

Production, transmission et restitution temps réel d'une scène sonore dans un format audio 3D flexible

Sébastien Moreau¹ Jérôme Daniel¹ Adil Chraa¹

¹ France Télécom R&D - DIH / IPS

2, avenue Pierre Marzin, 22 307 Lannion cedex – France

{sebastien.moreau, jerome.daniel, adil.chraa}@francetelecom.com

Résumé

Cet article présente une plateforme de production, transmission et restitution temps réel d'une scène sonore 3D. Parmi les formats audio existants, le format High Order Ambisonics (HOA) a été retenu pour sa souplesse d'utilisation. Des outils d'exploitation, inexistant pour ce format encore récent, ont été élaborés pour développer la plateforme temps réel. Cette dernière est constituée d'un microphone haute résolution, d'un module de transmission/réception par réseau IP, et d'un décodeur qui permet la restitution sur divers systèmes de diffusion. L'implantation a été effectuée avec la technologie VST de Steinberg et s'adapte à beaucoup de logiciels audio grand public. La plateforme réalisée peut concerner la diffusion musicale, radiophonique, etc.

Mots clefs

High Order Ambisonics, spatialisation sonore, format audio.

1 Introduction

Nous présentons ici un système temps réel de spatialisation sonore dont le but est de transmettre et de reconstruire un champ sonore 3D naturel dans un lieu distant du champ original. Un format de représentation du champ acoustique a été choisi parmi ceux existants en fonction de critères objectifs tels que la souplesse d'utilisation et la quantité d'information à transmettre.

2 Représentation d'une scène sonore 3D : quel format ?

Une technique de spatialisation sonore permet de représenter les informations spatiales et temporelles d'un champ acoustique dans le but de les stocker, transmettre, ou reproduire. Ceci pose le problème de la définition d'un format de représentation du champ acoustique.

2.1 Considérations préliminaires

La Figure 1 représente les principes de bases de la spatialisation sonore : l'encodage spatial d'une scène

sonore dans un format de représentation, puis le décodage de la scène formatée permettant sa reproduction.

Le format de représentation contient les informations spatiales et temporelles du champ sonore nécessaires à sa reconstruction. Il décrit donc, dans les limites de ses capacités, un champ sonore. Il est constitué d'un ensemble de signaux temporels auxquels peuvent être associés des paramètres de spatialisation.

L'encodage spatial désigne le processus qui permet d'acquérir, de traduire une scène sonore dans un format spécifique de représentation.

Le décodage spatial concerne quant à lui l'adaptation des données formatées au système de restitution (casque, 5.1, etc.). Si le format est constitué de signaux destinés à alimenter directement un système de diffusion particulier, le décodage n'est pas nécessaire (binaural, stéréophonie).

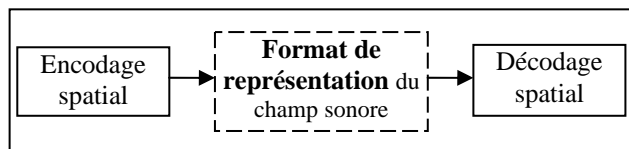


Figure 1 – Principe de base de la spatialisation sonore

2.2 Choix du format

Le choix du format de représentation d'une scène sonore est un choix décisif puisque de lui dépend la qualité potentielle de la reconstruction spatiale des sources sonores. Pour le développement de notre application temps réel, le choix s'est effectué selon trois critères :

- la possibilité de manipuler la scène sonore encodée ;
- la flexibilité par rapport au système de restitution ;
- la concision de la scène sonore encodée.

On entend par manipulation d'une scène sonore encodée le fait de modifier la position spatiale des sources sonores déjà présentes.

D'autre part, un format est dit flexible s'il peut s'adapter à divers systèmes de restitution pour la reconstruction de la scène sonore.

Enfin, la concision de la représentation est une contrainte imposée par les ressources matérielles limitées. Il est donc avantageux qu'un format de représentation contienne peu d'informations redondantes ou qu'il puisse être compressé

sans entraîner une dégradation de la qualité (du moins perceptive) de reproduction.

2.3 Comparaison des principaux formats existants

Le Tableau 1 permet de comparer les formats associés aux principales techniques de spatialisation existantes suivant les trois critères définis plus haut. Ces principaux formats sont : la *stéréophonie* (deux canaux ou plus) ; le *binaural* ; *High Order Ambisonics* (HOA) ; et le *tout paramétrique*.

| | Manipulation de la scène sonore | Dispositif de restitution | Taille de la scène encodée |
|---|-------------------------------------|----------------------------------|--|
| Stéréophonie (2.0, 5.1, etc.) | Impossible, scène sonore figée | fixe, lié au format | 1 signal par HP |
| Binaural | | | 2 signaux |
| HOA (ordre M) | globale (rotation, etc.) | adaptable, indépendant du format | 3D : $(M+1)^2$ 2D : $(2M+1)$ signaux |
| tout paramétrique | chaque source de façon indépendante | | 1 signal par source + paramètres de position |

Tableau 1 - Comparaison des principaux formats de représentation de scène sonore

Le format *stéréophonique* tel que défini ici, est constitué de signaux destinés à alimenter directement les haut-parleurs d'une configuration prédéfinie (2.0, 4.0, 5.1, etc.). Le format *binaural* est composé de deux signaux, appelés *signaux binauraux*, contenant les transformations subies par le son lors de son interaction avec le corps de l'auditeur (pavillons, tête et torse) et devant être reproduit au niveau des tympanes au moyen d'un casque.

Le format associé à la technologie *High Order Ambisonics* (HOA) est constitué de signaux appelés *signaux HOA* qui résultent de la décomposition en harmoniques sphériques d'un champ sonore en un point. Le nombre de signaux dépend de l'ordre M de la décomposition : $(M+1)^2$ en 3D et $(2M+1)$ en 2D. De l'ordre dépend également la résolution spatiale de la description.

Enfin, un format *tout paramétrique* contient autant de signaux que de sources sonores, auxquels sont associés des paramètres de position spatiale. La reconstruction est ensuite effectuée lors du décodage selon la technique de spatialisation choisie (*binaural*, *HOA*, *Wave Field Synthesis*, etc.) et le dispositif de restitution.

2.4 HOA : un format avantageux

Les formats *HOA* et *tout paramétrique* sont particulièrement intéressants car ils permettent de

représenter une scène sonore 3D indépendamment du système de restitution. Bien que le second paraisse offrir plus de possibilités en terme de manipulation de scène sonore, il comporte un nombre important de signaux lorsque le champ acoustique encodé est complexe. De plus il suppose la prise de son isolée des sources sonores et s'adapte donc mal à l'enregistrement naturel en 3D.

Le format HOA semble réaliser un bon compromis entre taille et souplesse d'utilisation. Il a donc logiquement été choisi pour développer notre application temps réel. Cependant, les outils de production qui lui sont associés sont à l'heure actuelle relativement limités. Nous en avons nous-même développés sur les bases théoriques suivantes.

3 High Order Ambisonics

Cette section introduit les éléments théoriques qui nous permettront d'exploiter le format HOA dans notre application temps réel. Seuls certains aspects correspondant aux fonctions requises seront abordés : l'encodage de champ sonore naturel, les rotations du champ sonore, et le décodage pour divers systèmes de restitution.

3.1 Série de Fourier-Bessel

La technologie *High Order Ambisonics* (HOA) se base sur une représentation harmonique spatiale du champ sonore, solution de l'équation d'onde exprimée en coordonnées sphériques (azimut θ , élévation δ , rayon r). Sur une zone exempte de source sonore, cette solution est la *série de Fourier-Bessel* :

$$p(\theta, \delta, kr) = \sum_{m=0}^{\infty} j_m^m j_m(kr) \sum_{0 \leq n \leq m, \sigma = \pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \delta). \quad (1)$$

Les fonctions radiales $j_m(kr)$ sont les *fonctions de Bessel sphériques* de première espèce d'ordre m (k étant la longueur d'onde). Les fonctions angulaires $Y_{mn}^{\sigma}(\theta, \delta)$ sont les *fonctions harmoniques sphériques*. Et enfin, les signaux B_{mn}^{σ} exprimés ici dans le domaine fréquentiel, sont les *signaux HOA* qui décrivent le champ sonore sur la zone considérée [1].

Notons que les harmoniques sphériques sont orthogonales deux à deux, ce qui se traduit mathématiquement par l'équation suivante :

$$\int_S Y_{mn}^{\sigma}(\theta, \delta) Y_{m'n'}^{\sigma'}(\theta, \delta) dS = \delta_{mm'} \delta_{nn'} \delta_{\sigma\sigma'} \quad (2)$$

où $\delta_{mm'}$ est le symbole de Kronecker, l'intégration se faisant sur la sphère unité.

En tronquant la série (1) à un ordre fini M ($m \leq M$), on obtient une approximation du champ sonore sur une zone de l'espace d'autant plus grande que M est élevé.

3.2 Encodage de champ sonore naturel 3D

L'encodage de champ naturel au format HOA consiste à extraire en un point d'un champ acoustique les signaux

B_{mn}^σ . La stratégie adoptée consiste à mesurer la pression sonore à la surface d'une sphère rigide de rayon a fixé. Celle-ci peut s'exprimer de la façon suivante :

$$p_a(\theta, \delta) = \sum_{m=0}^{\infty} W_m(ka) \sum_{0 \leq n \leq m, \sigma = \pm 1} B_{mn}^\sigma Y_{nm}^\sigma(\theta_q, \delta_q) \quad (3)$$

avec
$$W_m(ka) = \frac{i^{m-1}}{(ka)^2 h_m^-(ka)}$$

En utilisant la *propriété d'orthogonalité* des harmoniques sphériques (2), on peut calculer d'après l'équation (3) les signaux HOA (projection sur la base des harmoniques sphériques) :

$$B_{mn}^\sigma = \frac{1}{W_m(ka)} \int_S p_a(\theta_q, \delta_q) Y_{nm}^\sigma(\theta_q, \delta_q) dS. \quad (4)$$

En pratique, la pression sonore n'est mesurée qu'en un nombre fini Q de positions à la surface de la sphère. On effectue alors une approximation de la décomposition harmonique jusqu'à un ordre M maximal restreint :

$$B_{mn}^\sigma = \frac{1}{W_m(ka)} \sum_{q=1}^Q p_a(\theta_q, \delta_q) Y_{nm}^\sigma(\theta_q, \delta_q), \quad m \leq M. \quad (5)$$

Le nombre de signaux HOA estimé est alors de $K=(M+1)^2$. Pour une bonne estimation, on choisit Q tel que $Q \geq K$. D'autre part, l'équation (5) suppose que la base échantillonnée des harmoniques sphériques (représentée par les positions des microphones) est orthogonale. Dans des conditions plus générales, on préférera à cette *méthode de projection*, une méthode d'estimation basée sur une *approximation aux moindres carrés* [2].

Quelque soit la méthode choisie, les signaux HOA sont obtenus par *matriçage* puis *égalisation* $EQ_m = 1/W_m(ka)$ des signaux microphoniques. Cette égalisation nécessite, notamment en basse fréquence, d'être limitée en amplitude pour sa mise en œuvre [2].

3.3 Rotation de la scène sonore

Lorsque la scène sonore pivote (suivant trois degrés de liberté), le groupe formé par les $(2m+1)$ signaux HOA de même ordre m est transformé de la façon suivante :

$$b'_m = R_m \cdot b_m, \quad (6)$$

avec $b'_m = [b'_{mm}, b'_{m,m-1}, \dots, b'_{m,m}, b'_{m,m-1}, \dots, b'_{m0}]^t$ et

$$b_m = [b_{mm}, b_{m,m-1}, \dots, b_{m,m}, b_{m,m-1}, \dots, b_{m0}]^t.$$

b'_m est le vecteur contenant les $(2m+1)$ signaux HOA, exprimés dans le domaine temporel, résultant de la manipulation, R_m la matrice de dimensions $(2m+1) \times (2m+1)$ caractérisant la rotation, et b_m le vecteur contenant les $(2m+1)$ signaux HOA de départ, exprimés également dans le domaine temporel.

La rotation de scène sonore est ici destinée à être utilisée en association avec un *head-tracker* pour corriger les mouvements de tête en écoute binaurale (avec casque). Nous ne considérons donc ici que les rotations dans le

plan azimutal (rotation gauche-droite). Dans ce cas particulier, chaque couple de signaux (b_{mn}^l, b_{mn}^{l-1}) , m et n étant fixés, s'obtiennent pour une rotation d'angle θ' d'après la relation :

$$\begin{bmatrix} b_{mn}^l \\ b_{mn}^{l-1} \end{bmatrix} = \begin{bmatrix} \cos(n\theta') & -\sin(n\theta') \\ \sin(n\theta') & \cos(n\theta') \end{bmatrix} \begin{bmatrix} b_{mn}^1 \\ b_{mn}^{-1} \end{bmatrix}. \quad (7)$$

Nous pouvons ainsi calculer la matrice de rotation en fonction de l'ordre m considéré. Par exemple, pour l'ordre $m = 2$, la matrice caractérisant une rotation θ' en azimut est la suivante :

$$R_2 = \begin{bmatrix} \cos 2\theta' & -\sin 2\theta' & 0 & 0 & 0 \\ \sin 2\theta' & \cos 2\theta' & 0 & 0 & 0 \\ 0 & 0 & \cos \theta' & -\sin \theta' & 0 \\ 0 & 0 & \sin \theta' & \cos \theta' & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (8)$$

3.4 Décodage

Le *décodage* consiste à adapter une scène sonore encodée au format HOA à un dispositif particulier de L haut-parleurs, que l'on choisit concentrique. Les haut-parleurs sont au moins aussi nombreux que les K signaux HOA.

On considère que les haut-parleurs émettent des ondes planes (champ lointain) qui contribuent à la reconstruction des signaux HOA suivant l'équation matricielle :

$$b = C \cdot s, \quad (9)$$

où $b = (b_{00}^+, b_{11}^+, b_{11}^-, \dots, b_{mm}^\sigma)^t$ est le vecteur constitué par les signaux HOA (domaine temporel), C est la matrice contenant les gains $Y_{mn}^\sigma(\theta_l, \delta_l)$ associés à la direction l de chaque haut-parleur, et enfin $s = (s_1, s_2, \dots, s_L)^t$ est le vecteur contenant les signaux temporels émis par les haut-parleurs.

Ces signaux sont obtenus en inversant l'équation (9), soit par combinaison (*matriçage*) des signaux HOA :

$$s = D \cdot b \quad (10)$$

avec

$$D = C^t (C C^t)^{-1}.$$

D est appelé la *matrice de décodage*, c'est la matrice pseudo-inverse de C .

Si on désire effectuer un décodage pour le casque d'écoute, il est nécessaire d'ajouter une étape de simulation binaurale des haut-parleurs : on applique aux signaux issue de (10) des filtres binauraux $h_G(\theta, \delta)$ et $h_D(\theta, \delta)$ caractérisant le canal acoustique entre les haut-parleurs et les oreilles gauche et droite.

Pour plus d'efficacité, on synthétise les opérations de *matriçage* et de *filtrage* binaural sous la forme d'un ensemble f de filtres s'appliquant directement sur les signaux HOA. On obtient :

$$f_G = h_G \cdot D \text{ et } f_D = h_D \cdot D, \quad (11)$$

avec $f_G = [f_{G00}^1, \dots, f_{Gmm}^\sigma]$, $f_D = [f_{D00}^1, \dots, f_{Dmm}^\sigma]$,

$$h_G = [h_G(\theta_1, \delta_1), \dots, h_G(\theta_L, \delta_L)], \quad h_D = [h_D(\theta_1, \delta_1), \dots, h_D(\theta_L, \delta_L)].$$

Le décodage consiste alors à effectuer les opérations de filtrage suivantes (* désigne l'opération de convolution) :

$$g = \sum_{m,n,\sigma} b_{mn}^{\sigma} * f_{Gmn}^{\sigma} \quad \text{et} \quad d = \sum_{m,n,\sigma} b_{mn}^{\sigma} * f_{Dmn}^{\sigma}, \quad (12)$$

où g et d sont les signaux binauraux gauche et droit.

4 Plateforme temps réel

La technologie HOA a donc été choisie pour réaliser un système d'acquisition, transmission, et restitution temps réel d'une scène sonore 3D naturelle. Plusieurs applications sont visées : diffusion sur Internet d'émissions radiophoniques spatialisées, de concerts musicaux ; partage d'ambiance sonore, etc.

4.1 Description

La Figure 2 montre les différents éléments constituant la plateforme développée. Un *Encodeur* encode un champ sonore naturel 3D au format HOA. La scène encodée est ensuite transmise au *décodeur* qui permet de l'adapter à divers systèmes de restitution.

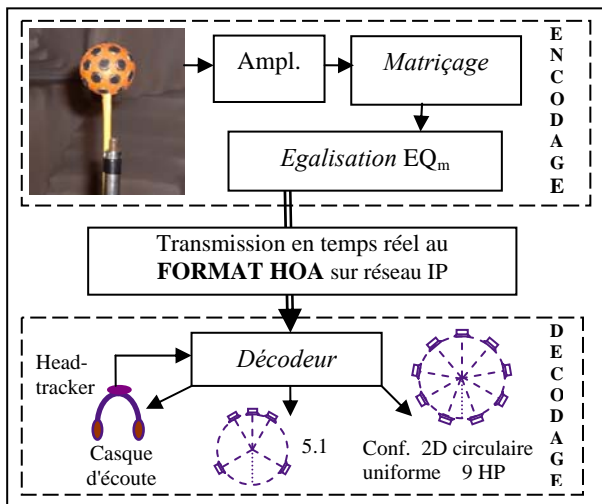


Figure 2 - Description de la plateforme temps réel

L'encodage du champ sonore 3D s'effectue grâce à un microphone pouvant estimer les signaux HOA jusqu'à l'ordre 4. Il est constitué de 32 capsules disposées à la surface d'une sphère rigide de rayon 2,6 cm. Les 32 signaux acquis sont matricés puis égalisés (section 3.2) de façon à obtenir un champ sonore encodé au format HOA indépendant du système microphonique.

Le décodage adapte la scène sonore encodée à divers systèmes de restitution : 5.1, configuration 2D circulaire uniforme de 9 haut-parleurs, casque d'écoute avec système de *head-tracking*, ce dernier donnant au décodeur en temps réel les informations de position de tête.

4.2 Implantation

La technologie VST (*Virtual Studio Technology*) de Steinberg a été choisie pour l'implantation des différents

modules de la plateforme décrite ci-dessus. Cette technologie permet d'intégrer dans un grand nombre de logiciels audio professionnels et grand public (Cubase, Wavelab, Nuendo, etc.), sous forme de *plugins*, les traitements relatifs à la technologie HOA.

Quatre *plugins VST* ont été développés :

- Un premier concerne les traitements de matricage/égalisation des signaux microphoniques.
- Un *plugin serveur* permet ensuite l'envoi de la scène encodée sur le réseau vers une adresse IP.
- Un *plugin client* a en charge la réception, sur un ordinateur distant, de la scène transmise.
- Enfin, un dernier *plugin* effectue le décodage de la scène sonore encodée. Il prend comme paramètre la nature du dispositif de restitution choisi.

4.3 Transmission de la scène sonore 3D

La transmission de la scène sonore 3D encodée est une étape délicate puisqu'elle doit se faire en temps réel et concerne une quantité d'information importante : à l'ordre 4, 25 signaux HOA sont nécessaires en 3D.

Il peut donc s'avérer indispensable de réduire cette quantité d'information. Plusieurs stratégies sont étudiées. La première suppose que l'on connaisse à l'avance le dispositif de restitution et consiste à effectuer le décodage avant transmission, réduisant ainsi le nombre de canaux au nombre de haut-parleurs ou à deux pour le casque d'écoute. Il est également possible de réduire la résolution spatiale (ordre $M < 4$, restriction 3D → 2D). On peut enfin avoir recours à la compression audio (MP3, AAC).

Différents tests de transmission sur un LAN et sur Internet sont actuellement menés. Ils permettront d'évaluer, en fonction du débit disponible, l'impact de la réduction des informations transmises sur la qualité de restitution.

5 Conclusion

Un système d'acquisition, transmission et restitution en temps réel d'une scène sonore 3D a été développée et est actuellement en cours de test. Elle est basée sur la technologie *High Order Ambisonics* dont l'intérêt réside dans la souplesse d'utilisation, notamment pour la restitution. Un certain nombre d'outils, encore inexistant pour cette technologie récente, ont été conçus pour mener à bien ce projet. Les applications possibles sont multiples : diffusion radiophonique, musicale, partage d'ambiance sonore, etc.

Références

- [1] J. Daniel, R. Nicol, et S. Moreau. Further Investigations of Higher Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging, 114^{ème} convention de l'AES, Amsterdam, 2003.
- [2] S. Moreau et J. Daniel. Study of Higher Order Ambisonic Microphone. Dans *Actes de la conférence CFA/DAGA'04*, Strasbourg, Mars 2004.