CFA/VISHNO 2016

Séparation de sources appliquée à un contenu ambisonique : localisation et extraction des champs directs

M. Baqué[a], A. Guérin[a] et M. Melon[b]
[a]Orange Labs, 4 rue du Clos Courtel, BP 91226, 35512 Cesson-Cévigné Cedex, France
[b]LAUM - Université du Maine, Avenue Olivier Messiaen, 72085 Le Mans Cedex 9, France
mathieu.baque@orange.com

Dans un précédent article, il a été montré que l'algorithme d'Analyse en Composantes Indépendantes (ACI) appelé *Entropy Rate Bound Minimization* (ERBM) permettait de localiser et d'extraire de manière aveugle des sources sonores dans des contenus multicapteurs ambisoniques simulés numériquement. Il a été également mis en évidence que cette décomposition du champ sonore restait efficiente en présence de premières réflexions et de réverbération tardive simulées par lancer de rayons. Dans cet article, l'étude est prolongée sur des signaux réels enregistrés avec un microphone ambisonique du marché dans des conditions acoustiques variables, de faiblement à moyennement réverbérantes. L'application d'ERBM à ces signaux permet une localisation précise des sources présentes dans le mélange, quels que soient le temps de réverbération et l'énergie des premières réflexions. Sur la base de cette localisation, une étape de formation de voie sous contrainte permet de réaliser une séparation des champs directs associés à chaque source, avec un score signal-sur-interférence compris entre 10 et 25 dB.

1 Introduction

L'audio 3D est actuellement en pleine expansion : formats de diffusion avec Auro3D ou Dolby ATMOS, codeurs nouvelle génération avec l'AC4 [1] ou MPEGH-3D [2, 3] permettant de compresser et transporter ces nouveaux formats. Et au centre, la question de la création des contenus devient prégnante : notamment, comment capter et mixer des scènes sonores immersives réalistes ? Le format multicanal ambisonique et son extension aux ordres élevés HOA [4] (pour *Higher Order Ambisonics*), présente de multiples avantages : il fournit une représentation spatiale du champ sonore indépendante du système de restitution et permet des transformations simples de la scène sonore comme la rotation ou la focalisation. La captation de scène audio 3D est facilitée par la disponibilité sur le marché d'un nombre croissant de microphones ambisoniques -des sphères de capteurs avec une résolution spatiale dépendant du nombre de capsules- ouvrant la possibilité d'enregistrer nativement dans ce format.

En contre-partie au caractère générique du format HOA, la manipulation avancée comme le déplacement, l'atténuation, voire la suppression d'une source particulière n'est pas immédiate, ce qui limite les possibilités de l'ingénieur du son. La réalisation de ce type de traitements nécessite une décomposition de la scène sonore en une somme d'objets, appelés aussi sources sonores. L'Analyse en Composantes Indépendantes (ACI) apparaît alors comme une approche pertinente pour réaliser l'analyse de scène et l'extraction des sources.

La première partie de cet papier expose le formalisme mathématique associé à la séparation de sources par ACI. Les caractéristiques du système de captation utilisé et les conditions acoustiques des différentes prises de son sont ensuite détaillées. Enfin, les résultats obtenus en terme de localisation des sources et de rapport signal-sur-interférence sont discutés dans la dernière partie.

2 Analyse en composantes indépendantes

Le cas théorique d'un enregistrement HOA de N sources sonores en conditions anéchoïques est représenté par l'équation suivante :

$$\mathbf{x}^d(t) = \mathbf{Y}^d \cdot \mathbf{s}(t) = \sum_{i=1}^{N} \mathbf{Y}_i^d \cdot s_i(t), \tag{1}$$

où $\mathbf{x}^d(t)$ est le vecteur colonne des observations du champ direct, $\mathbf{s}(t)$ le vecteur colonne des N sources sonores et $\mathbf{Y}^d$ la matrice d'encodage HOA. Pour un microphone idéal et des sources suffisamment lointaines, chaque colonne $\mathbf{Y}_i^d$ de la matrice $\mathbf{Y}^d$, qui est alors une matrice de gain, contient les coefficients d'encodage HOA onde plane de la source i :

$$\mathbf{Y}_i^d = \begin{bmatrix} Y_{00}^1\,(W) \\ Y_{11}^1\,(X) \\ Y_{11}^{-1}\,(Y) \\ Y_{10}^1\,(Z) \\ Y_{22}^1\,(U) \\ \dots \end{bmatrix}_i = \begin{bmatrix} 1 \\ \sqrt{3}\,\cos\theta_i\,\cos\phi_i \\ \sqrt{3}\,\sin\theta_i\,\cos\phi_i \\ \sqrt{3}\,\sin\phi_i \\ \sqrt{5}\,\frac{\sqrt{3}}{2}\,\cos 2\theta_i\,\cos^2\phi_i \\ \dots \end{bmatrix}, \tag{2}$$

où le couple (θ_i, ϕ_i) représente la position en coordonnées sphériques de la source i relativement au microphone et Y_{mn}^{σ} est la fonction harmonique sphérique d'ordre m et de degré σn.

Sous l'hypothèse d'indépendance des sources, et dans le cas où le nombre d'observations est supérieur au nombre de sources, la matrice $\mathbf{B}^d$, définie comme la pseudo-inverse de $\mathbf{Y}^d$, peut être identifiée par des méthodes d'ACI.

L'extraction des sources estimées $\hat{\mathbf{s}}(t)$ est alors effectuée suivant l'Eq. 3 :

$$\hat{\mathbf{s}}(t) = \mathbf{B}^d \cdot \mathbf{x}(t) \tag{3}$$

Dans l'article [5], il a été montré que des approches classiques telles que JADE [6], EFICA [7] ou encore des algorithmes plus récents comme ERBM [8], peuvent résoudre ce problème de séparation de sources en identifiant correctement $\mathbf{B}^d$.

Un cas plus réaliste peut être modélisé en considérant les signaux enregistrés $\mathbf{x}$ comme la somme des contributions du champ direct $\mathbf{x}^d$ donné par l'équation 1, du champ réverbéré $\mathbf{x}^r$ composé de premières réflexions hautement corrélées avec le champ direct et d'un réverbération tardive décorrélée, représenté par l'équation suivante :

$$\mathbf{x}(t) = \mathbf{x}^d(t) + \mathbf{x}^r(t) \tag{4}$$

Dans [5], il a été montré que, sur des signaux contenant un champ réverbéré simulé par lancer de rayons, les algorithmes classiques comme JADE ou EFICA échouent à identifier $\mathbf{B}^d$, principalement à cause des 1$^{\text{ères}}$ réflexions corrélées avec leurs champs directs respectifs. Ces réflexions peuvent être considérées comme des sources secondaires, l'hypothèse d'indépendance des sources n'est donc plus respectée.

L'approche classique pour effectuer la séparation complète est de considérer le mélange convolutif et de travailler en bandes de fréquences, ce qui permet d'identifier une matrice de mélange (complexe) propre à

chaque sous-bande. Cependant, si cette approche permet théoriquement de traiter des mélanges réverbérants, elle souffre généralement d'artefacts audibles lors de la reconstruction des signaux pleine-bande. De plus, lorsque l'on vise des applications comme la réorganisation de la scène sonore (ou l'égalisation des niveaux des différentes sources), il n'est pas nécessaire d'isoler la contribution totale de chaque source : seul le premier front d'onde est nécessaire. En effet, l'effet de précédence nous indique que la localisation d'une source sonore par l'auditeur dépend principalement de la direction d'arrivée du premier front d'onde. Aussi, l'extraction du champs direct de chaque source est une condition suffisante pour pouvoir effectuer une manipulation spatiale des objets qui soit cohérente du point de vue de l'auditeur [9].

Aussi, la décomposition en objets pour ce type d'application peut se restreindre à tenter d'estimer uniquement la matrice de mélange $\mathbf{B}^d$, sans avoir à procéder à une déconvolution.

Dans [5], il a été montré que, sur des effets de salle simulés, l'algorithme ERBM était toujours capable d'identifier la matrice $\mathbf{B}^d$. La seule contrainte est que le retard de la première réflexion par rapport au champ direct soit supérieur à 3-4 millisecondes, ce qui n'est pas problématique en pratique tant que le microphone n'est pas placé trop proche d'une surface réfléchissante. On se propose ici d'étendre l'étude au cas réel de signaux enregistrés avec un microphone ambisonique, l'Eigenmike^{TM}, ne présentant donc pas un encodage parfait et de mesurer l'impact de l'erreur d'encodage d'une part sur la localisation des sources, et d'autre part sur la séparation des sources. Le paragraphe suivant s'attache à caractériser le microphone, notamment sa déviation dans les basses fréquences par rapport à un encodage parfait.

3 Limites d'un microphone ambisonique réel

3.1 Imperfections microphoniques

Le microphone Eigenmike (Fig. 1) utilisé ici est composé d'un sphère rigide de 8.4 cm de diamètre sur laquelle sont disposées 32 capsules omnidirectionnelles régulièrement espacées, à partir desquelles peuvent être générées 25 fonctions de directivité (ambisonie à l'ordre 4).

Figure 1 – Microphone Eigenmike

Quatre d'entre-elles sont représentées en fonction de la fréquence sur la Fig. 2 : La composante $n°1$ (ordre 0 omnidirectionnel), les composantes $n°2$ et $n°3$ (deux premières fonctions harmoniques sphériques d'ordre 1), et la composante $n°5$ (première harmonique sphérique d'ordre 2).

La première limitation, liée au repliement spatial, impacte les hautes fréquences, et plus précisément les longueurs d'onde inférieures à deux fois la distance inter-capteurs. Dans le cas de l'Eigenmike, cela fixe la limite haute de la bande de validité de l'encodage ambisonique du microphone à environ 8 kHz. Cette limitation est mise en évidence par les directivités à 10 kHz des différents ordres, qui s'éloignent sensiblement des profils théoriques donnés par les courbes à 1 kHz.

L'autre limitation, qui dépend de l'ordre, impacte l'estimation des harmoniques sphériques dans les basses fréquences. L'estimation des harmoniques nécessite de calculer des gradients entre des capsules et/ou des directivités, gradients qui sont très faibles pour des longueurs d'onde grandes devant la taille du réseau de capsules. Il en résulte une estimation particulièrement bruitée qui se dégrade avec l'ordre comme illustré sur les courbes à 200 Hz de la Fig. 2. Sur les 3 premiers cadrans, on observe une légère déviation des directivités d'ordre 0 et 1 à 200 Hz par rapport à la directivité idéale représentée ici par celle à 1 kHz : cercle pour l'ordre 0 (canal omnidirectionnel), figure de 8 pour l'ordre 1. Quant à l'ordre 2 (cadran bas-droit de la Fig. 2), la directivité à 200 Hz ne correspond plus à la forme idéale en trèfle à 4 feuilles.

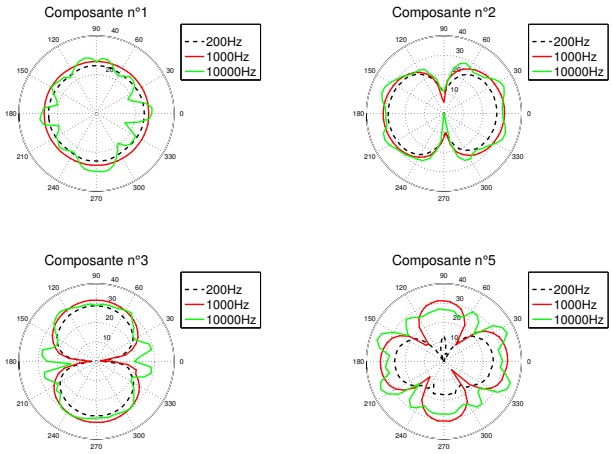

Figure 2 – Directivités des composantes ambisoniques synthétisées dans le plan azimutal pour différents ordres et différentes fréquences

D'après l'expression trigonométrique des coefficients d'encodage (Eq.2) et l'expression générale des fonctions harmoniques sphériques d'ordres supérieurs, on établit la relation suivante que l'on nomme critère d'encodage onde plane :

$$c_{op} = \frac{(2m'+1)\cdot\sum_{n\sigma}(Y^{\sigma}_{mn,i})^2}{(2m+1)\cdot\sum_{n\sigma}(Y^{\sigma}_{m'n,i})^2} \qquad (5)$$

qui est supposée être égale à 1 pour tous les ordres m et m'. L'éloignement de c_{op} de sa valeur théorique permet de quantifier la déviation de l'encodage microphonique par rapport à un encodage parfait. Un indicateur similaire est également utilisé dans [10].

En prenant comme référence le coefficient d'encodage omnidirectionnel (ordre 0), on observe sur la Fig. 3 que la limite basse de validité de l'encodage microphonique est bien fonction de l'ordre : environ 200 Hz pour l'ordre 1, 800 Hz pour les ordres 2 et 3, 1500 Hz pour l'ordre 4 tandis que la limite haute se situe à environ 8 kHz pour tous les ordres.

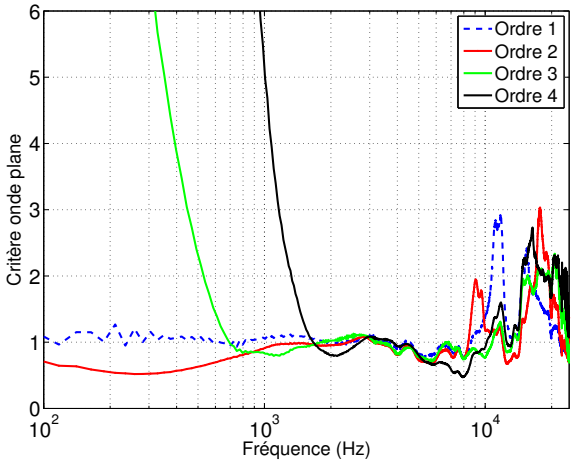

FIGURE 3 – Critère onde plane c_{op} de l'Eigenmike en fonction de la fréquence et de l'ordre. Source à la position angulaire (0°,0°).

3.2 Conséquence sur la décomposition en objets

La décomposition ou séparation qui est formalisée par l'équation 3 peut être interprétée comme une opération de *beamforming* : la $i^{ème}$ ligne permet d'extraire la $i^{ème}$ source par une formation de voie sous-contrainte consistant à placer des zéros dans la direction des sources interférentes tout en conservant un gain unité dans la direction de la source d'intérêt. Ainsi, l'exploitation des ordres supérieurs fournis par l'Eigenmike pour faire une décomposition par formation de voie apparaît au premier abord intéressante car cela permet i) de former des lobes plus fins qu'à l'ordre 1 et donc d'avoir une meilleure résolution spatiale, ii) de fiabiliser l'analyse et de séparer plus de sources grâce à un plus grand nombre d'observations. Cependant, les Fig. 2 et 3 montrent que l'encodage des ordres supérieurs ou égal à 2 est fluctuant en dessous de 1 kHz et nécessite une approche en sous-bandes pour séparer des sources dans cette zone de fréquence. Aussi, on se limitera ici à utiliser l'ordre 1 (4 canaux ambisoniques) qui présente des directivités quasi-constantes dans la bande [200 Hz-8 kHz]. Dans le paragraphe suivant, une description des environnements acoustiques réverbérants dans lesquels ont été captés les contenus est effectuée.

4 Base de données

La base de données de signaux ambisoniques a été générée par convolution de sources de voix mono avec des SRIR (*Spatial Room Impulse Response*) mesurées à l'aide du microphone Eigenmike. Celui-ci a été disposé au centre d'une salle, avec ou sans traitement acoustique comme présenté sur les Fig. 4 et 5. Ces salles possèdent des TR_{60} de respectivement 120 et 360 ms. La salle réverbérante, en dépit de son TR_{60} relativement faible dû à son faible volume ($\sim$ 60 m^3), présente des premières réflexions très énergétiques dues aux surfaces vitrées. A l'aide de haut-parleurs disposés dans la salle et préalablement égalisés, 8 jeux de SRIR dans le plan horizontal ont été mesurées aux angles $[0°, \pm 45°, \pm 90°, \pm 135°, 180°]$ pour la création des contenus ambisoniques.

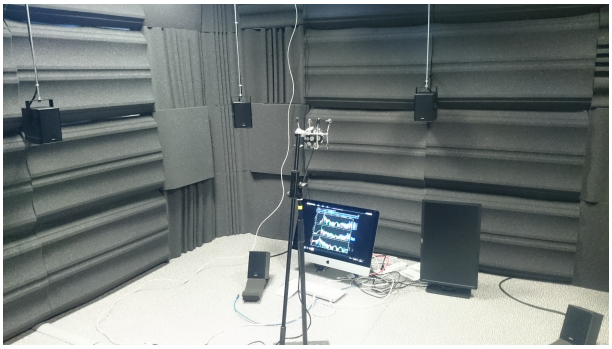

FIGURE 4 – Salle mate - $TR_{60} = 120\ ms$

FIGURE 5 – Salle réverbérante - $TR_{60} = 360\ ms$

Les signaux de parole utilisés pour les simulations ont été normalisés et échantillonnés à 16 kHz, les directivités de l'Eigenmike n'étant pas valides au-dessus de 8 kHz. Chaque source est convoluée par le jeu de SRIR correspondant à la direction d'arrivée souhaitée, puis les contributions de toutes les sources sont additionnées pour générer la scène ambisonique complète. Dans cet article, nous considérons le cas d'un mélange de 3 sources de parole dont les directions d'arrivée sont données dans le Tableau 1.

TABLEAU 1 – Directions d'arrivée des sources de parole

	azimuth (°)	élévation (°)
source 1	0	0
source 2	135	0
source 3	-45	0

Le paragraphe ci-après présente les performances de la décomposition en objets du champ sonore capté à l'aide d'ERBM.

5 Décomposition en objets des contenus ambisoniques

Comme précisé dans la section 2, on va donc s'intéresser ici à l'estimation de la matrice de gains $\mathbf{B}^d$. Dans le cas d'un microphone réel, la matrice $\mathbf{Y}^d$ est une matrice de filtres, néanmoins dans la bande de fréquences de validité de l'encodage microphonique on peut raisonnablement l'approcher par une matrice de gains.

5.1 L'algorithme ERBM

L'algorithme ERBM est dérivé d'EBM (*Entropy Bound Minimization*) [11], méthode d'ACI qui recherche la matrice de séparation $\hat{\mathbf{B}}$ minimisant l'entropie des sources estimées $\hat{s}_i(t)$ normalisées. Le calcul de l'entropie nécessite la connaissance de la densité de probabilité des signaux à extraire, densité non connue a priori. Aussi, EBM se base sur un ensemble de couples de fonctionnelles dérivables (V_k, G_k) pour estimer une entropie maximale $H_k(\hat{s}_i(t))$ adaptée à différentes distributions : paire/impaire, unimodale/bi-modale, etc. Pour chaque couple de fonctions, l'entropie maximale est donnée par :

$$H_k(\hat{s}_i) = 0.5\log(2\pi e) - V_k(E[G_k(\hat{s}_i)]), \tag{6}$$

où $E[.]$ est l'opérateur espérance, $0.5\log(2\pi e)$ l'entropie d'un signal gaussien de variance unité et les fonctions G_k sont un ensemble d'estimateurs statistiques des signaux (kurtosis, variance, etc). La matrice $\hat{\mathbf{B}}$ est estimée de manière itérative par l'algorithme du gradient, en choisissant le couple (V_k, G_k) fournissant la borne d'entropie la plus faible.

ERBM, pour *Entropy Rate Bound Minimization*, est une amélioration d'EBM proposée par Li et Adali pour des sources colorées [8]. Les sources sont considérées comme des processus auto-régressifs (AR) excités par des bruits indépendants et identiquement distribués. Ce modèle, bien que grossier, notamment en ce qui concerne l'excitation, est un bon prédicteur pour les signaux vocaux et musicaux. Dans ERBM, les paramètres AR de chaque source sont estimés à chaque itération du gradient et utilisés pour blanchir les signaux de sortie. EBM est alors appliqué aux signaux blanchis, la matrice de séparation et les paramètres AR sont ensuite optimisés alternativement jusqu'à une convergence de l'algorithme. Il a été observé dans [5] que le blanchiment opéré par ERBM permettait de rendre le champ direct plus indépendant de ses réflexions, et fiabilisait, en présence d'un champ réverbéré simulé, l'estimation de la matrice $\mathbf{B}^d$, en s'affranchissant de l'influence des premières réflexions.

5.2 Implémentation

Dans le but d'une utilisation temps-réel, l'algorithme ERBM est implémenté trame-à-trame, en décomposant le champ ambisonique en trames de 500 ms, avec un recouvrement de 50%, et en estimant une matrice de séparation $\hat{\mathbf{B}}_k$ pour chaque trame k. Pour résoudre les ambiguïtés de signe, d'amplitude et de permutation inhérentes aux méthodes de séparation aveugle de sources, on utilise les informations dérivant de l'encodage ambisonique.

En pratique, on utilise la matrice de mixage estimée $\hat{\mathbf{Y}}_k^d = \hat{\mathbf{B}}_k^{-1}$, où chaque colonne contient l'expression des coefficients d'encodage d'une source estimée. Les ambiguïtés de signe et d'amplitude sont résolues en normalisant chaque colonne par son premier terme qui est le gain estimé relatif au canal omnidirectionnel (W) supposé égal à 1. Les signaux sont extraits à partir de la pseudo-inverse de cette matrice de mélange normalisée.

L'ambiguïté de permutation des signaux extraits de la trame k est résolue en recherchant les couples (i, j) minimisant la distance L_2 entre les vecteurs $\hat{\mathbf{Y}}_{i,k}^d$ et les colonnes de la matrice d'encodage $\mathbf{Y}_j^d$. Les signaux de la trame k sont alors permutés en conséquence.

La reconstruction des sources estimées complètes à partir des sources estimées par trame est réalisée par la méthode d'*Overlap-and-Add*, en utilisant des fenêtres d'apodisation dépendant du taux de recouvrement des trames et du fenêtrage éventuel des signaux avant analyse.

5.3 Critères d'évaluation

On s'intéresse ici à la séparation des champs directs des différentes sources. Pour évaluer cette séparation, on mesure pour chaque trame un rapport signal-sur-interférence (SIR), ratio entre le champ direct de la source d'intérêt extraite et les champs directs résiduels des sources interférentes.

Pour chaque couple de sources (i, j), on définit $\mathrm{SIR}_{ij}^{\mathrm{input}}$ et $\mathrm{SIR}_{ij}^{\mathrm{output}}$ qui représentent les rapports signal sur interférence en entrée et en sortie de l'algorithme de décomposition en objets :

$$\mathrm{SIR}_{ij,k}^{input} = 10\log_{10}\left(\frac{\mathrm{E}\left[s_i^d(t)^2\right]_{t\in trame\ k}}{\mathrm{E}\left[s_j^d(t)^2\right]_{t\in trame\ k}}\right) \tag{7}$$

$$\mathrm{SIR}_{ij,k}^{output} = 10\log_{10}\left(\frac{\mathrm{E}\left[\hat{s}_i^d(t)^2\right]_{t\in trame\ k}}{\mathrm{E}\left[\hat{s}_{j\ in\ i}^d(t)^2\right]_{t\in trame\ k}}\right) \tag{8}$$

où E[.] est l'opérateur espérance, k le numéro de la trame, $s_i^d(t)$ le champ direct de la $i^{ème}$ source, $\hat{s}_i^d(t)$ le champ direct de la source i dans le signal extrait $\hat{s}_i(t)$ et $\hat{s}_{j\ in\ i}^d(t)$ le résidu du champ direct de la source j dans la source extraite $\hat{s}_i(t)$, i.e. après focalisation dans la direction de la source i.

La comparaison de $\mathrm{SIR}^{\mathrm{output}}$ et $\mathrm{SIR}^{\mathrm{input}}$ est un indicateur de la performance de la séparation.

La localisation des sources estimées est également un critère d'évaluation pertinent et directement lié à la matrice de séparation estimée par l'étape d'ACI. Grâce au lien direct entre les coefficients d'encodage ambisoniques et la direction d'arrivée des sources (Eq. 2), on déduit à chaque trame k la direction d'arrivée estimée de chaque source en utilisant les coefficients d'encodage de l'ordre 1 données par la matrice $\hat{\mathbf{Y}}_k^d$ (*pour alléger la notation, l'indice de trame k est volontairement omis*) :

$$\begin{cases} \hat{\theta}_i = \tan^{-1}\left(\frac{\hat{Y}_{11,i}^{-1}}{\hat{Y}_{11,i}^{1}}\right) \\ \hat{\phi}_i = \tan^{-1}\left(\frac{\hat{Y}_{10,i}^{1}}{\sqrt{(\hat{Y}_{11,i}^{1})^2+(\hat{Y}_{11,i}^{-1})^2}}\right) \end{cases} \tag{9}$$

5.4 Performances

Pour un enregistrement effectué dans la salle mâte (Fig. 4), La Fig. 6 donne une représentation en diagramme en boîte des SIR en relatifs à l'extraction de la source 1, c'est-à-dire pour les couples de sources $(1, 2)$ et $(1, 3)$. Les limites des boîtes sont déterminées par les 1^{er}, $2^{ème}$ (médiane) et $3^{ème}$ quartiles, les moustaches délimitant l'intervalle contenant 99.3% des données (hypothèse d'une distribution gaussienne).

Les rapports d'énergie dans le contenu initial $\mathrm{SIR}_{1j}^{\mathrm{input}}$ (vert) montrent une distribution autour de la valeur médiane 0 dB, ce qui est logique car les sources de départ sont normalisées. La variabilité est due à la nature non-stationnaire et intermittente des signaux de parole et à la taille relativement faible de la trame d'analyse (500 ms).

Les scores en sortie SIR_{1j}^{output} montrent une atténuation des champs directs des sources interférentes, entre 20 et 25 dB en moyenne. Concernant l'atténuation du champ direct de la source 3, l'amélioration est moindre et la variance du SIR_{13}^{output} est plus importante. Cela s'explique par le fait que l'erreur de localisation en azimuth de la source 3 est plus importante que pour les autres sources (cf. Tab. 2) : le biais est de 5°, à comparer avec les 1.5° ou 0.7° des sources 1 et 2.

Dans le cas de la salle réverbérante (Fig. 5), on retrouve les mêmes tendances (Fig. 7) : le résidu associé au champ direct de la source 2 est toujours très faible (atténuation de 20 dB en moyenne), tandis que celui de la source 3 n'est plus atténué que d'une dizaine de dB. Les mesures d'erreur de localisation du Tab. 3 permettent d'expliquer partiellement ces différences de performances : globalement, les erreurs de localisation augmentent pour toutes les sources, avec un maximum de 20° pour la source 3. On remarque également que la source 2 reste toujours relativement bien localisée (biais de 2° en azimuth) par rapport aux deux autres sources : cela s'explique par la configuration de la scène et la résolution angulaire de l'ordre 1 ambisonique. La source 2 est espacée des autres sources de 135° et 90°, tandis que les sources 1 et 3 ne sont espacées que de 45°. Or, si l'on observe la Fig. 8 qui illustre le type de directivité que l'on peut obtenir avec l'ordre 1, on voit qu'il est relativement difficile de discriminer les sources 1 et 3. Au final, les performances d'ERBM sont altérées par la présence de l'effet de salle -1ères réflexions et réverbération-, associée à une variation très faible de l'entropie en fonction de la direction d'arrivée, du fait de la faible directivité de l'ordre 1. la Fig. 8 montre également que, du fait de la largeur du lobe principal, ce biais dans la localisation se traduit en une chute de performances importante : la source 3, localisée à environ 65° au lieu de 45°, n'est plus atténuée que de 7 à 8 dB (la source 2, correctement localisée, est atténuée de 30 dB).

De premières écoutes informelles ont cependant suggéré que l'atténuation des champs directs interférents, même dans le cas le moins favorable, était suffisant pour déplacer une source sans modifier la perception de la direction d'arrivée des autres sources.

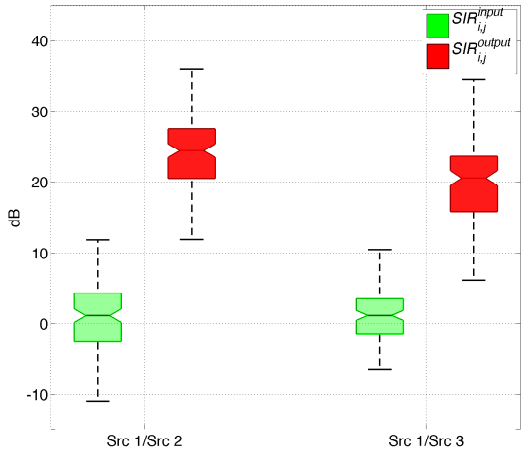

FIGURE 6 – SIR (dB) pour la salle mate (TR de 120 ms)

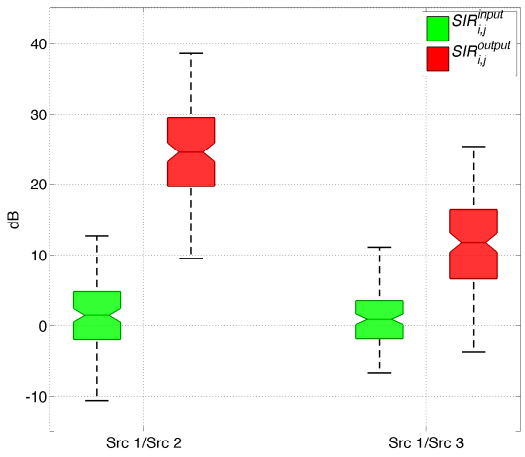

FIGURE 7 – SIR pour la salle réverbérante (TR de 360 ms)

TABLEAU 2 – Erreur de localisation - salle mâte

Erreur de localisation	azimuth (°)		élévation (°)	
	biais	écart-type	biais	écart-type
s_1	1.5	2.1	-0.9	1.8
s_2	0.7	2.2	-2.3	9.6
s_3	-4.9	8.5	-0.7	12.5

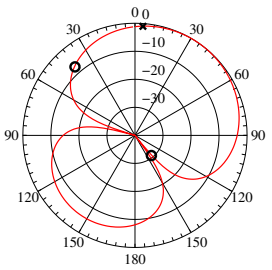

FIGURE 8 – Exemple de beamforming. × : DOA de la source extraite, ○ : DOA des sources interférentes

6 Conclusion

Dans cet article, on s'est intéressé à la localisation et à la séparation de sources captées en conditions réelles par un microphone ambisonique à l'ordre 1. Les résultats montrent que l'algorithme ERBM permet d'identifier les directions d'arrivée des sources dans différents environnements plus ou moins réverbérants, avec un biais à la localisation qui augmente avec le taux de réverbération. En tout état de cause, même si la séparation des champs directs des sources s'en trouve dégradée, des écoutes informelles montre que cela reste suffisant pour permettre le déplacement d'une source par exemple. Cela ouvre la voie à la manipulation de contenus ambisoniques. Toutefois, ces résultats nécessiteront d'être confirmés part des tests perceptifs, tests qui pourrait en outre permettre de fixer les limites d'utilisation de la méthode proposée dans cet article en fonction de l'application visée (rotation de source ou rehaussement de niveau).

Tableau 3 – Erreur de localisation - salle réverbérante

Erreur de localisation	azimuth (°)		élévation (°)	
	biais	écart-type	biais	écart-type
s_1	8.9	8.4	-0.9	15.1
s_2	-1.9	6.4	-1.5	14.4
s_3	-19.8	13.3	-0.6	11.6

Références

[1] Dolby. Dolby ac-4. [Online]. Available : http://www.dolby.com/us/en/technologies/AC-4.html

[2] MPEG. Iso/iec 23008-3. [Online]. Available : https://www.iso.org/obp/ui/#iso:std:iso-iec:23008:-3:ed-1:v2:en

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h 3d audio - the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, Aug 2015.

[4] J. Daniel, "Representation de champs acoustiques, application à la transmission et à la reproduction de scenes sonores complexes dans un contexte multimedia," Ph. D. Thesis, University of Paris VI, France, 2000.

[5] M. Baqué, M. Melon, and A. Guérin, "Separation of direct sounds from early reflections using the entropy rate bound minimization algorithm," in *AES 60th Conference on Dereverberation and Reverberation of Audio, Music, and Speech, 2016*, 2016.

[6] J.-F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *International Conference on Acoustics, Speech, and Signal Processing, 1990. ICASSP-90.* IEEE, 1990, pp. 2655–2658.

[7] Z. Koldovsky, P. Tichavsky, and E. Oja, "Efficient variant of algorithm FastICA for independent component analysis attaining the cramer-rao lower bound," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, Sep. 2006.

[8] X.-L. Li and T. Adali, "Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization," in *International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE*. IEEE, 2010, pp. 1934–1937.

[9] A. Brown, G. Stecker, and D. Tollin, "The precedence effect in sound localization," *J. Assoc. Res. Otolaryngol.*, no. 16, pp. 1–28, 2015.

[10] N. Epain, C. T. Jin, and A. van Schaik, "Blind source separation using independent component analysis in the spherical harmonic domain," in *2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, 2010.

[11] X.-L. Li and T. Adali, "Independent Component Analysis by Entropy Bound Minimization," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5151–5164, Oct. 2010.