Evaluation de systèmes d'encodage ambisonique aux ordres supérieurs.

Stéphanie Bertet[1], Jérôme Daniel[1], Laetitia Gros[1], Etienne Parizet[2], Olivier Warusfel[3]
[1] *France Télécom R&D, 22300 Lannion, France, courriel : stephanie.bertet@francetelecom.com,*
[2] *LVA, INSA-Lyon, 69000 Lyon, France, courriel : parizet@lva-insa.fr*
[2] *IRCAM, 75004 Paris, France, courriel : warusfel@ircam.fr*

Introduction

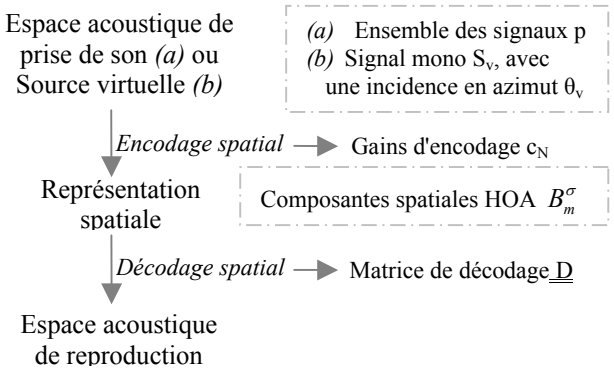
Différentes techniques de captation et reproduction d'une scène sonore 3D existent. Par exemple, pour la restitution stéréophonique frontale sur 2 haut-parleurs ou panoramique sur 5 haut-parleurs, il existe une variété de couples et arbres microphoniques qui leur sont dédiés. Nous nous intéressons ici à une approche plus « générique », dite « ambisonique », et plus généralement son extension aux ordres supérieurs, nommée « Higher Order Ambisonics » (HOA). Basée sur une décomposition du champ acoustique en « harmoniques sphériques », elle présente des propriétés intéressantes telles que: la scalabilité spatiale (la résolution spatiale de restitution peut être différente de celle d'enregistrement), la possibilité de manipuler le champ sonore avant diffusion, une flexibilité quant aux dispositifs d'écoute (restitution sur plusieurs configurations de haut-parleurs possibles, 5.1, restitution par casque). Le champ acoustique encodé peut provenir d'une synthèse de sources virtuelles ou d'un enregistrement avec un microphone HOA.

L'encodage HOA de champs sonores naturels (c'est-à-dire obtenus grâce à un enregistrement) est important pour constituer du contenu réel, réaliste, immersif. Il repose sur une prise de son par un réseau de microphones et ne peut réaliser que de façon approximative les formules d'encodage théoriques (par exemple il y a des artefacts spatiaux dû à l'espacement des capsules microphoniques). En 1993, le microphone Sound Field, ambisonique 1er ordre basé sur les théories de Gerzon, est commercialisé. Différentes études au sein de France Telecom R&D [1 2] ont abouti à 2 prototypes HOA, un prototype d'ordre 2, boule microphonique comprenant 12 capteurs et un prototype d'ordre 4, boule microphonique avec 32 capteurs.

Afin d'évaluer les effets de l'ordre ambisonique sur la résolution spatiale perçue et en particulier sur la localisation dans le plan horizontal, un test d'écoute est mis en œuvre.

L'approche ambisonique

Le traitement du champ (ou de la source virtuelle) peut se décomposer en deux étapes: l'encodage et le décodage.



L'encodage ambisonique est basé sur une décomposition du champ acoustique en « harmoniques sphériques », ou décomposition de Fourier-Bessel. En 2D, elle peut être représentée par une décomposition en harmoniques cylindriques (1).

$$p(r,\theta_r) = J_0(kr).B_0^+ + \sum_{m=1}^{M} 2j^m J_m(kr)\left[B_m^+ \cos m\theta_r + B_m^- \sin m\theta_r\right] \quad (1)$$

où les J_m (kr) représentent les fonctions de Bessel.

En tronquant la décomposition à un ordre M, on retient un nombre fini de signaux qui constituent le format d'encodage spatial ambisonique. Le champ sonore ainsi encodé est décrit avec une résolution spatiale d'autant plus fine que l'ordre M est grand avec une zone de reproduction élargie en fonction de la fréquence. Il en est de même du champ sonore reproduit, moyennant un décodage spatial et une diffusion sur un dispositif de haut-parleurs appropriés.

Soit une source virtuelle S_v, d'incidence θ_v; les composantes ambisoniques encodées jusqu'à l'ordre M sur le plan horizontal sont présentées équation 2.

$$B_m^\sigma \begin{cases} B_0^{+1} = 1.S_v \\ \quad = \vdots \\ B_m^{+1} = \sqrt{2}\cos m\theta_v.S_v \\ B_m^{-1} = \sqrt{2}\sin m\theta_v.S_v \end{cases} \quad (2)$$

où B_0^σ représente la composante omni (ordre 0) et B_m^σ les composantes d'ordre supérieur *m*.

Pour la reproduction sur haut-parleurs, les signaux S_n, à diffuser sur les n haut-parleurs sont obtenus par application d'une matrice de décodage D sur le vecteur de composantes spatiales B_m^σ : **S** = **D**.**B** [3].

La matrice de décodage « basique » est composée de gains à appliquer aux signaux encodés. Le décodage maxr_E « concentre » les contributions énergétiques des HP dans la direction voulue. La Figure 1 (b) montre la reconstruction des lobes de directivité d'un microphone virtuel équivalent correspondant au décodage « basique » ou « max r_E ».

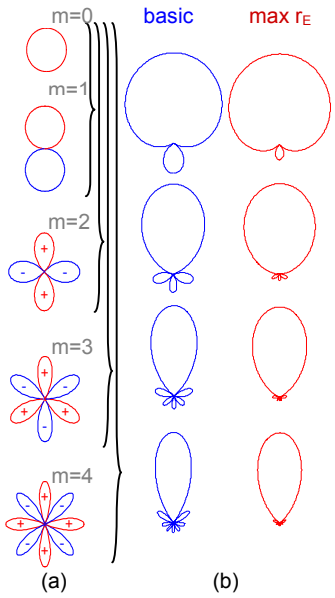


Figure 1. (a) Directivités des composantes encodées d'ordre *m*=0 à 4. (b) Directivités d'un microphone virtuel situé au centre du système de restitution, reconstruites après décodage basique ou max r_E des composantes ambisoniques associées.

Pour une reconstruction par un système de restitution sur haut-parleurs, dans le plan azimutal, des composantes spatiales d'ordre M, un nombre minimum de N=2M+2 haut-parleurs est recommandé.

Pour un enregistrement HOA, l'estimation des composantes spatiales repose en pratique sur une captation discrète du champ acoustique (i.e. par un réseau fini de microphones) associée à un traitement par une matrice de filtres. Cet échantillonnage spatial induit des artefacts d'encodage: d'une part l'aliasing spatial en haute fréquence, lié à l'écart entre capsules, d'autre part une directivité réduite (moindre résolution spatiale) en basse fréquence à cause des dimensions réduites de la boule microphonique [2]. Un compromis est donc à trouver sur la taille de la sphère.

Ecoute spatialisée

Notions de localisation

Les performances d'un système de reproduction 3D sont souvent évaluées à travers sa capacité à rendre l'effet de localisation d'une source sonore. Pour cela, on s'intéresse aux indices de localisation induits par une source sonore réelle. La position de la source est définie par son azimut, son élévation et sa distance dans un système de coordonnées sphériques centré sur la tête comprenant les plans médian, horizontal et frontal [4]. Une source placée sur le plan horizontal (hors du plan médian) sera à une distance différente entre l'oreille ipsilatérale et contralatérale provoquant des différences interaurales. Les différences interaurales de temps (ITD), pour les basses fréquences jusqu'à 1500Hz [5], sont prépondérantes sur les différences interaurales de niveau (ILD) opérant plutôt en haute fréquence. Les sources placées sur le cône de symétrie autour de l'axe interaural, appelé cône de confusion ont un ITD identique qui peut provoquer des confusions avant-arrière [5]. Pour lever ces ambiguïtés, l'audition spatiale exploite les modifications du spectre induites par le pavillon de l'oreille, qui varient en fonction de l'incidence du son: ce sont les indices spectraux.

Le flou de localisation d'une source varie selon la nature de la source : contenu fréquentiel et durée, pour une source située à l'azimut 0 dans le plan horizontal, le flou de localisation varie approximativement de 1° à 4° [4] ; et selon sa position. Considérant les signaux large bande, plusieurs études [4 6] montrent une précision de localisation entre ±3° et ±10° que la source soit à 0° devant l'auditeur, à 90° ou à 180°.

Reconstruction des indices de localisation avec des systèmes de haut-parleurs

La reproduction de scène sonore permet une reconstruction plus ou moins fidèle des indices de localisation. La dégradation (par rapport à une écoute naturelle) est quantifiable grâce à deux critères de qualité du système de reproduction, introduits par Gerzon, les vecteurs vélocité et énergie [6]. Notons que ces critères ne donnent pas un accès à la qualité de reconstruction des indices spectraux mais plutôt à celle des indices interauraux (ITD et ILD). Le module du vecteur énergie r_E ($r_E \leq 1$) peut-être interprété comme un facteur de réduction de l'effet de latéralisation [7]. Il dépend de l'ordre M. L'angle $\alpha_E = \text{acos}(r_E)$ caractérise le flou de la source virtuelle reconstruite [2].

Le test d'évaluation

L'étude présentée participe à une investigation plus globale sur l'impact d'une résolution spatiale « objective » (caractérisée notamment par l'ordre ambisonique) sur une résolution spatiale subjective et plus généralement l'ensemble des qualités spatiales perçues.

Le présent test est élaboré pour évaluer la qualité d'encodage des systèmes de prise de son HOA.

Les systèmes testés

Trois microphones ambisoniques ont été mesurés: le microphone Sound Field, dont les signaux sont encodés en B-format (ambisonique 1er ordre), le microphone 12 capsules (où les capsules sont positionnées en dodécaèdre sur une sphère semi-rigide de 7cm de diamètre), microphone ambisonique d'ordre 2 et le microphone 32 capsules (capsules positionnées en pentaki-dodécaèdre sur une sphère identique à celle utilisée pour le 12 capsules [2]) : microphone ambisonique d'ordre 4. Avec un traitement focalisé sur le champ horizontal (restitution 2D), la fréquence d'aliasing est autour de 7,5kHz pour le microphone 32 capsules et 4.4kHz pour le microphone 12 capsules.

Les réponses impulsionnelles des trois microphones ont été mesurées en chambre anéchoique tous les 5° sur le plan azimutal [2]. Les réponses ont été interpolées tous les degrés pour avoir une précision de 1° pour le test, vérifiant que cette interpolation n'introduit pas d'artefacts sur la bande passante audible. Les composantes ambisoniques des systèmes HOA (ordre 2 et ordre 4) sont calculées avec les réponses des microphones mesurés. Grâce aux mesures, l'influence des capsules, les différences inter capsules et les erreurs de positionnement potentielles sur la sphère sont prises en compte pour une définition optimale dans l'encodage ambisonique [2]. Les signaux du microphone SoundField enregistrés sont encodés en B-format directement.

Afin de tester l'influence des capsules positionnées hors du plan horizontal pour une restitution 2D, un système d'ordre 3 a été encodé utilisant les 8 capsules du plan horizontal du microphone 32 capsules.

Un cinquième système testé, un encodage ambisonique « idéal » d'ordre 4 (simulation d'une source virtuelle) est ajouté aux systèmes de prise de son en tant que référence.

Le décodage utilisé pour tous les systèmes est une combinaison d'un décodage « basic » (jusqu'à la fréquence limite de reconstruction basse fréquence, qui croît avec l'ordre M) et max r_E (pour optimiser la localisation dans le domaine haute fréquence complémentaire). Dans cette étude, nous nous concentrons sur l'encodage, le décodage est le même pour tous les systèmes. Une fréquence de coupure, fixe, optimale pour l'ordre 1 est choisie.

Le système de restitution

Un dodécagone de 1.5m de rayon est composé de 48 haut – parleurs Studer (4 haut-parleurs par côté) montré Figure 2**Erreur ! Source du renvoi introuvable.**. Chaque haut – parleur est séparé de 7,5°. Un rideau acoustiquement transparent cache le cercle à l'auditeur.

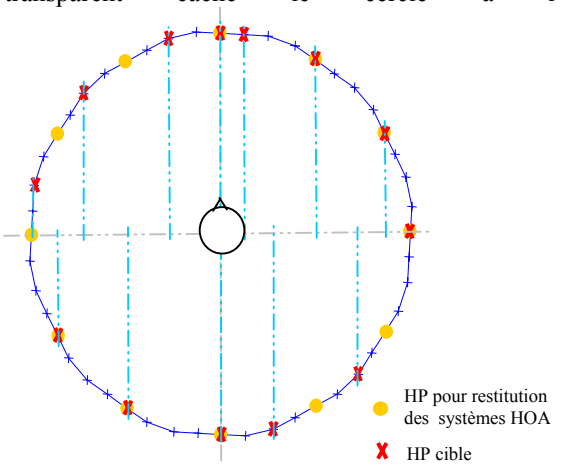


Figure 2. Système de restitution. Les 48 haut-parleurs sont représentés par les croix bleues. Les haut-parleurs "cible" sont représentés par les croix rouge et les haut-parleurs servant à la restitution ambisonique sont représentés par les ronds jaunes.

Les haut-parleurs ont été mesurés à la place de l'auditeur. Leurs réponses ont été déconvoluées aux signaux pour compenser à la fois l'influence des haut-parleurs et la concentricité imparfaite des haut-parleurs.

12 haut-parleurs sont utilisés pour la restitution des systèmes ambisoniques afin de bénéficier d'une configuration régulière.

13 sources cibles sont choisies parmi les 48 haut-parleurs disponibles, privilégiant l'évaluation de la zone d'écoute frontale. Le placement des sources est montré Figure 2. 7 positions de cible sont situées sur des haut-parleurs utilisés pour la spatialisation ambisonique.

Stimuli

Du bruit large bande est utilisé pour couvrir une plage fréquentielle sur laquelle l'ensemble des indices de localisation auditive sont sollicités.

Les réponses des cinq systèmes correspondant aux 360 positions possibles du pointeur sont convoluées avec un bruit uniformément masquant.

Le stimulus cible est un train d'impulsions de bruit blanc modulé en amplitude. Le stimulus cible et le stimulus du pointeur acoustique sont différents de manière à éviter l'appariement perceptif par distance spectrale et non par distance spatiale. Les deux sources sonores sont égalisées en niveau.

La méthode de pointage

Dans la littérature, nous pouvons trouver plusieurs méthodes de report utilisées pour les tests de localisation. Wenzel et Wightman demande aux auditeurs de reporter oralement l'angle perçu de la source [8 9] ce qui nécessite un apprentissage très long. Seeber [11] utilise un pointeur laser qui nécessite l'aide de la vision, et par conséquent une évaluation uniquement frontale des systèmes. Afin d'éviter les biais causés par les méthodes « classiques » de report de localisation, une méthode d'appariement avec pointeur acoustique [3 10] est utilisée. Cependant, le pointeur n'est pas un haut-parleur physique à ajuster à la position de la source virtuelle [10]. Le pointeur est le système ambisonique à tester. Il s'ajuste à la position de la cible située sur un des haut-parleurs physiques du cercle grâce à un bouton rotatif. L'utilisation d'un encodage ambisonique comme pointeur permet d'avoir une résolution d'ajustement proche de l'audition (1°) et non une résolution physique limité au placement des haut-parleurs.

Protocole

Le sujet, placé au centre, déplace une source, spatialisée suivant l'un des modes d'encodage testés, (le pointeur) pour ajuster la direction perçue à celle d'une source réelle (la cible). L'ajustement se fait à l'aide d'un bouton rotatif que le sujet tient dans la main. Les signaux de rotation sont transmis numériquement puis interprétés depuis une interface développée sous Max MSP.

Les fichiers étant prétraités, le pointeur ne peut pas être déplacé pendant la durée du stimulus. La rotation du bouton rotatif est effective au prochain son joué. Le pointeur a donc une durée relativement courte, 150 ms, pour ne pas gêner

l'auditeur par l'impossibilité de bouger le son « en continu ». Les deux signaux (cible et pointeur) sont présentés consécutivement, après un silence de 50 ms. L'auditeur a 25 répétitions de la séquence cible – pointeur pour positionner le pointeur sur la cible. Si l'auditeur ajuste les deux sources avant la fin des 25 répétitions, il peut passer à la position suivante en appuyant sur un bouton déclenchant la séquence suivante. Les positions du son cible sont tirées aléatoirement parmi les 13 positions choisies. La position initiale du pointeur est choisie aléatoirement entre ± 20° et ± 60° autour de la cible. Chaque système est présenté 3 fois par position afin de valider les réponses de l'auditeur.

Le sujet est au milieu du cercle. La tête de l'auditeur n'est pas fixée mais on lui demande de s'orienter vers une marque indiquant le 0°. Il lui est demandé de ne pas détourner la tête en direction de la source. Une séquence d'apprentissage est effectuée présentant à l'auditeur un échantillon des systèmes testés.

Déroulement du test

Les tests se déroulent à France Télécom R&D. A ce jour, 8 auditeurs, 7 hommes et 1 femme, âgés de 23 à 46 ans, ont passé le test.

Analyse

Les premiers résultats sont en cours d'analyse. Leur tendance est conforme aux résultats attendus : le système d'ordre 4 « idéal » est le meilleur en terme de localisation, le système ordre 4 mettant en œuvre le microphone 32 capsules a de bons résultats avec une plus grande variabilité. Les angles perçus sont identiques aux positions de la cible à 5° près pour le système idéal d'ordre 4, pour les systèmes d'ordre 4 et d'ordre 2 (mettant en œuvre le microphone 12 capsules) à 6° près.

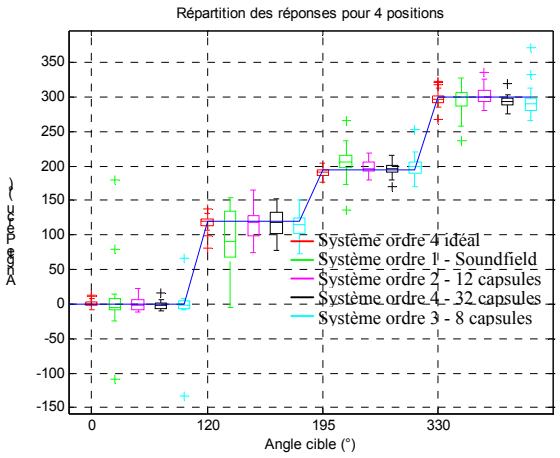


Figure 2. Résultats représentés par la médiane et les interquartiles à 25% et 75% des 5 systèmes respectivement : ordre 4 idéal, microphone SoundField (ordre 1), microphone 12 capsules (ordre 2), microphone 32 capules (ordre 4) et 8 capsules (ordre 3), pour 4 positions de la cible : 0°, 120°, 195° et 330°.

Les résultats peuvent varier jusqu'à ±30° par rapport à la position cible pour le système « idéal », ±40° pour le système d'ordre 4 « 32 capsules » et ±50° pour le système d'ordre 2. Les plus grandes erreurs d'angle ont été perçues avec le système d'ordre 1, mettant en œuvre le microphone SoundField (un angle moyen perçu de 90° pour une position de la cible à 120°), avec une grande variabilité dans les résultats). Les résultats de 4 positions de la cible sont montrés Figure 2. Pour tous les systèmes un flou de localisation est visible pour les sources latérales. Quelques confusions avant arrière sont constatées.

Une analyse plus approfondie sur le temps de réponse des auditeurs, l'influence inter individu reste à finaliser. L'effet d'apprentissage est à évaluer.

Conclusion

Le test met en œuvre le SoundField, les 2 prototypes HOA, d'ordre 2 et 4 et un système « idéal » avec une restitution par 12 haut-parleurs. Une tendance se dessine après les premiers résultats. On tend à observer que les performances de pointage (donc de localisation) s'améliorent avec la résolution d'encodage et lorsque l'encodage est dénué d'artefact (ordre 4 idéal).

On espère qu'une analyse plus complète permettra de corréler ces résultats à une caractérisation objective des systèmes [2].

Références

[1] Study of Higher Order Ambisonic Microphone. Moreau S. et Daniel J., CFA/DAGA, Strasbourg, 2004.

[2] 3D Sound Field Recording with Higher Order Ambisonics – Objective Measurements and validation of a 4th order Spherical Microphone. Moreau S., Daniel J. et Bertet S., convention AES, Paris, 2006.

[3] Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions. Daniel, J., Rault, J.B., and Polack, J.D. AES 105th Conv. 1998.

[4] Spatial Hearing. Blauert J., 1999.

[5] An introduction to the psychology of hearing, 4th edition. Moore B., 1997.

[6] Two-dimensional sound localization by human listeners. Makous J. J. Acoust. Soc. Am. 1990, 2188-2200.

[7] Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. Daniel, J. Paris 6, 2001.

[8] Headphone simulation of free field listening II : Psycophysical validation. Whigthman. F. J. Acoust. Soc. Am. 1989, 868-878.

[9] Localization using nonindividualized head-related transfer functions. Wenzel E. J. Acoust. Soc. Am. 1993, 111-123.

[10] Localization of virtual Sources in Multichannel Audio Reproduction. Pulkki V. IEEE. 2005. 105-119.

[11] A new method for localization studies. Seeber B. Acta Acustica. 1997.