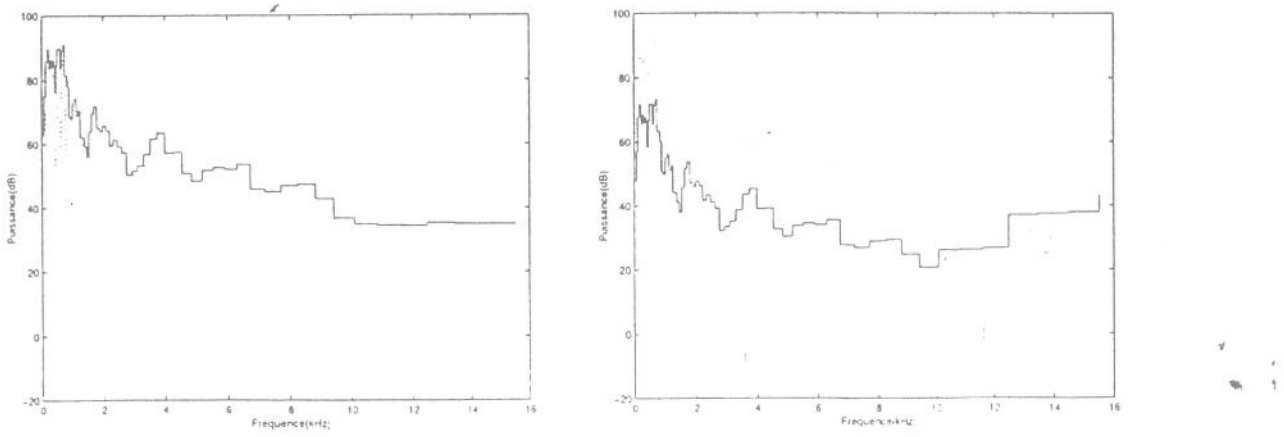

FIG. 13. — Grandeurs intermédiaires du modèle psychoacoustique. À gauche, énergie (pointillés) et énergie étalée (trait plein). À droite, énergie étalée (pointillés) et seuil de masquage (trait plein).

Internal variables of the psychoacoustic model. On the left : energy (dotted) and spread energy (solid). On the right : spread energy (dotted) and masking threshold (solid).

7. Signal to Mask Ratio

$$(35)\quad PE = \frac{1}{2}\sum_{p}(k^{p}_{max}(p) - k^{p}_{min}(p) + 1)\log_2\left[\frac{e^{p}(p)}{S^{p}_{m}(p)}\right]$$

Le document normatif propose de calculer une variante de l'équation 35, pour un bloc de N échantillons. Elle est égale à l'expression théorique à un facteur multiplicatif près, valant 0.602. En comparant PE calculée pour $N=2048$ à un seuil empirique, on décide :

– $N=2048$ si $PE_{2048} <$ seuil
– $N=256$ si $PE_{2048} \geq$ seuil

Ce critère répond plus à un impératif de réduction du nombre de bits de codage qu'à une analyse de la stationnarité du signal. En effet, un signal stationnaire n'implique pas nécessairement une valeur d'entropie perceptuelle faible, bien que ce soit souvent le cas.

IV.5. Influence de la longueur de fenêtre variable

Comme nous l'avons déjà évoqué, le codeur AAC travaille avec deux tailles de MDCT : $N=2048$ ou 256. Les fenêtres de longueur 256 sont, de plus, regroupées par série de 8 afin de conserver un débit bloc constant. Cette particularité a une influence sur la structure du modèle psychoacoustique, d'autant que ces valeurs de N sont valables quelle que soit la fréquence d'échantillonnage. La durée d'une fenêtre d'analyse varie donc avec F_e, de même que la signification fréquentielle des coefficients de DFT. C'est pourquoi la plupart des constantes du modèle psychoacoustique sont définies dans des tables. Pour chaque valeur de F_e, on dispose de tables pour les fenêtres longues, et d'autres tables pour les fenêtres courtes. En particulier, les limites des partitions et des sous-bandes, ainsi que le seuil d'audition absolu sont ainsi variables.

À titre d'exemple, nous donnons dans les tables I et II les caractéristiques de résolution temporelle et fréquentielle ainsi que le nombre de partitions et de sous-bandes (à comparer aux 24 bandes critiques) pour deux fréquences d'échantillonnage classiques.

TABLEAU I. — Résolutions temporelles et fréquentielles (DFT et MDCT) dans le codeur AAC pour $F_e=32$ et 44,1 kHz.

Time and frequency resolution (DFT and MDCT) of an AAC coder, for $F_e=32$ and 44.1kHz.

F_e	32 kHz	44,1 kHz
Durée d'une fenêtre $N=2048$	64 ms	46,4 ms
Durée d'une fenêtre $N=256$	8 ms	5,8 ms
Résolution fréquentielle $N=2048$	15,6 Hz	21,5 Hz
Résolution fréquentielle $N=256$	125 Hz	172,3 Hz

Enfin, on remarque que le choix du paramètre N pour la fenêtre courante n'intervient qu'après le calcul des rapports signal à masque, c'est-à-dire en fin du modèle psychoacoustique. Or le calcul des seuils de masquage a besoin de connaître la valeur de N. Pour résoudre cette contradiction, on doit calculer en parallèle les seuils de masquage pour les deux valeurs de N, et donc fournir au modèle psychoacoustique simultanément une fenêtre de 2048 échantillons, et huit fenêtres de 256 échantillons successives, soit les deux versions d'un même bloc. En revanche, la sortie ne propose qu'un seul jeu de rapports signal à masque car la décision a déjà eu lieu. La MDCT peut ensuite être effectuée. Cependant cette architecture, suggérée dans le document normatif, induit une difficulté de mise en œuvre lors de l'insertion des fenêtres de transitions de la MDCT (lire aussi la section 3.4).

TABLEAU II. — Intervalles fréquentiels dans le codeur AAC pour F_e = 32 et 44,1 kHz.

Number of frequency intervals of an AAC coder, for $F_e=32$ and 44.1 kHz.

F_e	32 kHz	44,1 kHz
Nombre de partitions $N=2048$	66	66
Nombre de partitions $N=256$	44	44
Nombre de sous-bandes $N=2048$	51	49
Nombre de sous-bandes $N=256$	14	14

V. QUANTIFICATION ET CODAGE SANS BRUIT

V.1. Principes et notations

Les étapes décrites dans les sections précédentes peuvent être considérées comme un prétraitement du signal audionumérique. La quantification représente l'étape de compression proprement dite. Son but est d'exprimer les composantes de la MDCT sous forme d'entiers binaires, en commettant une certaine erreur dite de quantification. Plus le nombre de bits utilisés pour représenter une composante sera faible, plus le signal sera comprimé.

De plus, l'opération de quantification doit essayer de satisfaire simultanément deux contraintes qui s'opposent l'une à l'autre :

– Respecter les exigences du modèle psychoacoustique, soit contenir la DSP de l'erreur de quantification en dessous du seuil de masquage. On la nomme *contrainte psychoacoustique*.
– Le nombre de bits de codage utilisés par unité de temps doit rester inférieur au débit de sortie net du codeur. En effet, le flux binaire doit conserver un débit fixe, qui est défini comme un paramètre de la transmission. Il s'agit du débit brut. Ensuite, le flux binaire inclut des informations de contrôle et des données représentant les composantes quantifiées. La part du débit brut disponible pour les données

est appelée débit net, c'est le débit utile pour la quantification. Il s'agit donc de la *contrainte de débit*.

Or, en général, plus la DSP de l'erreur de quantification est faible, plus le nombre de bits de codage nécessaire est important. On a donc affaire à une procédure d'optimisation sous double contrainte, particulièrement complexe. D'autant qu'il arrive souvent que les deux contraintes ne puissent pas être satisfaites simultanément, en particulier lorsque le débit brut spécifié est faible. La qualité sonore de l'ensemble codeur-décodeur n'est alors plus optimale. Toutefois, nous n'aborderons pas dans cet article la question délicate de la gestion de la sous-optimalité

La stratégie d'optimisation pour satisfaire ces deux contraintes n'est pas normalisée par MPEG. La norme spécifie uniquement la syntaxe du flux binaire, c'est-à-dire le mode de représentation des composantes quantifiées. Une stratégie possible est proposée dans le document normatif. Elle sera décrite par la suite. Il s'agit d'un algorithme itératif à deux boucles imbriquées.

De plus, la représentation binaire des composantes quantifiées ne fait pas appel à un code à longueur fixe, mais à un jeu de codes à longueur variable, qui font partie de la famille des codes de Huffman. Il s'agit d'un codage sans bruit, en bloc, utilisant des tables prédéfinies. L'algorithme de quantification est généralement itératif. À chaque étape, il est nécessaire de mesurer le nombre de bits de codage utilisés, pour évaluer la contrainte de débit. Il faut donc inclure l'opération de codage sans bruit à l'intérieur des boucles. D'autre part, pour estimer la DSP de l'erreur de quantification à chaque étape, et évaluer la contrainte de masquage, la méthode choisie consiste à décoder le signal puis à calculer effectivement l'erreur par soustraction du signal reconstruit au signal d'origine. On dit qu'il s'agit d'un codage de type *analyse par la synthèse*. Dans les codeurs plus anciens, par exemple le codeur MPEG-1, couche I et II, l'allocation de bits est faite en *boucle ouverte* : On alloue des bits en fonction du rapport signal à masque fourni par le modèle psychoacoustique, suivant des tables de correspondance calculées a priori. On ne contrôle donc pas directement le bruit de quantification.

Comme une quantification indépendante de chaque composante serait trop coûteuse en terme de débit, on regroupe les composantes en sous-bandes fréquentielles. Il s'agit de groupes de composantes adjacentes, de largeur variable : Les sous-bandes les plus étroites (4 composantes) se trouvent dans les basses fréquences, puis leur largeur augmente en allant vers les aigus, afin de suivre la sensibilité fréquentielle de l'oreille. Toutefois, à cause des dimensions des tables de Huffman, elles regroupent toujours un multiple de 4 composantes. Leur définition dépend de la fréquence d'échantillonnage et de la taille de fenêtre. À 32 et 44,1 kHz, en fenêtre longue, la sous-bande la plus large comporte 32 composantes. (voir aussi la section IV.5). La figure 14 représente les modules des 1024 composantes MDCT pour une fenêtre longue, ainsi que le seuil de masquage et les limites des sous-bandes. Le signal est le même que précédemment, toujours échantillonné à 32 kHz.

On rappelle les notations adoptées dans les sections précédentes : $y_k(m)$ est la k-ème composante en sortie du banc de filtres d'analyse correspondant à la m-ème fenêtre. Comme le codeur AAC n'exploite pas la corrélation entre $y_k(m-1)$ et $y_k(m)$ dans le module de quantification[8] et comme il faudra spécifier les composantes relativement à des sous-bandes, on modifiera dans cette section les notations sous la forme $y(sb, r)$ où sb spécifie une sous-bande et r une composante dans cette sous-bande.

Contrairement aux codeurs MPEG1 couche I et II, où l'allocation de bits utilise un jeu de quantificateurs prédéfinis, le codeur AAC, comme le MPEG1 couche III, utilise le principe de mise à l'échelle des composantes spectrales : Tout quantificateur scalaire uniforme est équivalent à un quantificateur scalaire uniforme unitaire (de pas 1) précédé d'un gain qui réalise la mise à l'échelle des composantes. Le quantificateur scalaire uniforme unitaire peut être simplement implémenté par l'opération d'arrondi à l'entier le plus proche.

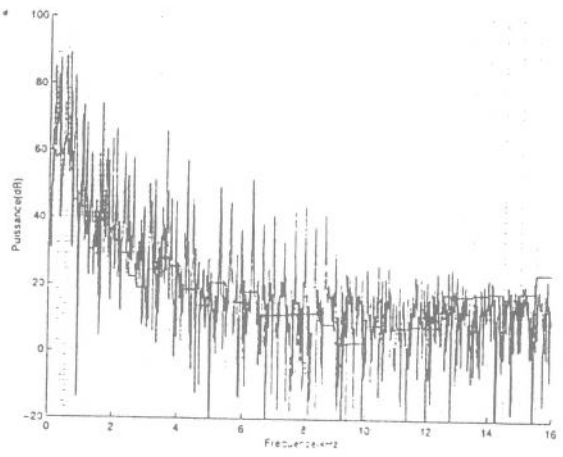

FIG. 14 — Coefficients de la MDCT, seuil de masquage et limites des sous-bandes, en échelle fréquentielle linéaire.

MDCT coefficients, masking threshold and subbands boundaries, with a linear frequency scale.

Chaque composante $y(sb, r)$ est donc quantifiée de façon scalaire suivant la formule :

$$(36)\quad y_q(sb, r) = \text{signe}\left[y(sb, r)\right] \text{round}\left(\left|y(sb, r)\right| 2^{-\frac{1}{4}\mu(sb)}\right)^{\gamma}$$

où *round* est la fonction réalisant l'arrondi, γ est une constante prenant la valeur 3/4. Les paramètres entiers

8. Cette corrélation est exploitée dans le module optionnel de prédiction décrit dans la section 6.4

$$g(x) = \left(\frac{x}{2^{\frac{1}{4}\mu(sb)}}\right)^{3/4} = \left(\frac{x}{2^{\frac{3}{4}\mu(sb)}}\right)^{3/4}$$

15/20

$\mu(sb)$ sont appelés les *facteurs d'échelle*. On constate que le choix de ces paramètres, définis par sous-bande, suffit à décrire entièrement l'opération de quantification. Le but de l'algorithme d'optimisation est donc de déterminer les valeurs optimales de ces paramètres. Le codage sans bruit à l'aide de dictionnaires de Huffman est alors effectué sur ces valeurs quantifiées (et éventuellement sur les signes).

On reconstruit les entrées du banc de filtres de synthèse au décodage par une formule inverse de la précédente :

$$(37) \qquad \hat{y}(sb, r) = \text{signe}\,[y(sb, r)]\ |y_q(sb, r)|^{\frac{1}{\gamma}}\ 2^{\frac{1}{4}\mu(sb)}$$

V.2. Codage sans bruit

Avant d'aborder le problème central qui est l'optimisation des facteurs d'échelle, examinons le module de codage sans bruit. Les entrées de ce module sont les M composantes quantifiées $y_q(sb, r)$ ainsi que les facteurs d'échelle $\mu(sb)$. Le codage de Huffman a pour rôle d'une part de calculer le nombre de bits nécessaire pendant le déroulement du processus itératif d'optimisation, et d'autre part de déterminer les mots de code en sortie du processus. Dans cet article, nous n'aborderons pas les questions relatives à la construction du train binaire.

V.2.1. Codage des facteurs d'échelle

Les facteurs d'échelle $\mu(sb)$ sont également codés par un code de Huffman. Pour cela, on définit un *gain global*, g :

$$(38) \qquad g = \mu(1)$$

Le gain global est codé en binaire naturel sur 8 bits. Ensuite, les facteurs d'échelles sont définis de manière différentielle au moyen des paramètres Δ :

$$(39) \qquad \forall\, sb \neq 1 \quad \Delta(sb) = \mu(sb) - \mu(sb-1)$$

On forme alors les indices correspondants :

$$(40) \qquad index\,(sb) = \Delta(sb) + 60$$

Ces indices sont directement utilisés comme entrée des tables de Huffman. On remarque toutefois que les dictionnaires codent les indices entiers entre 0 et 120. La valeur différentielle entre deux facteurs d'échelle de deux sous-bandes adjacentes est donc limitée entre -60 et 60.

V.2.2. Codage des composantes quantifiées

On montre, figure 15 à gauche, la valeur des 1024 composantes $|y(sb, r)|$ de la MDCT suivant une échelle linéaire. Cette information est équivalente à celle de la figure 14 exprimée en dB. On montre, figure 15 à droite, les valeurs quantifiées $|y_q(sb, r)|$ correspondant à un certain choix de $\mu(sb)$ et après convergence du processus itératif. On remarque que la dynamique des valeurs $|y(sb, r)|$ est forte, comprise entre 0 et 15000 environ. En revanche, la dynamique des valeurs quantifiées $|y_q(sb, r)|$ est nettement plus réduite (entre 0 et 13). Ces nouvelles valeurs sont mieux adaptées à une représentation sous forme d'entiers binaires.

Les tracés de la figure 15 montrent aussi que cette opération de contraction n'est pas linéaire. C'est le rôle de la constante γ dont la valeur a été prise égale à 3/4. En effet, en observant les données de la table III on constate que les fortes valeurs sont nettement plus réduites que les valeurs faibles ce qui est assez logique compte tenu du comportement logarithmique de l'oreille en puissance.

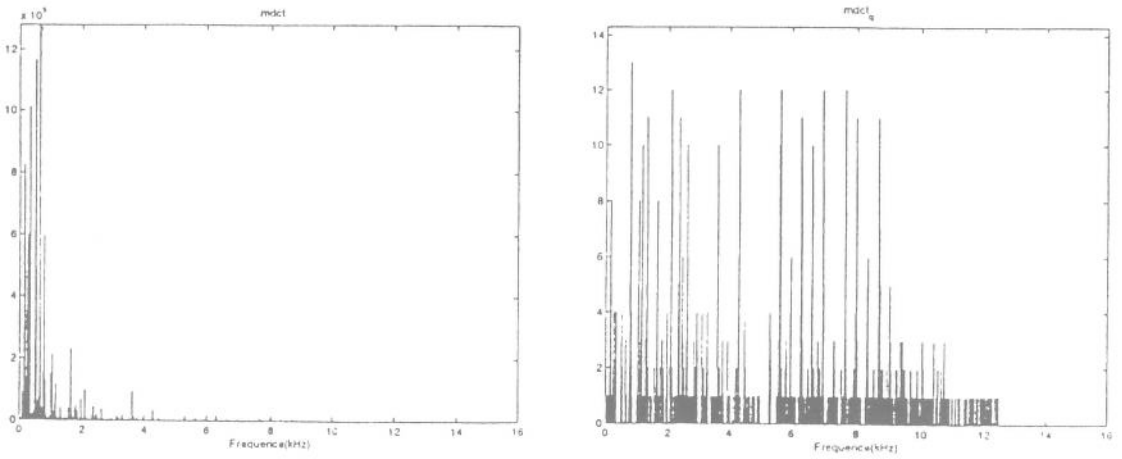

FIG. 15 — M = 1024 composantes $|y(sb, r)|$ de la MDCT à gauche. Résultat de la quantification $|y_q(sb, r)|$ à droite.

On the left : M = 1024 MDCT coefficients $|y(sb, r)|$. On the right : the same coefficients, quantized : $|y_q(sb, r)|$.

TABLEAU III. — Influence du paramètre $\gamma = 3/4$.

Influence of γ parameter value: $\gamma = 3/4$.

x	1	10	100	1000	10000
$y = x^{3/4}$	1	6	32	178	1000
x/y	1	1.7	3.1	5.6	10

Le codage des composantes $y_q(sb, r)$ se fait par sous-bande. On détermine la valeur absolue maximale des composantes dans chaque sous-bande et on lui attribue un dictionnaire d'Huffman selon cette valeur maximale. Cette correspondance est définie dans la table IV. Il existe 12 dictionnaires correspondant chacun à la valeur maximale pouvant être codée. Le choix du dictionnaire par ce critère de valeur maximale est transmis comme une information dans la chaîne binaire. Deux dictionnaires sont particuliers. Le dictionnaire caractérisé par la valeur maximale 0 signale uniquement que toutes les composantes quantifiées dans la sous-bande considérée sont nulles et que, par conséquent, elles ne seront ni codées ni transmises. Le dernier dictionnaire est destiné au codage des composantes d'une sous-bande dont la valeur maximale est comprise entre 12 et 16. Pour les valeurs supérieures à 16 un code d'échappement est utilisé en combinaison avec ce dictionnaire. Il existe deux dictionnaires pour chacune des valeurs maximales 1, 2, 4, 7 et 12. Chacun d'eux représente une fonction de distribution de probabilité différente. Le choix entre les deux est toujours le dictionnaire optimal en terme de nombre de bits alloués. Les composantes de chaque sous-bande sont regroupées et codées par blocs de 2 ou 4, selon le dictionnaire choisi (table IV colonne 3), autrement dit, on attribue un seul mot de code par 2 ou par 4 composantes consécutives. Les dictionnaires no 1, 2, 5 et 6 gèrent directement le signe des composantes dans le mot de code attribué (table IV colonne 2). Les autres dictionnaires traitent le signe séparément en le concaténant au mot de code, ce qui nécessite un bit supplémentaire par composante (moins si une ou plusieurs composantes sont nulles).

TABLEAU IV. — Caractéristiques des dictionnaires de Huffman.

Huffman codebooks specifications.

No dictionnaire	Composantes signées	Nombre composantes	Valeur max
0			0
1	Oui	4	1
2	Oui	4	1
3	Non	4	2
4	Non	4	2
5	Oui	2	4
6	Oui	2	4
7	Non	2	7
8	Non	2	7
9	Non	2	12
10	Non	2	12
11	Non	2	16

Prenons l'exemple de 2 composantes $|y_q(sb, r)|$ et $|y_q(sb, r+1)|$ appartenant à une sous-bande où la valeur maximale de l'ensemble des composantes dans cette sous-bande serait égale à 7. On a le choix entre les dictionnaires no 7 et 8. On donne, à titre d'exemple, ces deux dictionnaires, table V représentés sous forme matricielle. Ils donnent en fonction des valeurs prises par $|y_q(sb, r)|$ et $|y_q(sb, r+1)|$ le nombre de bits nécessaire pour coder ces deux informations. On rappelle qu'il doit aussi exister deux autres matrices fournissant les mots de code associés à chaque couple $|y_q(sb, r)|$ et $|y_q(sb, r+1)|$. Elles ne sont pas reproduites ici. On remarque que peu de bits sont nécessaires pour de faibles valeurs des composantes, ce qui paraît assez logique, que manifestement les fonctions de distribution de probabilités sont différentes et que ces matrices ne sont pas tout à fait symétriques, ce qui semble plus bizarre ! Le dictionnaire no 7 correspond manifestement au cas où les 4 couples {00,01,10,11} sont très probables.

TABLEAU V. — Exemple des dictionnaires de Huffman n° 7 et 8. La première colonne spécifie les valeurs possibles de $|y_q(sb, r)|$ et la première ligne les valeurs possibles de $|y_q(sb, r+1)|$. Les valeurs dans la matrice spécifient le nombre de bits nécessaires pour coder ces deux informations.

Example: Huffman codebooks #7 and 8. First column sets $|y_q(sb, r)|$ values, and first line $|y_q(sb, r+1)|$ values. Matrix values specify the number of bits required for coding both coefficients.

	0	1	2	3	4	5	6	7
0	1	3	6	7	8	9	10	11
1	3	4	6	7	8	8	9	9
2	6	6	7	8	8	9	9	10
3	7	7	8	8	9	9	10	10
4	8	8	9	9	10	10	10	11
5	9	8	9	9	10	10	11	11
6	10	9	9	10	10	11	12	12
7	11	10	10	10	11	11	12	12

	0	1	2	3	4	5	6	7
0	5	4	5	6	7	8	9	10
1	4	3	4	5	6	7	7	8
2	5	4	4	5	6	7	7	8
3	6	5	5	6	6	7	8	8
4	7	6	6	6	7	7	8	9
5	8	7	6	7	7	8	8	10
6	9	7	7	8	8	8	9	9
7	10	8	8	8	9	9	9	10

V.3. Algorithme d'optimisation

Une stratégie pour le choix des facteurs d'échelle consiste à les séparer en une partie fixe (β) et une partie variable ($\alpha(sb)$, toujours positive) :

$$\mu(sb) = \beta - \alpha(sb) \quad (41$$

La première est modifiée dans une première boucle, afin de respecter la contrainte de débit, alors que la seconde est modifiée dans une seconde boucle qui contient la première, afin de respecter la contrainte psychoacoustique par mise en forme spectrale de l'erreur de quantification.

On remarquera que les facteurs α et β sont directement liés à la puissance de l'erreur de quantification et au nombre de bits consommés, mais suivant des lois complexes et localement non-monotones. En effet, dans

chaque sous-bande *sb*, la puissance de l'erreur de quantification est donnée par :

$$(42) \qquad \sigma_Q^2(sb) = \frac{1}{\text{card}(sb)} \sum_{r \in sb} [|\hat{y}(sb, r)| - |\hat{y}(sb, r)|]^2$$

et la contrainte psychoacoustique pour cette sous-bande s'écrit :

$$(43) \qquad \sigma_Q^2(sb) \leq = \sigma_{mask}^2(sb)$$

Or $|\hat{y}(sb, r)|$ dépend d'une opération d'arrondi. D'autre part, le nombre de bits consommé dépend de la sortie des tables de Huffman, qui reste assez imprévisible. Toutefois, en moyenne sur un grand nombre de réalisations, on peut dire que lorsque $\mu(sb)$ augmente, $\sigma_Q^2(sb)$ augmente, et le nombre de bits consommé par la sous-bande diminue, sans qu'il soit possible de prévoir ponctuellement l'effet d'une variation de $\mu(sb)$.

Décrivons maintenant cet algorithme d'optimisation plus en détail. La boucle interne récupère la valeur des paramètres $\alpha(sb)$ et β déterminés aux itérations précédentes, réalise une quantification non-uniforme puis un codage de Huffman et en déduit le nombre de bits nécessaires. Elle modifie ensuite le paramètre β en l'augmentant si le nombre de bits nécessaires est supérieur au nombre de bits disponibles ou en le diminuant dans le cas contraire d'une quantité entière $\Delta\beta$. Cette quantité $\Delta\beta$ est initialisée à une valeur importante (64) à l'entrée de la boucle. Elle est ensuite systématiquement divisée par 2 à chaque itération. La boucle interne est terminée lorsque $\Delta\beta = 0$.

La boucle externe est responsable de la mise en forme du bruit de quantification. Le codeur calcule la puissance de l'erreur de quantification dans chaque sous-bande, puis évalue la contrainte de masquage. Si elle n'est pas satisfaite pour une ou plusieurs sous-bandes, on augmente les paramètres $\alpha(sb)$ correspondant. Ceci risque évidemment de rendre le nombre total de bits utilisés pour le codage supérieur au nombre de bits disponibles. D'où l'appel de la boucle interne de contrôle du débit pour réduire éventuellement le nombre de bits utilisés. Le codeur cherche ainsi à trouver un compromis entre débit et masquage psychoacoustique.

Contrairement au paramètre β qui peut augmenter ou diminuer, les paramètres $\alpha(sb)$ ne peuvent qu'augmenter. L'augmentation prévue dans la norme est de 1,5 dB. Cette valeur est assez raisonnable puisque l'oreille détecte des variations de puissance de l'ordre de 1 dB[9].

Ces étapes sont répétées itérativement jusqu'à ce qu'une ou plusieurs conditions de sortie de la boucle externe soient vérifiées. La condition principale est évidemment le masquage du bruit dans toutes les sous-bandes, mais elle n'est pas toujours possible surtout lorsque le débit visé est faible. Une autre condition de sortie est relative au codage différentiel des facteurs d'échelle et impose que la différence entre les facteurs d'échelles de deux sous-bandes adjacentes ne soit pas supérieure à 60. Comme l'algorithme d'analyse par synthèse n'est pas nécessairement convergent, on rajoute d'autres conditions, comme celle, par exemple, qui consiste à fixer une valeur limite pour les facteurs d'échelle.

On montre, figure 16 (à gauche), le spectre de l'erreur de quantification comparé au seuil de masquage à la fin de l'algorithme d'optimisation. On constate que, dans cette fenêtre d'analyse, la contrainte psychoacoustique n'est pas rigoureusement vérifiée, mais le spectre de l'erreur est très proche du seuil de masquage. (dans ces simulations le débit a été choisi égal à 64 kbit/s). On montre également (à droite) la valeur optimale des facteurs d'échelle.

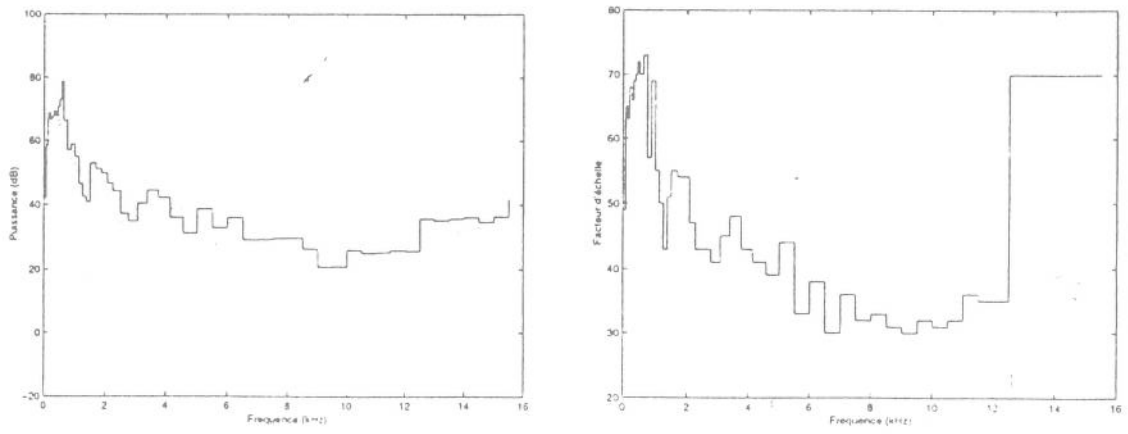

FIG. 16. — A gauche, seuil de masquage (en trait plein) et spectre de l'erreur de quantification (en pointillés) à la fin de l'algorithme d'optimisation. À droite, valeur optimale des facteurs d'échelle, $\mu(sb)$ en trait plein, et valeur du facteur β en pointillés. Les facteurs $\alpha(sb)$ correspondent donc à l'écart entre ces deux courbes.

On the left: masking threshold (solid) and quantization error spectrum (dotted). On the right : optimal value of scalefactors $\mu(sb)$ *(solid) and value of* β *parameter (dotted).* $\alpha(sb)$ *parameters thus corresponds to the distance between these curves.*

9. Comme $20 \log_{10} 2^{1/4} = 1,5$, ceci explique la division par 4 qui apparaît dans la formule 36. (Il en est de même pour β).

VI. MODULES OPTIONNELS

Pour terminer, nous allons évoquer rapidement les quatre modules optionnels d'un codeur ou d'un décodeur AAC. Ils permettent d'améliorer l'efficacité du codage, mais augmentent la charge de calcul, tant au codage qu'au décodage. Ces modules prennent place entre le banc de filtres et la quantification. Ils sont présentés dans l'ordre de leur utilisation dans le codeur. Dans le décodeur, ils suivent naturellement l'ordre inverse.

VI.1. Mise en forme temporelle du bruit

Une fois obtenus les coefficients MDCT pour la fenêtre courante, ce module permet d'appliquer un pré-filtrage linéaire au signal, de manière indépendante pour chaque fenêtre. Il n'intervient que lorsque le gain de prédiction, obtenu par une analyse LPC standard, dépasse une valeur seuil, donc lorsque le signal est de nature prédictible. Les coefficients du filtre sont simplement ceux obtenus avec l'analyse LPC. On essaye donc de blanchir le signal avant codage. Ces coefficients sont transmis au décodeur dans le flux de bits, afin d'appliquer le filtrage inverse. Le filtre de codage est de type non-récursif. Le filtre de décodage est donc de type auto-régressif (ou tout-pôles).

VI.2. Stéréo d'intensité

Ce module est utilisé en configuration stéréophonique. On dispose donc d'un canal L (Left) et d'un canal R (Right). En s'appuyant sur une déficience naturelle de la perception auditive en hautes fréquences, on considère que les coefficients MDCT, dans sous-bandes hautes des canaux L et R, sont identiques, à des coefficients multiplicatifs près, appelés *coefficients de couplage*. Au décodage, les coefficients MDCT de L et R sont donc déduits d'un seul canal, prenant la place du canal L, en y adjoignant les coefficients de couplage qui prennent alors la place des facteurs d'échelles du canal R. On a donc économisé les bits qui auraient servi au codage des coefficients MDCT quantifiés du canal R.

VI.3. MS Stéréo

Il s'agit d'une seconde option de codage des paires de canaux stéréo, toujours notés L et R. Cette fois, on ne fait aucune approximation sur l'interdépendance des signaux L et R, mais on transforme ces canaux en M et S, pour Middle et Sides, avec une opération de matriçage simple : $M = \frac{1}{2}(L + R)$, et $S = \frac{1}{2}(L - R)$. En effet, dans la majorité des signaux stéréophoniques, L, et R sont semblables. Dans ce cas, la majeure partie de la puissance se retrouve sur le canal M, au détriment du canal S. Ce matriçage s'avère effectivement efficace, mais implique de modifier la procédure de calcul des seuils de masquage, car les signaux n'ont plus la même signification. Cela suppose une variante de modèle psychoacoustique que nous n'aborderons pas dans cet article. Il est à noter que l'option MS Stéréo s'applique à toutes les sous-bandes, et exclut naturellement l'emploi de la stéréo d'intensité.

VI.4. Prédiction

Cet outil permet de soustraire au signal avant codage une composante prédictible. Pour ce faire, on utilise un prédicteur linéaire au second ordre dans le domaine transformé : On note $y(m)$ les coefficients MDCT à la fenêtre m, et $y_q(m)$ les coefficients quantifiés. Un prédicteur P est associé à l'équation de prédiction suivante :

$$y^{\text{pred}}(m) = P(y_q(m-1), y_q(m-2)) \tag{44}$$

Les coefficients de P sont obtenus par une estimation aux sens des moindres carrés. Alors, on ne code que le signal correspondant à l'erreur de prédiction :

$$e(m) = y(m) - y^{\text{pred}}(m). \tag{45}$$

La prédiction ne s'applique toutefois qu'en fenêtres longues, car le signal est sensé être stationnaire, et uniquement pour les sous-bandes correspondant aux basses fréquences.

Au décodage, on suppose déjà obtenus les blocs de coefficients reconstruits $\hat{y}_q(m-1)$ et $\hat{y}_q(m-2)$. Des informations contenues dans le flux de bits permettent de décrire le prédicteur utilisé au codage, P. Alors, on calcule le spectre reconstruit à la fenêtre m selon :

$$\hat{y}_q(m) = \hat{e}_q(m) + P(\hat{y}_q(m-1), \hat{y}_q(m-2)) \tag{46}$$

On remarquera qu'il s'agit d'un second module de prédiction, qui s'ajoute à la mise en forme temporelle du bruit. Toutefois, ces deux outils sont complémentaires car le premier travaille à l'intérieur d'un bloc de coefficients MDCT, alors que celui-ci exploite la corrélation entre blocs successifs.

VII. CONCLUSION

Par sa structure modulaire et ses choix techniques, le codeur MPEG-2 AAC est actuellement le plus abouti et le

plus évolutif des codeurs audio à haute qualité. En outre, il a été prévu pour un très grand choix de débits et de fréquences d'échantillonnage, ce qui en fait pour ainsi dire un codeur audio universel pour les signaux de musique, ou pour la voix à très haute qualité. De plus, des tests de qualité récents [22] ont comparé un codeur-décodeur (codec) AAC optimisé, développé par l'institut Fraunhofer, à d'autres codecs usuels, à savoir des MPEG-1 layer II et III, ainsi qu'un un Dolby AC-3. Les tests d'écoute aveugles, en configuration stéréophonique et avec une fréquence d'échantillonnage de 32 kHz, pour des débits binaires bruts (donc pour les deux canaux) compris entre 64 et 192 kbits/s, ont clairement établi la supériorité de ce codec AAC sur ses concurrents. De plus, on a observé que la meilleure qualité sonore absolue (c'est-à-dire sans considération de débit) était obtenue avec le codec AAC à 128 kbits/s, devançant nettement le Dolby AC-3 à 192 kbits/s. Il faut toutefois préciser que n'étaient pas pris en compte les facteurs de complexité algorithmique et de délai de reconstruction.

Manuscrit reçu le 28 février 2000
accepté le 21 juin 2000

BIBLIOGRAPHIE

[1] 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.

[2] NOLL (P.), "MPEG digital audio coding", *IEEE Signal Processing Magazine*, pp. 59-81, September 1997.

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

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

[5] HERRE (J.) and SCHULTZ (D.), "Extending the MPEG-4 AAC codec by perceptual noise substitution", *Audio Engineering Society Convention Preprints*, May 1998, 104th Convention, Preprint n° 4720.

[6] JAYANT (N.) and NOLL (P.), Digital coding waveforms, *Prentice Hall*, 1984.

[7] JAYANT (N.), JOHNSTON (J.), and SAFRANEK (R.), "Signal compression based on models of human perception", *Proceedings of the IEEE*, vol. 81, n° 10, pp. 1385-1422, October 1993.

[8] JOHNSTON (J.), "Transform coding of audio signals using perceptual noise criteria", *IEEE Journal on Selected Areas in Communications*, vol. 6, n° 2, pp. 314-323, February 1988.

[9] PERREAU-GUIMARES (M.), *Optimisation des ressources binaires et modélisation psychoacoustique pour le codage audio*, PhD thesis, Université de Paris V, Juin 1998.

[10] VAIDYANATHAN (P.), "Multirate digital filters, filter banks, polyphase networks and applications : A tutorial", *Proceedings of the IEEE*, Jannary 1990.

[11] PRINCEN (J.) BRADLEY (A.), "Analysis/synthesis filter bank design based on time domain aliasing cancellation", *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. 34, n° 5, pp. 1153-1161, October 1986.

[12] MALVAR (H.), *Signal processing with lapped transforms*, Artech House, 1992.

[13] RAO (K.) and YIP (P.), Discrete cosine transform: algorithms, advantages, applications, *Academic Press*, 1990.

[14] CROCHIERE (R.) and RABINER (L.), Multirate Digital Signal Processing, *Prentice-Hall*, 1983.

[15] ZWICKER (E.), FELDTKELLER (E.), *Psychoacoustique, l'oreille récepteur d'information*, Masson, Collection technique et scientifique des télécommunications. Traduit de l'allemand par C. Sorin, 1981.

[16] MOORE (B.), An introduction to the psychology of hearing, *Academic Press*, Second edition, 1982.

[17] GREEN (D.), An Introduction to Hearing, *Hillsdale*, New-Jersey, USA: LEA, 1976.

[18] BOTTE (M.), CANÉVET (G.), DEMANY (L.), SORIN (C.), *Psychoacoustique et perception auditive*, Série Audition INSERM/SFA/CNET, 1989.

[19] HUMES (L.), and JESTEADT (W.), "Models of the additivity of masking," *J. Acoust. Soc. Am.*, vol. 85, n° 3, pp. 1285-1294, March 1989.

[20] VELDHUIS (R.), "Bit rates in audio source coding", *IEEE Journal on Selected Areas in Communications*, vol. 10 n° 1 pp. 86-96, 1992.

[21] BERGER (T.), Rate distortion theory: A mathematical basis for data compression, *Prentice-Hall*, 1971.

[22] SOULODRE (G.), GRUSEC (T.), LAVOIE (M.), THIBAULT (L.), "Subjective evaluation of state of-the-art 2-channel audio codecs", *Audio Engineering Society Convention Preprints*, May 1998, 104th Convention, Preprint n° 4740.