Structure de codage audio spatialisé à scalabilité hybride

Adil MOUHSSINE[1] Abdellatif BENJELLOUN TOUIMI[1]

[1] France Télécom Recherche et Développement – TECH/SSTP
2, Avenue Pierre Marzin
22307 Lannion Cedex - France

{adil.mouhssine, abdellatif.benjellountouimi}@orange-ftgroup.com

Résumé.

Cet article présente une nouvelle structure de codage audio 3D permettant d'assurer une scalabilité hybride (scalabilité de la qualité intrinsèque perceptive et scalabilité de la qualité spatiale et du système de rendu sonore utilisé). Ce nouveau codeur se base sur l'architecture du codeur monophonique MPEG2/4-AAC en lui ajoutant une transformée spatiale permettant de mieux répartir l'information spatiale contenue dans la scène 3D à coder et un module de test de pertinence des éléments résultant de cette transformée. Un module d'ordonnancement du flux binaire est enfin appliqué à la sortie du codeur.

Mots clefs

Codage audio 3D, Ambisonic, MPEG2/4-AAC, scalabilité hybride, angles de Gerzon.

1 Introduction

Le son 3D ou le son spatial permet d'offrir une nouvelle expérience d'écoute à l'utilisateur en lui ajoutant le confort basé sur la localisation des sources contenus dans la scène sonore. Une scène sonore 3D peut être transmise et reproduite sur différents systèmes de rendu. De tels systèmes dépendent essentiellement du terminal de restitution utilisé : simples écouteurs, deux haut-parleurs d'un PC, un système home cinéma à 5 haut-parleurs ou encore des systèmes plus sophistiqués comme l'Ambisonic.

Les codeurs audio 3D visent à réduire le débit de codage nécessaire d'une scène sonore riche en information spatiale tout en garantissant une qualité auditive globale optimale. Dans ce cas la qualité auditive globale désigne:

- la qualité auditive intrinsèque relative à la qualité du signal audio reconstruit par rapport au signal original.
- la qualité auditive spatiale relative à la précision des sources sonores reconstruite selon un système de reproduction.

La plupart des codeurs audio monophonique existants peuvent être adapté pour le codage des signaux multicanaux. Cependant, MPEG Surround [6] représente l'état de l'art actuel dans le domaine. Il est basé sur une approche paramétrique pour le codage des signaux multicanaux. En effet, le processus de codage extrait les paramètres spatiaux à partir des signaux multicanaux qui seront codés et transmis en parallèle des données du signal. Une deuxième opération consiste à effectuer un "down-mix" sur l'ensemble des signaux puis le codage du signal résultant à l'aide d'un codeur monophonique (MPEG-4 AAC, HE-AAC …). Le processus de décodage consiste principalement en la synthèse de la scène sonore 3D à l'aide des paramètres spatiaux transmis et du signal mono ou stéréo décodé [6].

Le développement important et la multiplication des moyens d'accès au contenu multimédia riches en information spatiale ont posé le problème de l'hétérogénéité des réseaux et terminaux. Pour faire face à ce problème, des méthodes spécifiques de compression/représentation du contenu ont été mises au point. La scalabilité est une fonctionnalité permettant au codeur de devenir flexible et de produire des formats de flux binaires permettant de s'adapter facilement et de manière naturelle à ces hétérogénéités. Dans cet article nous présentons une technique permettant d'effectuer un codage audio 3D d'une scène sonore en garantissant une scalabilité du flux binaire en sortie pour une adaptabilité à différents réseaux de transmission et systèmes de rendu sonore. Après un bref rappel sur la représentation Ambisonic dans la section 2, la section 3 décrit la structure de ce codeur et le principe de base de son fonctionnement. La section 4 donne les détails de l'algorithme permettant d'ordonner le flux binaire pour obtenir cette scalabilité hybride.

2 Ambisonic

L'ambisonic est un format de représentation du champ acoustique. Une scène sonore peut être représentée par un ensemble de composantes ambisonics. Ces composantes permettent de stocker l'information spatiale relative au champ acoustique [1].

Considérant une onde plane S_θ, d'incidence θ et d'amplitude P_θ, qui se propage dans l'espace (Figure 1), la

décomposition de cette onde en harmoniques cylindriques en un point $M(r,\varphi)$ s'écrit sous la forme suivante:

$$S_\theta(r,\varphi) = P_\theta.\left[J_0(kr) + \sum_{1\le m\le\infty} 2.j^m J_m(kr).(\cos m.\theta.\cos m.\varphi + \sin m.\theta.\sin m.\varphi) \right]$$

où J_m est la fonction de Bessel d'ordre m, k est le nombre d'onde associé à S_θ, et r et φ sont les coordonnées polaires du point courant.

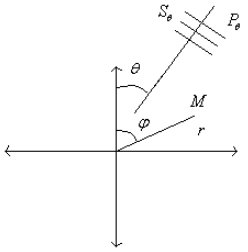

Figure 1: *Propagation d'une onde plane*

A partir de cette décomposition on peut considérer le signal S_θ comme la superposition d'une infinité de signaux élémentaires $(B_m)_{m\in\mathbb{N}}$ appelés composantes ambisonics [1]. Si on limite la décomposition à un ordre M, on obtient une approximation du champ acoustique entièrement décrite par les composantes $(B_0, B_1, \cdots, B_M)$. Ces composantes définissent une représentation ambisonic 2D, d'ordre M [1] [3] [4]. La restitution du champ originale se fera à l'aide d'une matrice de décodage dépendant seulement de la résolution spatiale et de la position du dispositif de reproduction (hauts parleurs) [4] [5]. Cette approximation peut être vue comme une opération d'encodage spatial que subit la scène sonore. En effet, de cette restriction du nombre de composantes ambisonic, dépend la résolution spatiale de la projection. Ainsi, pour avoir une représentation fine de la scène sonore, et donc une qualité spatiale meilleure, il est nécessaire d'effectuer un encodage ambisonic à un ordre M élevé. Lors du décodage, l'ordre de la matrice est aussi important pour garder la meilleure qualité de restitution possible. De ces deux paramètres dépend la précision de restitution [4] [5].

3 Structure de codage à scalabilité hybride

3.1 Architecture du codeur audio 3D à scalabilité hybride

Le codeur audio utilisé dans cet article est donné dans la Figure 2. Le schéma de ce codeur est constitué de manière générique des blocs suivants :

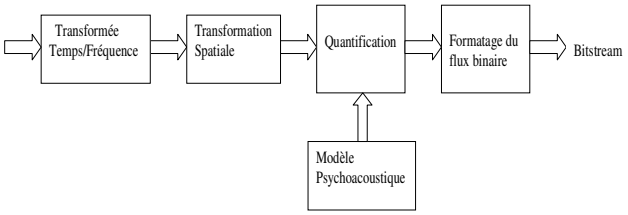

Figure 2 – *Schéma du codeur générique*

- Transformation temps/fréquence de type transformée MDCT caractérisée par sa longueur de fenêtre M.
- Encodage spatial : Il est effectué sur les signaux d'entrée et est définie par une transformée spatiale. On se focalisera sur le cas particulier de la transformée spatiale Ambisonic [1] dont la matrice est caractérisée par les paramètres $(N, p, Q, \{\theta_i\}_{1\le i\le N}, \{\varphi_i\}_{1\le i\le N})$ où :
  - N est le nombre de signaux d'entrée.
  - p l'ordre de l'encodage Ambisonic.
  - $Q = 2p+1$ le nombre de composantes résultant de l'encodage.
  - $\{\theta_i\}_{1\le i\le N}$ et $\{\varphi_i\}_{1\le i\le N}$ les angles définissant les positions dans l'espace des signaux d'entrée.
- Quantification.
- Formatage du flux binaire : ce bloc effectuera un réarrangement du flux binaire permettant une troncature facile. En effet, le but de cet étage de traitement est de rendre le flux binaire flexible et facile à adapter aux contraintes.
- Modèle psychoacoustique. Il faut noter que l'on doit tenir compte du domaine d'écoute final des signaux pour appliquer ce modèle en combinaison avec la quantification. Dans notre cas, nous utilisons un modèle psychoacoustique monophonique sur chaque signal de la scène sonore combiné avec une nouvelle méthode de quantification.

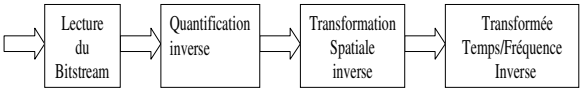

Figure 3 – *Schéma du décodeur générique*

Le schéma du décodeur (Figure 3) est constitué des blocs suivant :

- Bloc de lecture du bitstream.

- Quantification inverse.
- Décodage spatial : effectue une restitution de la scène sonore encodée sur un système de rendu donnée. La transformée de ce décodage dépend donc du système de rendu sonore du terminal [1] [4]. Il est caractérisé donc par les paramètres $\left(N', p', Q', \{\theta_i'\}_{1\le i\le Q'}, \{\varphi_i'\}_{1\le i\le Q'}\right)$.
- Transformée temps/fréquence inverse de type MDCT inverse.

3.2 Calcul d'un critère de la qualité auditive

Un critère objectif de la qualité auditive globale de notre chaîne de transmission sera la comparaison des MNR des différents signaux d'entrée. Cette solution n'a pas pour objectif de déterminer un critère optimal de masquage auditif des bruits de quantification dans le domaine d'écoute (celui imposé par le système de rendu sonore). Ceci est dû au fait qu'au niveau du codeur nous n'avons aucune information sur le type du système de rendu qui sera utilisé. Une solution sous-optimale adoptée consiste à utiliser une matrice de décodage Ambisonic calculée pour un système de haut-parleurs réguliers et dont le nombre est égal au nombre de sources en entrée. L'influence du décodage sur les signaux en sortie sera ainsi prise en compte au niveau du codeur.

Soit la chaîne de codage décrite dans la Figure 4.

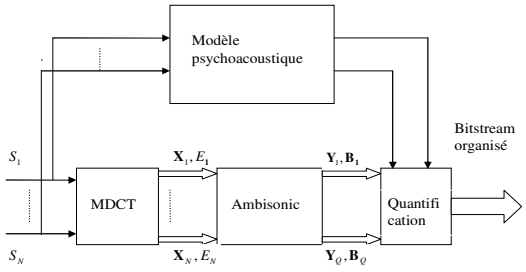

Figure 4 – *Système de codage des signaux d'entrée*

Le calcul du MNR dans le domaine du signal nécessite la connaissance des erreurs de quantification $(E_i)_{1\le i\le N}$ des signaux $(S_i)_{1\le i\le N}$ [2].

On note :

- $\mathbf{D}$ la matrice d'encodage Ambisonic.
- $\mathbf{E} = \begin{pmatrix} E_1 \\ \vdots \\ E_N \end{pmatrix}$ et $\mathbf{B} = \begin{pmatrix} B_1 \\ \vdots \\ B_Q \end{pmatrix}$ les bruits de quantification respectivement dans le domaine du signal et dans le domaine Ambisonic.
- $\mathbf{X}_j = \begin{pmatrix} X_{1,j} \\ \vdots \\ X_{N,j} \end{pmatrix}$ et $\mathbf{Y}_j = \begin{pmatrix} Y_{1,j} \\ \vdots \\ Y_{Q,j} \end{pmatrix}$, $1 \le j \le M$, les composantes fréquentielles des signaux d'entrées respectivement dans le domaine de la transformée MDCT et dans le domaine MDCT/Ambisonic, avec $\mathbf{Y}_j = \mathbf{D}.\mathbf{X}_j$.

On peut écrire que: $\mathbf{B} = \mathbf{D}.\mathbf{E}$, donc $\mathbf{E} = \mathbf{D}^{-1}\mathbf{B}$. Les différents bruits de quantification dans le domaine du signal sont ainsi déterminés. La courbe de masquage des signaux d'entrée étant connue au niveau du codeur, les rapports masque sur bruit (MNR) dans chaque sous bande peuvent être déterminés directement.

3.3 Calcul d'un critère de la qualité spatiale

3.3.1 Principe

L'idée est d'utiliser les critères de Gerzon [3] [4] pour déterminer, pour un seuil donné, si les directions de provenance des sources constituant la scène spatiale ont été respectées. Les critères de Gerzon sont utilisés généralement pour déterminer la matrice de décodage Ambisonic correspondant à une configuration donnée de haut-parleurs. En effet, ils fournissent des conditions suffisantes pour que la restitution spatiale des sources sonores soit optimale.

Dans notre cas ces critères sont exploités pour déterminer l'influence de la quantification sur le positionnement des sources après restitution pour une matrice de décodage donnée. Ainsi l'utilisation de cette méthode dans le cadre d'un codeur audio à scalabilité spatiale peut se combiner avec un algorithme de mise en forme du flux binaire en effectuant à chaque itération, et pour un seuil donné, un test sur le degré du respect de la direction de la source. Il est possible alors de construire un flux binaire adapté aux différentes contraintes imposées par le débit, le degré de la précision spatiale désiré et le système de rendu sonore.

3.3.2 Fonctionnement

Considérons la chaîne de codage définie dans la Figure 2. Le flux binaire ainsi constitué lors du codage peut être manipulé de façon à déterminer les éléments les moins pertinents au sens de la précision spatiale. Pour réaliser un codeur scalable au sens de la précision spatiale nous adoptons un critère de spatialisation basé sur les angles de Gerzon.

Soit le vecteur suivant: $\boldsymbol{\theta} = \begin{pmatrix} \theta_V \\ \theta_E \end{pmatrix}$ représentant l'angle de Gerzon généralisé. On définit la métrique $\Delta\boldsymbol{\theta}_j(n)$ définie par $\Delta\boldsymbol{\theta}_j(n) = \left\|\boldsymbol{\theta}_j(D_{n+1}) - \boldsymbol{\theta}_j(D_n)\right\|_2$, avec $\boldsymbol{\theta}_j(D_n)$ représente l'angle de Gerzon calculé dans la sous bande j pour le débit D_n et $\|.\|_2$ représente la norme 2 de $\mathbb{R}^2$.

Soit $\mathbf{P} = \begin{pmatrix} P_1 \\ \vdots \\ P_{Q'} \end{pmatrix}$ les signaux issues des différents hauts parleurs. Les vecteurs de Gerzon sont définis selon les formules suivantes [3] [4]:

$$\vec{V} = \begin{cases} x_V = \dfrac{\sum_{1\le i\le Q'} P_i \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i} = r_V \cos\theta_V \\ y_V = \dfrac{\sum_{1\le i\le Q'} P_i \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i} = r_V \sin\theta_V \end{cases} \quad (1)$$

$$\vec{E} = \begin{cases} x_E = \dfrac{\sum_{1\le i\le Q'} P_i^2 \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i^2} = r_E \cos\theta_E \\ y_E = \dfrac{\sum_{1\le i\le Q'} P_i^2 \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i^2} = r_E \sin\theta_E \end{cases} \quad (2)$$

Les vecteurs $\vec{V}$ et $\vec{E}$ représente respectivement les vecteurs de vélocité et d'énergie.

Note: Le calcul qui suit, utilisée pour déterminer les angles de Gerzon, ne prend pas en compte le couple (r_V, r_E). En effet, ce couple dépend seulement de la matrice de décodage Ambisonic utilisée. Selon les critères de Gerzon cette matrice doit permettre de s'assurer que le couple (r_V, r_E) tend vers une valeur proche de $(1,1)$ pour avoir une restitution spatiale parfaite de la scène sonore. Vu que ce couple est constant pour une matrice de décodage donnée, alors il est possible de choisir le cas parfait $(r_V, r_E) = (1,1)$.

On peut remarquer que sous les critères de Gerzon $\dfrac{\sum_{1\le i\le Q'} P_i \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i}$, $\dfrac{\sum_{1\le i\le Q'} P_i \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i}$, $\dfrac{\sum_{1\le i\le Q'} P_i^2 \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i^2}$ et $\dfrac{\sum_{1\le i\le Q'} P_i^2 \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i^2} \in [-1,1]$, donc il existe un unique couple d'angle (θ_V, θ_E) tel que :

$$\begin{cases} \cos\theta_V = \dfrac{\sum_{1\le i\le Q'} P_i \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i} \\ \sin\theta_V = \dfrac{\sum_{1\le i\le Q'} P_i \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i} \end{cases} \text{ et } \begin{cases} \cos\theta_E = \dfrac{\sum_{1\le i\le Q'} P_i^2 \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i^2} \\ \sin\theta_E = \dfrac{\sum_{1\le i\le Q'} P_i^2 \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i^2} \end{cases}$$

Ainsi

$$\begin{cases} \theta_V = sign\left(\dfrac{\sum_{1\le i\le Q'} P_i \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i}\right).\arccos\left(\dfrac{\sum_{1\le i\le Q'} P_i \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i}\right) \\ \theta_E = sign\left(\dfrac{\sum_{1\le i\le Q'} P_i^2 \sin\varphi'_i}{\sum_{1\le i\le Q'} P_i^2}\right).\arccos\left(\dfrac{\sum_{1\le i\le Q'} P_i^2 \cos\varphi'_i}{\sum_{1\le i\le Q'} P_i^2}\right) \end{cases} \quad (3)$$

Le calcul des angles de Gerzon dans une sous bande se fait en utilisant les signaux en sortie des hauts parleurs pour la sous bande correspondante. Ces signaux sont calculés à l'aide de la matrice du décodage Ambisonic, puis ils sont injectés dans les formules de Gerzon (3). L'angle de Gerson $\boldsymbol{\theta}$ est donc fonction des signaux $\mathbf{P}$.

D'autre part $\mathbf{P} = \mathbf{Q}\mathbf{Y}'$, avec $\mathbf{Q}$ est la matrice de décodage spatiale utilisée et $\mathbf{Y}'$ est le vecteur des composantes en sortie du bloc de la quantification inverse. Nous avons $\mathbf{Y}' = \mathbf{Y} + \mathbf{B}$ donc : $\mathbf{P} = \mathbf{Q}(\mathbf{Y} + \mathbf{B})$ (4)

A partir des équations (3) et (4) on peut déduire que les angles de Gerzon dépendent du bruit de quantification. Il est possible alors d'utiliser ces angles pour définir un critère sur la qualité spatiale.

4 Algorithmes d'ordonnancements

4.1 Algorithme basé sur le critère du MNR

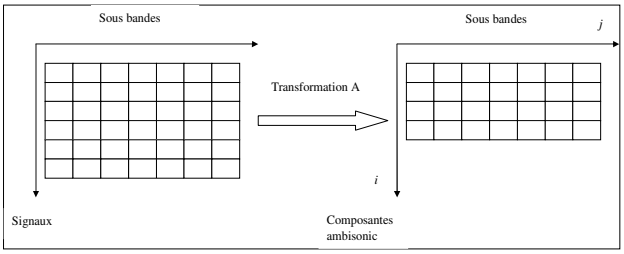

Figure 5 – *Application de la transformée Ambisonic sur un ensemble de signaux en sous bandes*

Soit D_n le débit à l'itération n et $D_{n+1} = D_n - \delta_n$ avec δ_n est le pas de l'itération à l'instant n. Les étapes de l'algorithme itératif se déroule comme suit :

1. Quantification de l'ensemble des composantes Ambisonic avec les débits D_n et D_{n+1}. Puis calcul des MNR résultants de chaque quantification.
2. Calcul de la matrice $\Delta\mathbf{MNR}(n) = \left|\mathbf{MNR}(D_n) - \mathbf{MNR}(D_{n+1})\right|$, avec $\mathbf{MNR}(D_n)$ la matrice des valeurs des MNR calculés dans chaque sous-bandes pour D_n.
3. Détermination du couple $(i_n, j_n) = \arg\min_{(i,j)\in\mathbb{E}_n} \left\|\Delta\mathbf{MNR}(n)\right\|$ avec $\mathbb{E}_n$ est l'ensemble des couples d'indice des composantes Ambisonic restantes à l'itération n. Le couple (i_n, j_n) représente alors la sous bande la moins pertinente au sens de la qualité auditive dans le domaine Ambisonic/fréquence.
4. Mise à la fin du flux binaire de la sous bande d'indice (i_n, j_n).
5. La sous bande d'indice (i_n, j_n) est supprimée dans la suite des itérations.

Les étapes 1 à 5 sont itérées tant de fois que nécessaire pour ordonner les composantes Ambisonic selon leur pertinence sur la qualité auditive.

4.2 Algorithme basé sur le critère des angles de Gerzon

Soit D_n le débit à l'itération n et $D_{n+1} = D_n - \delta_n$ avec δ_n est le pas de l'itération à l'instant n.

1. Quantification de l'ensemble des composantes Ambisonic avec les débits D_n et D_{n+1}. Puis calcul des angles de Gerzon résultant de chaque quantification.
2. Calcul du $\Delta\boldsymbol{\theta}_j(n) = \left\|\boldsymbol{\theta}_j(D_{n+1}) - \boldsymbol{\theta}_j(D_n)\right\|_2$, $1 \le j \le M$, avec $\boldsymbol{\theta}_j$ représente l'angle de Gerzon calculé dans la sous bande j.
3. Détermination de $j_n = \arg\min_{j \in [\![1, s_{Max}]\!]} \left(\Delta\boldsymbol{\theta}_j(n)\right)$, avec s_{Max} est le nombre total de sous bande. Soit $(Y_{i,j_n})_{1 \le i \le p}$ les sous bandes constituant la bande fréquentielle d'indice j_n.
    a. Suppression lors de l'itération courante de la sous bande d'indice Y_{i,j_n}.
    b. Calcul de l'angle de Gerzon dans ce cas là. Puis calcul de la différence $\Delta\boldsymbol{\theta}_{i,j_n}(n) = \left\|\boldsymbol{\theta}_{j_n}(Y_{i,j_n} = 0, D_n) - \boldsymbol{\theta}_{j_n}(D_n)\right\|_2$
    c. Restitution de la sous bande Y_{i,j_n} dans les itérations et suppression de la sous bande suivante Y_{i+1,j_n}, et on refait les étapes de b et c tant que $i \le p$.
    d. Détermination du $i_n = \arg\min_{i \in [\![1,p]\!]} \Delta\boldsymbol{\theta}_{i,j_n}(n)$.
4. On place la sous bande d'indice (i_n, j_n) à la fin du flux binaire et ne sera pas prise en compte dans la suite de l'algorithme. Cette sous bande est donc la moins pertinente au sens de la précision spatiale au cours de cette itération selon le critère choisi.

Les étapes 1 à 4 sont itérées tant de fois que nécessaire pour ordonner les composantes Ambisonic selon leur pertinence sur qualité spatiale.

4.3 Scalabilité hybride

Dans le cadre de la réalisation d'un codeur à scalabilité hybride, nous pouvons fixer dès le départ et selon les cas d'usages des critères qui peuvent favoriser une meilleure qualité auditive ou une meilleure qualité de restitution spatiale. Cela est dû principalement au fait que les deux algorithmes fonctionnent indépendamment. Ainsi en utilisant les deux algorithmes simultanément et en se basant sur les critères imposés par les contraintes de départ, on réalise un bitstream organisé d'une façon optimale qui peut être transmis entièrement ou juste partiellement sans trop de dégradation ni auditive ni spatiale. A chaque itération les algorithmes effectuent un ordonnancement des éléments les moins pertinents pour la qualité auditive et la qualité spatiale. Le flux binaire en sortie du codeur se retrouve ordonné pour les débits utilisés selon le critère de la pertinence. Donc ce nouveau codeur est capable de fonctionner pour les différents débits utilisés lors des itérations.

5 Conclusion

Dans cet article nous avons décrit un nouveau type de codeur audio 3D combinant un codeur monophonique et une transformation spatiale de type Ambisonic. Associé à de nouveaux algorithmes d'ordonnancement du flux binaire, ce codeur fournit une scalabilité hybride. Les algorithmes décrits fixent des critères sur la qualité auditive et spatiale qui peuvent être utilisés pour déterminer la pertinence des composantes d'un signal sur l'aspect immersif de l'utilisateur dans la scène sonore 3D.

Références

[1] J. Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Thèse de doctorat de l'université Paris 6, 31 juillet 2001.

[2] O. Derrien. *Optimisation de la quantification par modèles statistiques dans le codeur MPEG Advanced Audio Coder (AAC) - Application à la spatialisation d'un signal comprimé en environnement MPEG-4*. Thèse de doctorat de l'ENST, 22 novembre 2002.

[3] M. A. Gerzon. Criteria for Evaluating Surround-Sound Systems. *J. Audio Eng. Soc*., 25:400–408, June 1977.

[4] M. A. Gerzon. Ambisonics in Multichannel Broadcasting and Video. *J. Audio Eng. Soc.*, 33(11):859–871, November 1985.

[5] M. Poletti. The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems. *J. Audio Eng. Soc.*, 44(11):948–963, November 1996.

[6] *"Text of ISO/IEC FDIS 23003-1, MPEG Surround"*, ISO/IEC JTC 1/SC 29/WG 11 N8324, July 2006, Klagenfurt, Austria.