UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

TELECOM INT

Études de techniques d'extraction de l'information spatiale dans une scène sonore multicanal

Benjamin DUVAL

Rapport de stage de Master 2 SdI/MIS/ATIAM
Effectué dans les laboratoires de France Télécom R&D
15 mars - 29 septembre 2006

Stage encadré par David VIRETTE et Jérôme DANIEL

ircam Centre Pompidou

france telecom

Remerciements

Je tiens à remercier chaleureusement toute l'équipe son 3D de France Télécom R&D, en particulier mes 2 co-encadrants, David et Jérôme, pour leur accueil, leur disponibilité et leur gentillesse, et pour les généreuses et passionnantes explications. Merci également à Manu d'en face, Greg et Alex pour leur confiance, et Rozenn parce qu'elle le vaut bien.

Merci à Stouf et Micheline, super doctoranges et collègues de bureau fabuleux, pour tous les bons moments, en particulier les instants Bobby… Merci à tous les thésards et stagiaires du labo pour leur accueil, en particulier Jean-Claude l'ex-strapontiste.

Merci enfin à tous ceux qui m'ont proposé de découvrir leurs terribles séances de tests d'écoute ; les musiques du corpus de tests MPEG et les bruits blancs de Steph trottent encore dans ma tête…

Table des matières

Présentation de France Télécom R&D 1
Introduction 2
Chapitre 1 : Ambisonics 3
1.1. Principe 3
1.1.1. Du B-format à HOA 3
1.1.2. Représentation de scène sonore avec HOA 5
1.2. Encodage et décodage spatial 7
1.2.1. Encodage spatial 7
1.2.2. Décodage spatial 8
1.3. Critique de la technologie HOA 11
1.3.1. Bilan 11
1.3.2. Limites pratiques 11
Chapitre 2 : Analyser un contenu HOA 13
2.1. Objectifs 13
2.1.1. Restitution de la spatialisation à partir d'un downmix mono 13
2.1.2. Amélioration de la précision spatiale 13
2.2. Principe de l'extraction d'information spatiale avec une ACP 14
2.2.1. Information spatiale 14
2.2.2. Application de l'analyse en composantes principales 15
2.3. Méthodologie 16
2.3.1. Construction de scènes artificielles 16
2.3.2. Exploitation d'enregistrements naturels 18
2.3.3. Visualisations des résultats 18
2.3.4. Différents modes d'écoute 19
Chapitre 3 : Description des analyses 21
3.1. Détermination des directions principales 21
3.1.1. Base des algorithmes 21
3.1.2. Traitement du champ résiduel 22
3.2. Extension à la 3D 23
3.3. Analyse aux ordres supérieurs 23
3.3.1. Description 23
3.3.2. Illustration sur une scène virtuelle 24
3.4. Traitement des données 26
3.4.1. Algorithme des k-moyennes 26
3.4.2. Critères de pertinence 27

3.4.3. Lissages....29
3.5. Exploitation de toutes les composantes de l'ACP....29
Chapitre 4 : Résultats et applications....31
4.1. Illustration et critique des résultats obtenus....31
4.1.1. Avant-propos....31
4.1.2. Scène 2D artificielle à 4 sources fixes....31
4.1.3. Scène 2D artificielle à 4 sources dont 3 fixes et 1 mobile....32
4.1.4. Scène 3D artificielle à 1 source et bruit de fond....33
4.1.5. Scène 2D naturelle à 1 source....34
4.2. Quelques applications....34
4.2.1. Amélioration de la précision spatiale : synthèse de l'ordre 2 à partir de l'ordre 1....34
4.2.2. Restitution depuis un downmix mono : synthèse de l'ordre 1 à partir de W....36
4.3. Comparaison avec une méthode existante....37
4.4. Développement d'une application pour la conférence audio spatialisée....39
4.4.1. Position du problème....39
4.4.2. Solution proposée....40
4.4.3. Bilan....42
Conclusion....43
Annexe A : Développement du champ acoustique en harmoniques sphériques....45
A.1. Équation d'onde en coordonnées sphériques....45
A.2. Solutions de l'équation d'onde en coordonnées sphériques....46
Annexe B : Compléments théoriques sur le décodage HOA....53
B.1. Vecteur vélocité et vecteur énergie....53
B.2. Décodages optimisés....54
Annexe C : Analyse en composantes principales....55
C.1. Espace des données....55
C.2. Diagonalisation....57
C.3. Code Matlab....57
Annexe D : Analyse en sous-bandes....59
D.1. La transformée de Fourier à court terme (TFCT)....59
D.2. Synthèse par addition-recouvrement....60
Bibliographie....63

Table des figures

Figure 1.1 - Schéma global du principe de la technologie HOA 4
Figure 1.2 - Prototype de microphone HOA d'ordre 4 5
Figure 1.3 - Erreur de reconstruction d'une onde plane en fonction de l'ordre M et du produit kr. 6
Figure 1.4 - Représentation dans le plan horizontal dans une zone de diamètre 1m de la reconstruction d'une onde plane pour plusieurs ordres et fréquences. 6
Figure 1.5 - Encodage horizontal 8
Figure 1.6 - Prise de son virtuelle équivalente à un encodage au 1er ordre et un décodage basique sur une disposition en 2D hexagonale. 10
Figure 1.7 - Directivités théoriques et reconstruites dans le plan horizontal. 12
Figure 2.1 - Y en fonction de X, dans le cas d'un bruit blanc enregistré 15
Figure 2.2 - Y en fonction de X, dans le cas d'un son placé artificiellement 15
Figure 2.3 - Signes de W · X et W · Y en fonction du cadran 16
Figure 2.4 - Interface graphique du logiciel Bidule. 17
Figure 2.5 - Représentation de tables des couleurs 19
Figure 2.6 - Représentation de la pertinence 19
Figure 3.1 - Schéma général d'analyse 21
Figure 3.2 - Filtres utilisés, inspirés des filtres ERB 22
Figure 3.3 - Position des sources de la scène de test à 4 sources dont 1 mobile 24
Figure 3.4 - Angles d'une scène virtuelle ayant 1 source mobile et 3 sources fixes 25
Figure 3.5 - Angles de la même scène trouvés à partir d'une ACP sur les ordres 0 et 2. 25
Figure 3.6 - Angles d'une scène réelle constituée d'un bruit parcourant le cercle, trouvés à partir d'une ACP sur les ordres 0 et 1, 0 et 2, et 0 et 4 26
Figure 3.7 - Représentation temps-fréquence des angles 28
Figure 3.8 - Énergie de ce même son.. 28
Figure 3.9 - Norme relative de la 1ère composante 28
Figure 3.10 - Pondération appliquée à l'entrée des k-moyennes 29
Figure 3.11 - Directivités séparant idéalement 4 sources placées à 17, 73, 195 et -58° 30
Figure 4.1 - Angles théoriques et calculés pour une scène constituée de 4 sons fixes 32
Figure 4.2 - Angles théoriques et calculés de la scène à 4 sources dont 1 mobile. 32

Figure 4.3 - Azimut et élévation théoriques et calculés d'un son en 3D artificiel placé dans une ambiance bruyante. 33
Figure 4.4 - Angles théoriques et calculés d'un bruit blanc tournant autour du micro, enregistré en chambre sourde 34
Figure 4.5 - Principe de l'amélioration de la précision spatiale 35
Figure 4.6 - Directivités équivalentes à la séparation de 4 sources à l'ordre 1 35
Figure 4.7 - Principe de la restitution de l'ordre 1 à partir du signal mono W 36
Figure 4.8 - Schéma de principe de la méthode BCC 36
Figure 4.9 - X et Y, dans le cas d'une sinusoïde placée à 15° et 120° 37
Figure 4.10 - Angles trouvés par l'algorithme de [Pulkki and Faller, 2006] 38
Figure 4.11 - Angles trouvés par une ACP pour la même scène. 39
Figure 4.12 - Schéma de principe d'une conférence audio spatialisée 40
Figure 4.13 - Schéma amélioré (débit réduit) 40
Figure 4.14 - Analyse : principe de l'encodage dans le serveur 41
Figure 4.15 - Synthèse : principe du décodage dans le client 41
Figure A.1 - Système de coordonnées sphériques 45
Figure A.2 - Fonctions de Bessel sphériques $j_m(kr)$ pour les ordres 0 à 5 47
Figure A.3 - Représentation des harmoniques sphériques pour les ordres m = 0 à 3. 51
Figure B.1 - Décodages les plus appropriés par sous-bande et en fonction de l'étendue de la zone d'écoute 54
Figure D.1 - Schéma de principe de l'analyse par TFCT 60
Figure D.2 - Schéma de principe de la synthèse par la méthode overlap-add 61

Présentation de France Télécom R&D

Le groupe France Télécom est l'un des principaux opérateurs de télécommunications au monde. Créée en 1988 après le scindement des PTT (Poste, Télégraphe, Téléphone) en deux groupes, France Télécom a longtemps bénéficié du statut d'opérateur historique. Depuis 1997, France Télécom fonctionne en tant que société anonyme.

Le CNET, Centre National d'Étude en Télécommunication a été créé en 1944 dans le but de rétablir un réseau de télécommunication en France. En mars 2000, le CNET change de nom et devient France Télécom Recherche & Développement.

Le site France Télécom Division R&D à Lannion a été inauguré le 28 octobre 1963. L'équipe son 3D y mène des recherches sur la prise et la restitution du son, la spatialisation sonore ainsi que sur les environnements acoustiques virtuels. Elle participe également activement à la normalisation MPEG-4.

De nombreux travaux sont consacrés aux technologies de spatialisation binaurale, transaurale (avec la technique du stéréo-dipôle) et ambisonique. De nombreux brevets et publications découlent de ces recherches, qui trouvent déjà des applications dans la conférence audio spatialisée, et ouvrent des perspectives bien plus vastes.

Introduction

La reproduction sonore semble aujourd'hui être parfaitement maîtrisée. Mais l'est-elle vraiment ? Rien n'est moins sûr. Reproduire un son signifie "reproduire toutes ses propriétés telles qu'elles existent dans une situation d'écoute réelle" (tel que rappelé dans [Nicol, 1999]). Or le son est un phénomène à quatre dimensions : il évolue dans le temps et dans les 3 dimensions de l'espace. Si le problème de la reproduction temporelle peut être considéré comme résolu, la restitution spatiale est, elle, toujours à l'étude.

La restitution d'une scène sonore (c'est-à-dire d'un ensemble d'événements sonores placés à des positions et des instants précis) passe par la reproduction du champ acoustique dans une zone donnée, zone au moins assez grande pour qu'on puisse y placer ses 2 oreilles. L'auditeur pourra alors localiser précisément les sources constituant la scène à la fois en temps et dans l'espace.

L'objectif du stage présenté ici est de retrouver par le calcul ces informations à partir de l'analyse de signaux de représentation du champ sonore 3D, en particulier ceux issus de l'encodage spatial de scènes sonores réalisé par la technologie Higher Order Ambisonics (HOA). Il s'agit de retrouver dans une scène HOA les positions qu'occupaient les sources à chaque instant lors de l'encodage.

Cette analyse permet alors d'entrevoir des applications pour le codage et la transmission, et/ou l'amélioration de contenu audio 3D, la séparation de sources ou bien encore la conférence audio spatialisée. En effet, une telle description de la scène permettrait de caractériser de manière légère une partie de l'information transmise par les signaux HOA, et ouvrirait même l'opportunité d'un démixage.

Nous allons ici présenter la technologie *ambisonics* puis les éléments théoriques et pratiques de l'analyse des signaux HOA. Les éléments mathématiques sur lesquels repose ambisonics, à savoir le développement d'un champ acoustique en harmoniques sphériques, sont développés en Annexe A. La lecture préalable de cette annexe n'est pas indispensable à la compréhension du Chapitre 1 mais est vivement conseillée.

Chapitre 1

Ambisonics

1.1. Principe

1.1.1. Du B-format à HOA

L'approche ambisonique, développée dans les années 1970 par Gerzon [Gerzon, 1973] [Gerzon, 1985], a pour but de reconstruire au voisinage de la tête de l'auditeur le champ sonore et ses caractéristiques de propagation, c'est-à-dire au minimum le champ de pression et son gradient. Cette technique est catégorisée dans les techniques de panoramique d'amplitude ou à microphones coïncidents. Le format orignal, appelé B-format, peut en fait être vu comme une restriction au 1er ordre d'une décomposition du champ en harmoniques sphériques (voir Annexe A) ou cylindriques (dans le cas d'une restriction à 2 dimensions). Le format HOA (*Higher Order Ambisonics*) est une extension de cette décomposition à des ordres plus élevés permettant une reconstruction améliorée du champ (une meilleure approximation) dans une zone de restitution élargie.

Une grande spécificité de l'approche ambisonique est de contenir directement l'information spatiale en codant le champ acoustique en un point de l'espace indépendamment du système de restitution utilisé, contrairement aux systèmes de spatialisations classiques (stéréo, 5.1…). L'encodage directionnel d'une onde plane S d'incidence $\vec{u}$ se traduit par les équations :

$$\begin{cases} W = S \\ X = S\sqrt{2}\cos\theta\cos\delta \\ Y = S\sqrt{2}\sin\theta\cos\delta \\ Z = S\sqrt{2}\sin\delta \end{cases} \tag{1.1}$$

Le facteur de normalisation $\sqrt{2}$ a été introduit à l'origine pour s'assurer que les composantes W, X et Y aient une énergie moyenne égale dans le cas de l'encodage d'un champ diffus horizontal.

Le B-format peut être vu comme une troncature au premier ordre de la décomposition en série de Fourier-Bessel du champ sonore [Daniel, 2000]. Une troncature à un ordre plus élevé (*Higher Order* Ambisonics) apporte alors à la reconstruction une meilleure qualité, tant en terme de précision de l'image que de dimensions du *sweet spot*, zone où le champ original est restitué avec une tolérance donnée. Cette zone, très petite pour l'ordre 1 (il faut veiller à se placer bien au centre du dispositif) est nous le verrons nettement élargie aux ordres supérieurs. On atteint également une discrimination angulaire plus fine. Ces ordres supérieurs étant les termes suivants de la série de Fourier-Bessel, une représentation HOA inclut les ordres inférieurs, constituant donc une représentation hiérarchique (ou *scalable*) du champ sonore.

La chaîne de traitement d'un système ambisonique se décompose en plusieurs étapes (voir Figure 1.1) :

- l'encodage spatial, réalisé de manière artificielle pour spatialiser une scène synthétique, ou bien à partir de signaux naturels issus d'une structure multi-microphonique (voir Figure 1.2)

- une éventuelle transformation des signaux HOA (compression, ou transformations linéaires comme par exemple une rotation)

- le décodage spatial, matriçage des signaux pour une diffusion adaptée au dispositif d'écoute

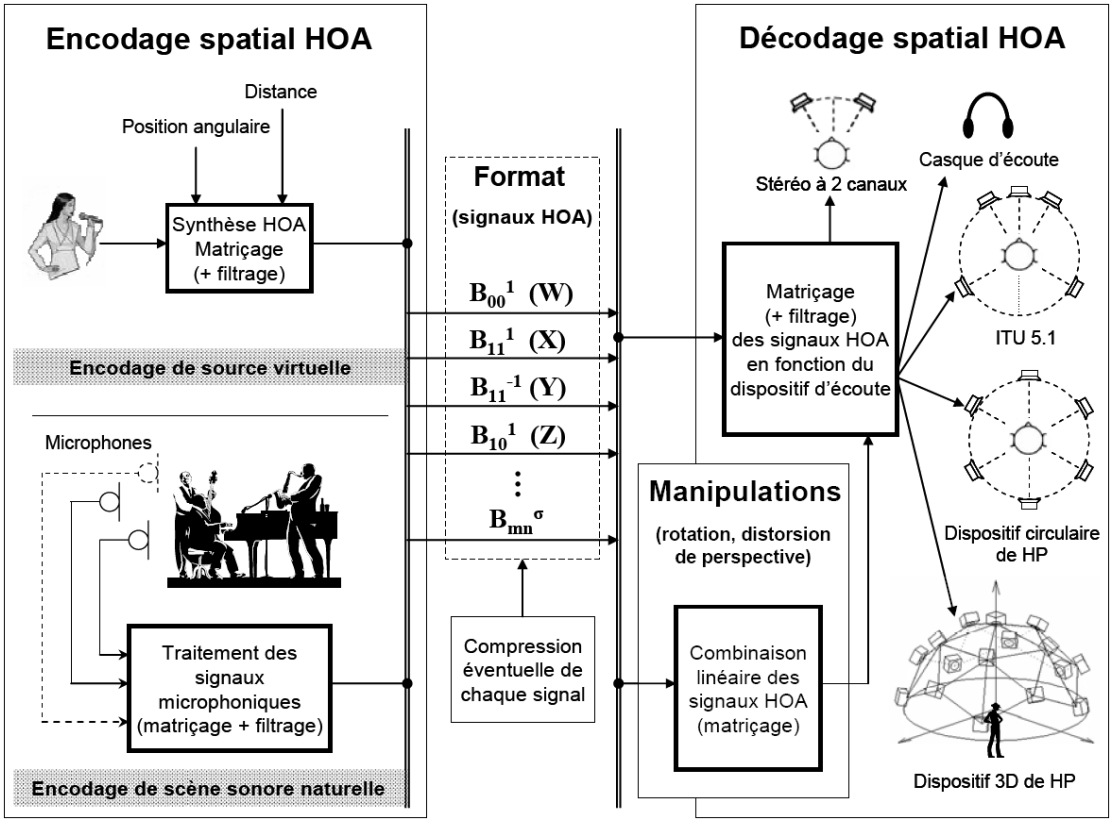

Figure 1.1 - Schéma global du principe de la technologie HOA [Moreau, 2006]

Figure 1.2 - Prototype de microphone HOA d'ordre 4 constitué de 32 capteurs omnidirectifs répartis sur une sphère en plastique rigide. [Moreau, 2006]

1.1.2. Représentation de scène sonore avec HOA

La série de Fourier-Bessel (A.16) en Annexe A) tronquée à l'ordre *M* donne une approximation de la représentation du champ sonore :

$$p_M(kr,\theta,\delta) = \sum_{m=0}^{M} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \tag{1.2}$$

La représentation d'ordre *M* est composée de $K = (M+1)^2$ signaux HOA en 3D, et $K = 2M+1$ signaux HOA si on se contente du plan horizontal.

Pour mesurer l'erreur de reconstruction, on peut utiliser le critère des moindres carrés normalisé (NMSE, pour Normalized Mean Square Error) :

$$e(kr) = \frac{\iint_S \left| p(kr,\theta,\delta) - p_M(kr,\theta,\delta) \right|^2 dS}{\iint_S \left| p(kr,\theta,\delta) \right|^2 dS} \tag{1.3}$$

Dans le cas particulier d'une onde plane, l'erreur de reconstruction est indépendante de l'angle d'incidence, et ne dépend que de l'ordre et du produit *kr*. En rappelant que *k* est le nombre d'onde $k = \omega/c = 2\pi f/c$, on s'aperçoit que l'erreur de reconstruction pour un ordre donné augmente quand on s'éloigne du centre du système (*r* augmente) ou lorsque la fréquence s'élève. On remarque qu'une erreur de 4% (-14 dB) est atteinte approximativement quand *M* = *kr*.

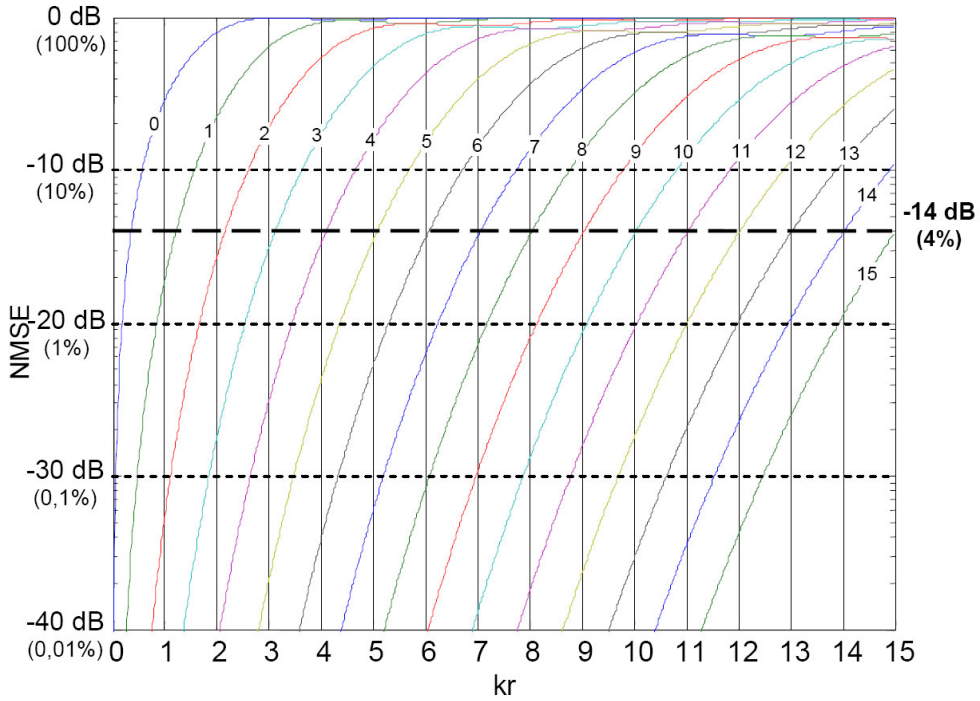

Figure 1.3 - Erreur de reconstruction d'une onde plane en fonction de l'ordre M et du produit kr [Moreau et al., 2006].

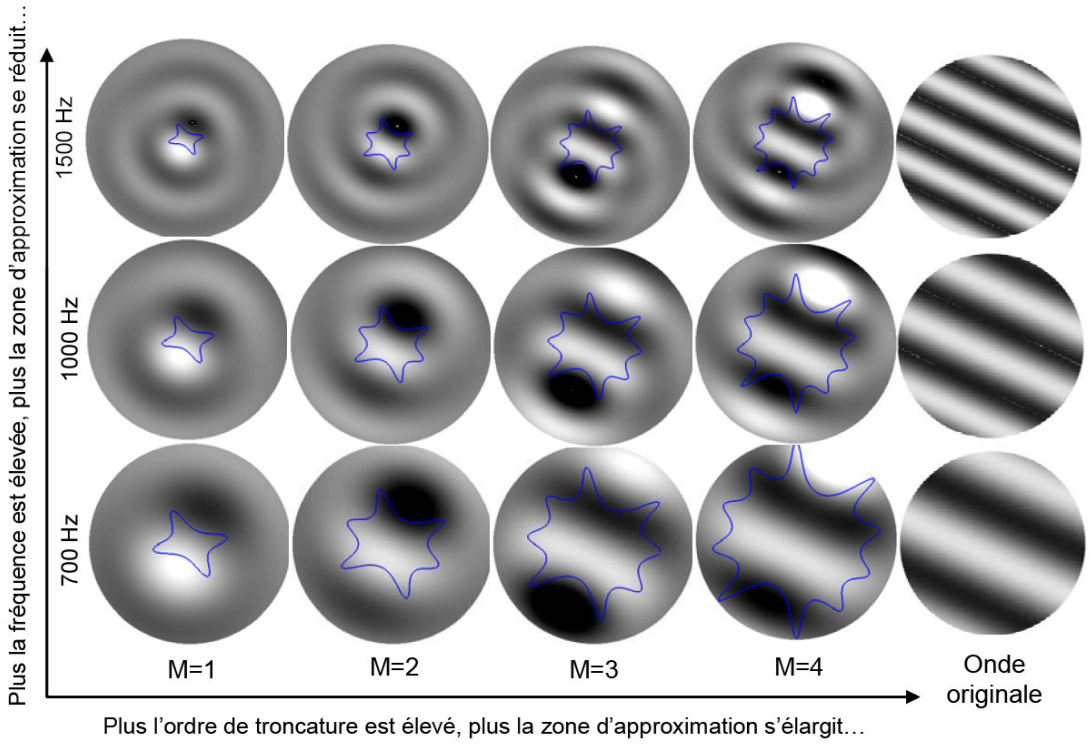

Figure 1.4 - Représentation dans le plan horizontal dans une zone de diamètre 1m de la reconstruction d'une onde plane (à droite) pour plusieurs ordres et fréquences. Les niveaux de gris représentent la valeur de la pression acoustique. Les courbes étoilées indiquent la limite des zones où l'erreur de reconstruction est inférieure à 7% (-11.5 dB) [Moreau, 2006]

1.2. Encodage et décodage spatial

1.2.1. Encodage spatial

D'un point de vue d'ingénieur du son, l'encodage spatial peut être vu comme un panoramique d'amplitude. Un preneur de son l'interprétera comme une capture par un jeu de microphones directifs coïncidents. Un mathématicien verra plutôt le lien avec la décomposition en harmoniques sphériques décrite en Annexe A.

Étudions le cas d'une onde plane provenant de la direction (θ_p, δ_p) avec une amplitude *S*. L'expression du champ de pression s'écrit alors :

$$p(kr, \theta, \delta) = Se^{jkr\cos\gamma} \tag{1.4}$$

soit encore

$$p(kr, \theta, \delta) = S\sum_{m=0}^{\infty}(2m+1)j^m j_m(kr)P_m(\cos\gamma) \tag{1.5}$$

où γ désigne l'angle entre la direction (θ, δ) du point d'observation et la direction (θ_p, δ_p) de provenance de l'onde. Il vient ensuite :

$$p(kr, \theta, \delta) = S\sum_{m=0}^{\infty} j^m j_m(kr)\sum_{n=0}^{m}\sum_{\sigma=\pm1} Y_{mn}^{\sigma}(\theta_p, \delta_p)Y_{mn}^{\sigma}(\theta, \delta) \tag{1.6}$$

En identifiant cette équation (1.6) avec la série de Fourier-Bessel (A.16), nous obtenons finalement l'expression de la transformée de Fourier sphérique d'un champ de pression acoustique engendré par une onde plane [Moreau, 2006] :

$$B_{mn}^{\sigma} = S\,Y_{mn}^{\sigma}(\theta_p, \delta_p) \tag{1.7}$$

Les coefficients B_{mn}^{σ} sont ainsi définis par la valeur des harmoniques sphériques dans la direction de provenance de l'onde plane, pondérée par l'amplitude *S* du signal transporté. Ce sont ces coefficients qui constituent le signal ambisonique. Des coefficients de normalisation α^{N3D} ou de semi-normalisation α^{SN3D} sont ensuite appliqués pour que les signaux soient d'énergie moyenne équivalente entre eux (α^{SN3D}) ou entre les différents ordres (α^{SN3D}) (voir Annexe A, Tableau A.1).

Ces fonctions d'encodage peuvent être simplifiées dans le cas d'une restriction à 2 dimensions (plan horizontal). On ne conserve alors que 2 composantes par ordre, celles qui décrivent le plan horizontal, c'est-à-dire vérifiant $m = n$. Les coefficients de normalisation sont alors différents, et l'angle δ est nul. Les composantes restantes sont alors :

ordre	notations		écriture
0	W	B_{00}^{1}	1
1	X	B_{11}^{1}	$\sqrt{2}\cos\theta$
	Y	B_{11}^{-1}	$\sqrt{2}\sin\theta$
2	U	B_{22}^{1}	$\sqrt{2}\cos(2\theta)$
	V	B_{22}^{-1}	$\sqrt{2}\sin(2\theta)$
m	-	B_{mm}^{1}	$\sqrt{2}\cos(m\theta)$
	-	B_{mm}^{-1}	$\sqrt{2}\sin(m\theta)$

Tableau 1.1 - Fonctions d'encodage ambisonique dans le plan horizontal, avec la convention de normalisation associée (N2D)

La figure suivante illustre l'encodage spatial aux ordres 0, 1 et 2 de 2 sources placées aux angles θ_v et $\theta_{v'}$ dans le plan horizontal :

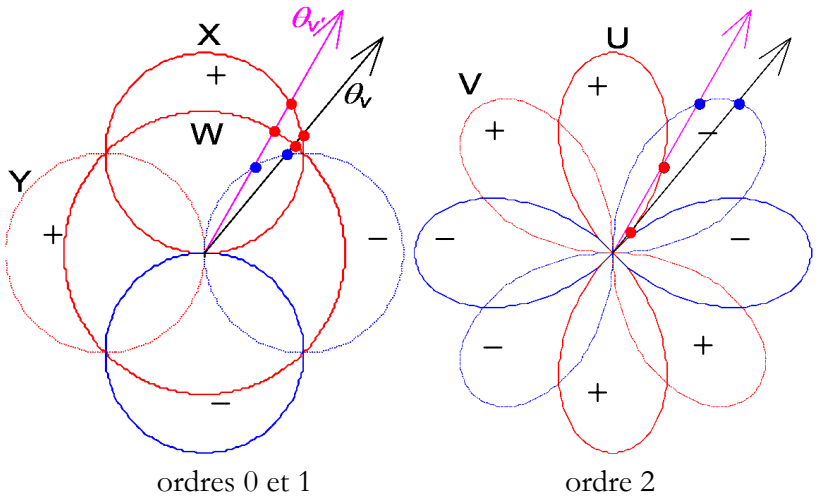

Figure 1.5 - Encodage horizontal [Moreau et al., 2006]. Le coefficient (gain) appliqué à chaque composante de chacune des 2 sources est donné par les intersections des flèches et des diagrammes des harmoniques.

Comme on peut le voir sur cette figure, un encodage au 1er ordre offre une discrimination angulaire peu appuyée, alors que les 2 sources, pourtant proches, auront des coefficients très différents à l'ordre 2 (les points d'intersection sont plus éloignés).

1.2.2. Décodage spatial

L'objectif du décodage est de restituer au mieux sur un système de diffusion donné le champ sonore HOA. Il faut ainsi calculer les signaux S_l à diffuser sur L haut-parleurs en fonction des composantes ambisoniques B_{mn}^{σ}. Ces L haut-parleurs peuvent être disposés dans le plan horizontal ou dans un espace tridimensionnel. Les décodages les plus courants s'effectuent sur une demi-sphère (cas 3D) ou un cercle (cas 2D) de

haut-parleurs. C'est une opération matricielle, qu'on peut formaliser dans le cas d'un décodage dans le plan horizontal par l'équation :

$$C \cdot S = B \tag{1.8}$$

où la matrice *C* contient les vecteurs harmoniques sphériques associés à chaque direction de haut-parleur et organisés suivant ses lignes, le vecteur *S* contient les signaux émis par les *L* haut-parleurs, et enfin le vecteur *B* contient les signaux HOA.

$$C = \begin{bmatrix} Y_{00}^{1}(\theta_1,\delta_1) & \cdots & Y_{00}^{1}(\theta_L,\delta_L) \\ Y_{11}^{1}(\theta_1,\delta_1) & \cdots & Y_{11}^{1}(\theta_L,\delta_L) \\ \vdots & Y_{mn}^{\sigma}(\theta_l,\delta_l) & \vdots \\ Y_{M0}^{1}(\theta_1,\delta_1) & \cdots & Y_{M0}^{1}(\theta_L,\delta_L) \end{bmatrix}, \quad S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_L \end{pmatrix}, \quad B = \begin{pmatrix} B_{00}^{1} \\ \vdots \\ B_{mn}^{\sigma} \\ \vdots \\ B_{M0}^{1} \end{pmatrix} \tag{1.9}$$

(θ_l,δ_l) repérant la position du $l^{ème}$ haut-parleur.

Il faut donc estimer S connaissant B et C. Le décodage se résume alors à un système de $(M+1)^2$ (en 3D) ou $2M+1$ (en 2D) équations linéaires à *L* inconnues. On ne peut trouver de solution exacte que si le nombre de haut-parleurs est au moins égal au nombre de signaux, c'est-à-dire si $L > (M+1)^2$ (en 3D) ou $2M+1$ (en 2D). C'est en outre une condition nécessaire pour pouvoir avoir une reproduction homogène de la scène sonore.

La matrice C n'est pas nécessairement carrée et inversible ; une solution générale consiste à calculer l'inverse généralisée de la matrice C, notée D, encore appelée pseudo-inverse de Moore-Penrose :

$$S = D \cdot B \tag{1.10}$$

avec

$$D = \text{pinv}(C) = C^t \cdot (C \cdot C^t)^{-1} \tag{1.11}$$

D est appelée matrice de décodage. $S = D \cdot B$ résout le système (1.8) si les haut-parleurs sont suffisamment nombreux $L > (M+1)^2$ (en 3D) ou $2M+1$ (en 2D)). S'ils sont en outre disposés sur une sphère (3D) ou un cercle (2D) de manière équidistante, c'est-à-dire s'ils sont les sommets d'un polyèdre ou polygone régulier, ils préservent alors la propriété d'orthonormalité des harmoniques sphériques. Ceci se traduit par :

$$C^t \cdot C = L \cdot I_L \tag{1.12}$$

où I_L est la matrice identité de taille $L \times L$. On a alors :

$$D = \text{pinv}(C) = \frac{1}{L} C^t \tag{1.13}$$

Cette propriété ne peut pas être mathématiquement vérifiée en 3D au-delà de l'ordre 2, autrement dit il n'existe pas de polyèdre régulier dont la position des sommets conserve strictement l'orthonormalité des harmoniques sphériques.

Visuellement, le décodage correspond à la reproduction du champ sonore tel qu'il aurait été capté par un ensemble de N microphones à directivités hyper-cardioïdes tournés chacun vers l'un des haut-parleurs. Leur directivité serait d'autant plus précise que l'ordre de reproduction est élevé.

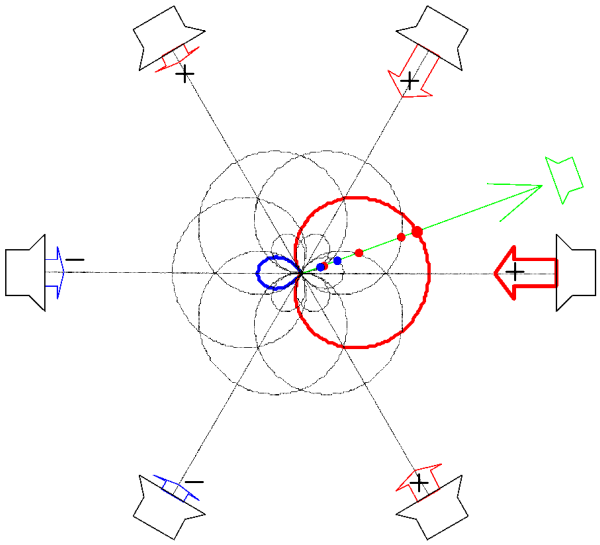

Figure 1.6 - Prise de son virtuelle équivalente à un encodage au 1er ordre et un décodage basique sur une disposition en 2D hexagonale [Moreau et al., 2006].
Le gain g_l appliqué à chaque haut-parleur pour une direction de source virtuelle donnée (en vert) est indiqué par la longueur des flèches et est reportée aux points d'intersection de la flèche verte et des diagrammes de directivité.

On peut remarquer que les diagrammes de directivité sont très larges par rapport à l'espacement angulaire des haut-parleurs. Le 2nd ordre offrirait des directivités plus fines et donc une sélectivité spatiale accrue. Il faudrait alors bien sûr davantage de haut-parleurs pour ne pas introduire des "trous" entre eux. On remarque également la présence non négligeable des lobes secondaires, utiles pour la reconstruction du champ acoustique (notamment en basses fréquences), mais qui peuvent être gênants dans certains cas, par exemple lorsque l'auditoire est large.

Ce décodage est appelé décodage basique car il n'optimise pas la restitution en accord une quelconque hypothèse, et cherche uniquement à reproduire fidèlement le champ acoustique. Pour améliorer la restitution dans des cas critiques, comme par exemple en-dehors du *sweet spot* (centre du dispositif), il faut optimiser le décodage en réduisant par exemple les lobes secondaires. Ainsi, le décodage *max-r_E* s'attache à maximiser l'énergie reçue de la provenance de la direction souhaitée, alors que le décodage *in-phase* annule complètement ces lobes secondaires (voir Annexe B).

1.3. Critique de la technologie HOA

1.3.1. Bilan

On a ainsi vu que la technologie HOA est une représentation intermédiaire du champ de pression acoustique, à mi-chemin entre la description directe du champ lui-même et des signaux à restituer. Cette technologie a de nombreux avantages comparée aux autres méthodes de reproduction de champ sonore (on trouvera notamment une étude détaillée des comparaisons entre HOA et la Wave Field Synthesis (WFS) dans [Daniel et al., 2003]) : elle permet d'éviter le repliement spatial (voir paragraphe suivant), et permet ainsi une reconstruction dans une gamme de fréquences très large.

La qualité de la restitution ambisonique n'est théoriquement limitée que part l'ordre de troncature de la série de Fourier-Bessel, à partir du moment où le nombre de haut-parleurs est suffisant. Ainsi, un ordre tendant vers l'infini (et un nombre de haut-parleurs au moins aussi grand) permettra une reproduction quasi-parfaite sur l'ensemble de la zone d'écoute.

Mais les principaux avantages de cette technologie sont d'une part d'être flexible, c'est-à-dire qu'on peut à loisir adapter le nombre de signaux transmis au canal de transmission. Une chute de débit n'entraînera pas de coupure de la transmission, mais simplement une perte de précision spatiale par l'abandon des composantes d'ordres les plus élevés. Et d'autre part, elle est indépendante du système de reproduction. Ainsi, un même signal HOA pourra être diffusé sur toutes les configurations de haut-parleurs ; il suffira pour cela de calculer la matrice de décodage associée à ce système.

1.3.2. Limites pratiques

Dans la pratique, cette technologie a toutefois des limitations. Si l'aliasing spectral est évité, on constate aux ordres les plus élevés l'apparition d'un repliement spatial lors des enregistrements naturels, à partir d'une fréquence liée à la distance maximale entre 2 capsules du microphone HOA (qui doit être au plus égale à la moitié de la plus petite longueur d'onde d'après Shannon) et au nombre total de capsules (qui doit être supérieur à $(M+1)^2$, M étant l'ordre auquel on souhaite enregistrer).

Le repliement spectral représente la présence importune dans des composantes ambisoniques de sources qui, de part leur position, ne devraient qu'apparaître avec cette importance que dans les composantes d'ordre plus élevé. Il est dû à l'intrusion des directivités d'ordres supérieurs dans les composantes harmoniques sphériques estimées. Comme on le voit sur la Figure 1.7, les directivités ne sont pas des cylindres (au sens mathématique du terme, la base n'étant pas forcément un cercle) mais varient en fonction de la fréquence, pour ressembler davantage dans le haut du spectre aux courbes de directivité de l'ordre supérieur. Ainsi, la directivité de l'ordre 1 ($2^{ème}$ graphique), qui devrait être un 8, ressemble davantage au-dessus de 10 000 Hz à la figure de directivité de l'ordre 4 (8 lobes).

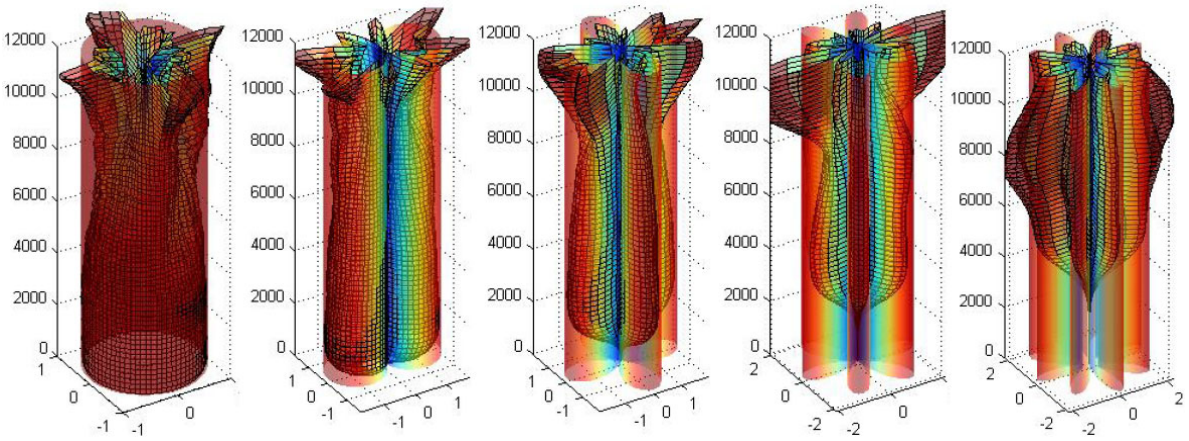

Figure 1.7 - Directivités théoriques (en transparence) et reconstruites (grille) dans le plan horizontal pour des ondes planes horizontales. L'axe vertical correspond à la fréquence, et les différents graphiques aux ordres 0,1,2,3 et 4 de gauche à droite.

À l'opposé, la taille réduite du microphone, si elle permet de réduire cet aliasing spatial, empêche la capture des basses fréquences aux ordres supérieurs (voir 2.3.2). La résolution spatiale des basses fréquences sera donc moindre, ce qui implique que, en ajoutant l'effet du repliement en hautes fréquences, les ordres élevés captés avec le micro HOA ne constituent qu'une plage du spectre assez peu étendue ; c'est cette seule plage qui pourra se targuer d'une excellente précision spatiale.

Chapitre 2

Analyser un contenu HOA

2.1. Objectifs

La réflexion a tout au long du stage été orientée vers deux objectifs : le premier est, dans un contexte de transmission, de reconstruire une scène spatialisée à partir du minimum d'informations ; le deuxième, appliqué à des ordres plus élevés, est d'arriver à resynthétiser les composantes de l'ordre immédiatement supérieur.

2.1.1. Restitution de la spatialisation à partir d'un downmix mono

La première approche vise donc à reconstruire une scène spatialisée à partir d'un downmix le plus léger possible. Un downmix est une réduction du nombre de canaux par fusion, l'exemple le plus simple étant le passage d'un signal stéréo à un signal mono construit comme la demi-somme des composantes gauche et droite. On tente dans cette approche de se restreindre au plus petit nombre de signaux possible, un signal étant très coûteux à transmettre, en ajoutant parallèlement à ce signal des informations complémentaires aidant à la reconstruction.

Le cas extrême est la réduction du flux à un signal mono. Dans le cas d'ambisonics, un signal mono comprenant toutes les sources est déjà présent : il s'agit de la composante omnidirectionnelle W. Il faut donc extraire de la scène encodée à un certain ordre M des informations permettant la reconstruction d'un ordre M' après transmission. Nous verrons que nous allons chercher à trouver les positions des sources grâce à une analyse en composantes principales (ou ACP, voir Annexe C) effectuée par sous-bandes (voir Annexe D).

2.1.2. Amélioration de la précision spatiale

La deuxième piste suivie a pour but de reconstituer un ordre ambisonique non transmis. On pourrait qualifier cette approche d'*upmix* ambisonique, car elle vise à reconstruire un ordre manquant. En effet, le format ambisonique étant flexible, il est bien

adapté à la transmission. Ainsi, si on souhaite transmettre une scène encodée à l'ordre 2 dans le plan horizontal, il faudra un canal d'un débit minimum (si l'on prend l'hypothèse de pas vouloir compresser les signaux) de 4,71 Mo/s. Un canal plus petit (par exemple 3.4 Mo/s) ne permettra théoriquement de transmettre qu'un ordre 1. Mais l'analyse des signaux peut permettre de déterminer les positions des sources, à partir desquelles on peut synthétiser des composantes du 2nd ordre approchant les composantes d'origine. On pourrait même imaginer pouvoir synthétiser un ordre qui n'avait pas été encodé à l'origine !

Pour cela, il est nécessaire de déterminer les positions des sources prédominantes dans le champ sonore, pour piloter leur encodage spatial aux ordres supérieurs et arriver à la synthèse d'une représentation de résolution spatiale accrue. Il s'agit dans le même temps d'isoler au mieux les signaux associés à ces différentes sources avant de leur appliquer l'opération d'encodage ; on se rapproche donc d'un problème de séparation de sources.

Pour parvenir à trouver ces positions, deux méthodes ont été testées. La première consiste comme précédemment à appliquer en sous-bandes une analyse en composantes principales. La seconde implique la formation de directivités à partir des vecteurs propres de la matrice de corrélation utilisée dans cette analyse.

L'analyse en composantes principales, ou ACP (voir Annexe C) est donc largement privilégiée dans nos approches. Elle permet en effet très facilement d'extraire des informations précises de signaux corrélés. Pour justifier ce choix, une autre méthode récente a également été testée et comparée à l'ACP (voir 1.1).

2.2. Principe de l'extraction d'information spatiale avec une ACP

2.2.1. Information spatiale

Nous avons donc tout d'abord testé l'analyse en composantes principales sur des signaux HOA. Nous avons choisi dans un premier temps de limiter l'analyse aux composantes horizontales du premier ordre, W, X et Y, avant de l'étendre à la 3ème dimension avec Z.

Dans un premier test, les signaux HOA sont analysés par une ACP dans leur intégralité. L'ACP est appliquée sur les signaux X et Y, qui contiennent l'information spatiale. En effet, X et Y correspondent aux signaux qui seraient captés par des microphones à figure en 8 placés sur l'axe avant-arrière et gauche-droite respectivement.

La Figure 2.1 montre la corrélation entre X et Y dans le cas d'un bruit blanc capté par le microphone ambisonique. Les points ne sont pas exactement alignés du fait des artefacts inhérents à la prise de son (réflexions sur les parois, approximations du microphone, etc.)

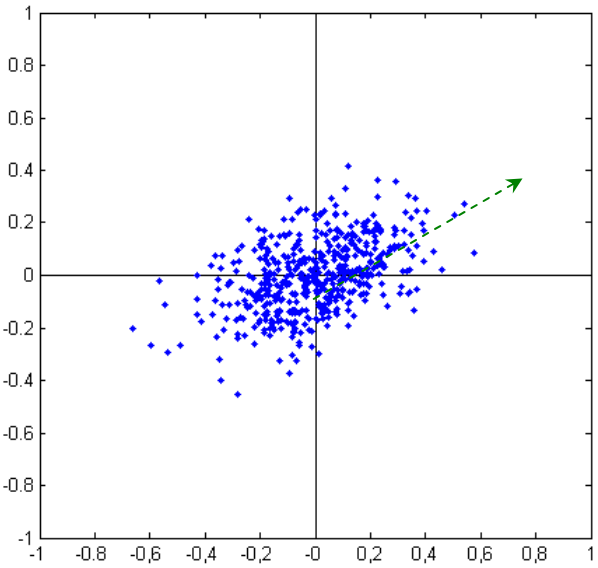

Figure 2.1 - Y en fonction de X, dans le cas d'un bruit blanc enregistré placé à 30°. On remarque que le nuage est étiré dans la direction de provenance du son, indiquée par la flèche en pointillés verts.

2.2.2. Application de l'analyse en composantes principales

La 1ère composante issue de l'ACP appliquée dans le plan (X, Y) pointe dans la direction dans laquelle X et Y ont la plus grande corrélation. Or comme X contient l'information placée sur l'axe avant-arrière et Y l'information de l'axe gauche-droite, cette direction est directement la direction de la source. Par exemple, un son situé droit devant dans l'axe (à 0°) aura une composante Y nulle, et par conséquent les échantillons seront dans le plan tous disposés sur l'axe horizontal. L'ACP prendra donc naturellement cet axe comme 1er vecteur propre, et il pointe bien vers la position de la source. De même, un son situé à 135° ($\frac{3}{8}\times 2\pi$ rad) aura la propriété Y = -X. Les échantillons seront donc placés sur la 2ème bissectrice, qui fait avec l'axe Y=0 un angle de 135° (voir Figure 2.2).

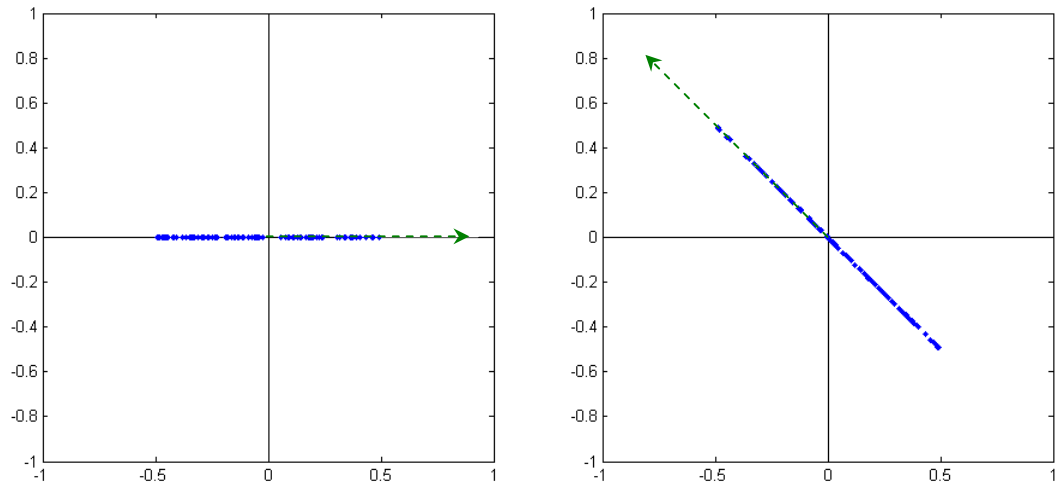

Figure 2.2 - Y en fonction de X, dans le cas d'un son placé artificiellement dans la direction indiquée par la flèche en pointillés verts, à 0° (à gauche) et 135° (à droite).

L'analyse de X et Y ne permet cependant de déterminer l'angle qu'à π près ; il reste une indétermination (on détermine la direction d'un axe qu'il reste à orienter). Pour lever cette ambiguïté, on intègre W à l'analyse, car cette composante, pourtant omnidirectionnelle et donc dépourvue d'information spatiale, apporte une référence pour les phases. Ainsi, notre son placé à 135° vérifiait (aux éventuels coefficients de normalisation près) Y = -X = W, alors qu'un son placé à 135 - 180 = -45° aurait la propriété Y = -X = -W.

Pour prendre en compte W dans l'analyse, on peut comparer directement les phases de ces 3 canaux. Algorithmiquement, on ramène l'angle trouvé à son modulo $\pi/2$, puis on détermine le vrai cadran par comparaison des phases de W, X et Y, c'est-à-dire en comparant le signe des produits W · X et W · Y. Ainsi, W · X > 0 signifie que W et X sont en phase, et réciproquement (voir Figure 2.3).

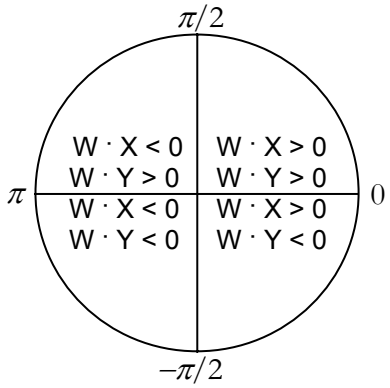

Figure 2.3 - Signes de W · X et W · Y en fonction du cadran

On peut également intégrer W à l'analyse en appliquant l'ACP sur le triplet (X, Y, W). Le problème d'ambiguïté est ainsi résolu, la direction recherchée étant celle du vecteur principal dans le demi-espace W > 0 projetée sur le plan (X, Y). Cette méthode s'avère plus robuste et surtout plus généralisable à des ordres plus élevés.

Cette analyse permet donc de déterminer la position d'une source unique. Pour pouvoir séparer plusieurs sources, nous avons appliqué l'ACP par sous-bandes, en supposant que les sources n'ont pas des supports fréquentiels strictement identiques. Le principe de cette analyse en sous-bandes est développé en Annexe D.

2.3. Méthodologie

2.3.1. Construction de scènes artificielles

Pour tester les algorithmes implémentés, nous avons dû créer des scènes sonores spatialisées au format HOA. Plusieurs types de scènes ont été utilisés, de complexités différentes. Les premières ont été des scènes virtuelles synthétisées avec Matlab. Les premières scènes, simplissimes, ne comportait qu'une sinusoïde ou un bruit blanc fixe placé dans le plan horizontal. Puis, pour tester l'analyse par sous-bandes, ont été créés des sons polyphoniques, composés de 2 sinusoïdes tout d'abord, puis de sons complexes (synthétisés par synthèse FM), dans des gammes fréquentielles distinctes ou se

chevauchant. Enfin, ces sons ont été rendus mobiles en faisant varier leur position dans le temps.

Nous avons pu ainsi constater à l'aide des sons fixes et de supports fréquentiels distincts la précision de l'analyse, et avec les sons mobiles et mêlés sa robustesse. Nous nous sommes limités sous Matlab à la génération de 2 sons simultanés et du 1er ordre, et un autre outil, Plogue Bidule, a été utilisé pour composer des scènes plus complexes.

Bidule est un studio modulaire virtuel ressemblant à MAX/MSP. Des plugins ont été développés à France Télécom pour permettre de manipuler des signaux HOA. Ainsi, des scènes complexes ont pu facilement être synthétisées à des ordres élevés. Des scènes à 3 et 4 sources, fixes ou dont 1 source mobile, ont été réalisées à partir de voix démixées, placées dans le plan horizontal ou en 3 dimensions.

Ces scènes étant artificielles, elles sont en un sens idéales : ne contiennent aucun signal réverbéré (effet de salle) ni diffracté (par un micro), l'encodage est fait sans artefacts, les ondes sont considérées comme planes et ont donc une direction bien définie. Si on est ici bien loin de la vraie vie, le gros avantage des scènes artificielles est qu'on connaît précisément à chaque instant la position des sources et qu'on peut ainsi vérifier le résultat de l'analyse.

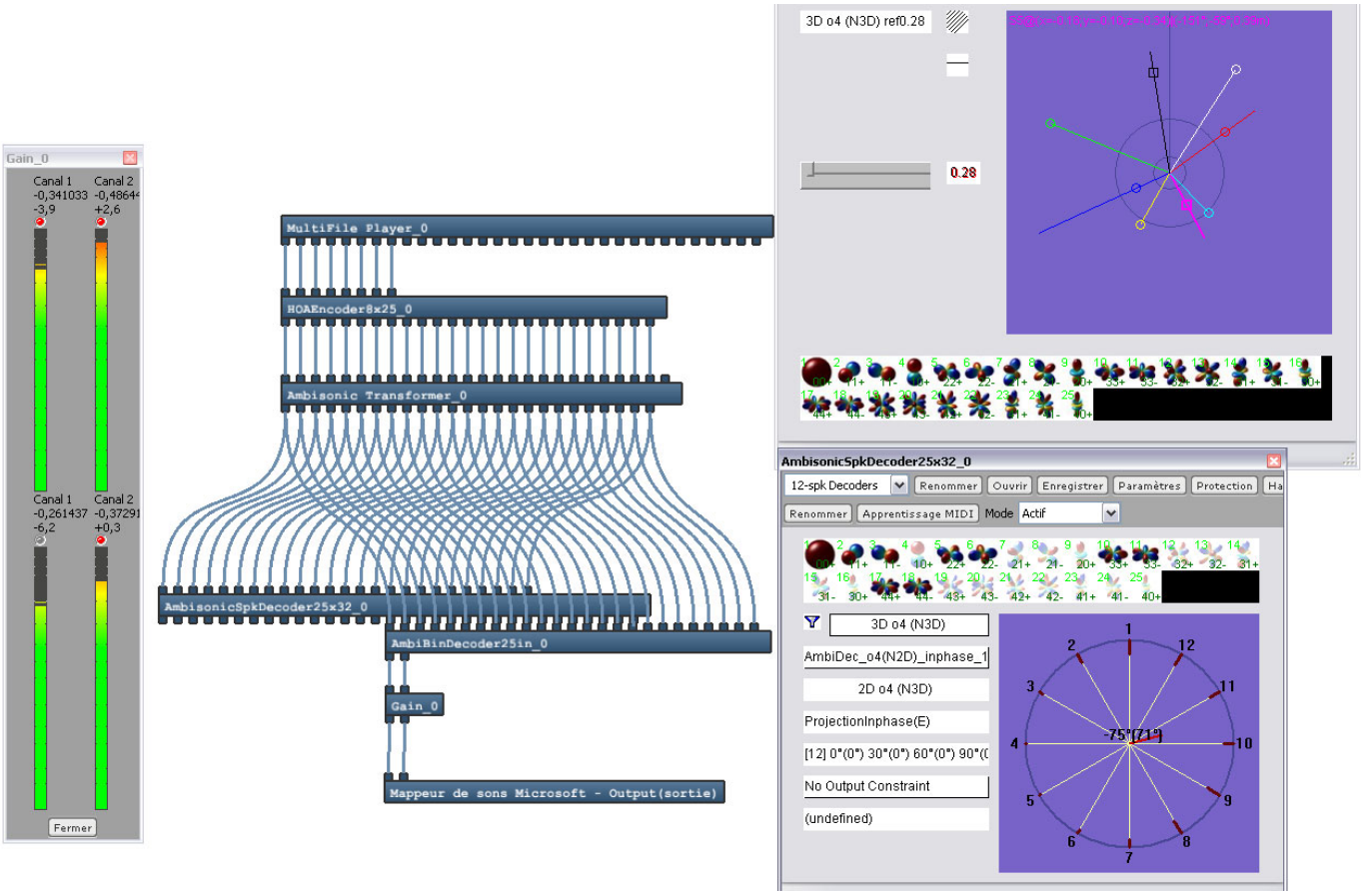

Figure 2.4 - Interface graphique du logiciel Bidule, avec à droite les contrôles des encodeurs et décodeurs HOA créés à France Télécom.

2.3.2. Exploitation d'enregistrements naturels

Une fois des résultats convaincants obtenus avec des scènes virtuelles, des enregistrements naturels ont été analysés. Ces scènes ont été enregistrées grâce au prototype de micro HOA du 4ème ordre développé à France Télécom (voir Figure 1.2). Ce processus implique des différences notables avec un encodage virtuel.

Le prototype de microphone contraint en effet à quelques compromis pour optimiser des critères incompatibles. La taille du microphone résulte d'un compromis entre la précision d'estimation des composantes en basses fréquences, qui nécessite une taille élevée, et la minimisation de l'aliasing spatial, qui ne peut être évité qu'avec des capsules très proches les unes des autres. Leur disposition est également irrégulière ; il n'existe en fait aucune façon de disposer régulièrement plus de 20 points sur une sphère. La configuration qui avait été retenue par les concepteurs [Moreau, 2006] est la "moins irrégulière possible" au sens de la préservation de l'orthonormalité des harmoniques sphériques : les 32 capsules sont réparties sur une sphère de 3,5 cm de rayon selon un Pentaki-dodécahèdre, polyèdre semi-régulier dual de l'icosaèdre tronqué (ballon de football).

Ces artefacts sont donc inhérents à l'utilisation d'un réseau de microphones. Mais le monde réel pose également des problèmes délicats. La complexité des modes de rayonnement, d'interaction avec l'environnement, réclame l'introduction de modèles d'encodage beaucoup plus complexes que le cas d'une onde plane en champ libre et rend l'information spatiale moins limpide et univoque au sein de la scène encodée (enregistrée), en comparaison avec une scène virtuelle. Les sources ne sont ainsi plus ponctuelles mais étendues, elles ne sont plus à l'infini mais à une distance finie du microphone (les ondes reçues ne sont donc plus planes). En outre, chaque source, en se réfléchissant sur les parois et les différents objets de la salle, crée une multitude de sources secondaires. Il est enfin souvent difficile de connaître avec précision la position d'une source dans l'espace lors d'un enregistrement.

Des enregistrements en chambre sourde ont été réalisés pour se passer des réflexions de l'environnement. Afin de connaître avec précision la position de la source, un haut-parleur diffusant un bruit blanc, nous avons placé le microphone sur une table tournante commandée par un plug-in sous Plogue Bidule.

2.3.3. Visualisations des résultats

L'observation des résultats a été effectuée de plusieurs manières, visuellement et par l'écoute. Les graphiques étant plus commodes, c'est par cette voie que les premiers essais étaient conclus. Plusieurs types de graphiques ont été produits, décrivant des paramètres différents.

Les représentations temps-fréquence permettent de représenter une variable à 1 dimension sous Matlab. Cette voie a été suivie pour visualiser les paramètres extraits de chaque sous-bande de chaque fenêtre, notamment les angles, l'énergie et la pertinence. Malheureusement, Matlab ne permet pas de représenter sur le même graphique une 4ème dimension, sous la forme d'une 2ème dimension colorimétrique (on prend généralement la

saturation). Cette dimension supplémentaire aurait permis de représenter simultanément les angles et leur pertinence.

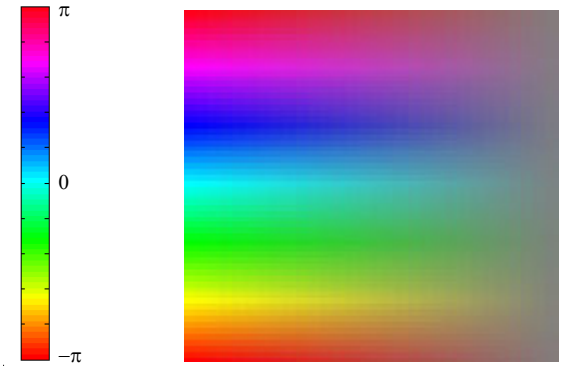

Figure 2.5 - Représentation de la table des couleurs dite "HSV" de Matlab (linéaire, à gauche) et de l'espace colorimétrique HSL (Hue Saturation Luminance, à droite)

Pour représenter simultanément angles et pertinence, il était donc nécessaire d'abandonner une dimension. C'est le découpage fréquentiel qui a été mis de côté, pour pouvoir représenter dans un plan temps-angles les pertinences. Celles-ci sont représentées soit par le niveau de gris des points, soit par leur taille.

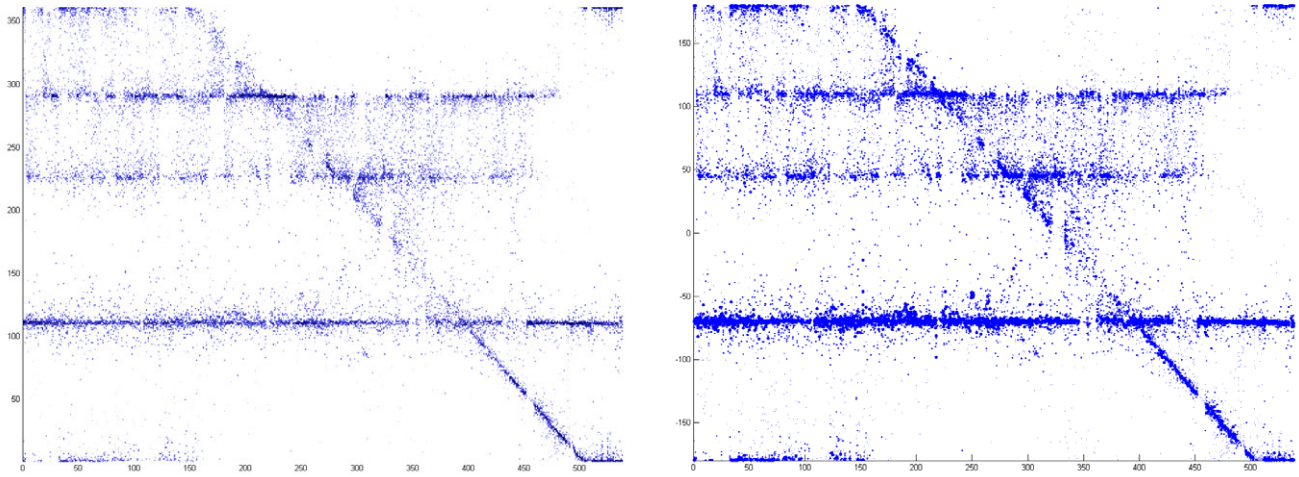

Figure 2.6 - Représentation de la pertinence des angles de la scène à 4 sources dont 1 mobile décrite en 3.3.2 (en abscisse : les trames, en ordonnées : les angles en degré). La pertinence est représentée à gauche par l'intensité du bleu, et à droite par la taille des points.

2.3.4. Différents modes d'écoute

La visualisation est un indice précis de résultat, mais l'écoute est tout de même indispensable pour vérifier la qualité globale de la synthèse. Plusieurs modes d'écoute sont possibles ; de ceux-ci, l'écoute au casque est la plus aisée. Une pseudo-stéréo est ainsi reconstituée à partir des signaux du 1er ordre[1] pour restituer les scènes restreintes au

[1] W+Y+0.2X pour le canal gauche et W-Y+0.2X pour le canal droit

plan horizontal. Pour plus de précision à la restitution ou pour des scènes en 3D, et sans perdre la facilité d'une écoute personnelle au casque, on peut appliquer des filtres HRTF (Head-Related Transfert Function) pour une reproduction binaurale.

Pour apprécier pleinement la spatialisation, la diffusion sur un système multi haut-parleurs est indispensable. Deux systèmes sont installés dans le laboratoire de France Télécom, un système 5.1 et un système circulaire de 48 haut-parleurs. Ces systèmes, installés dans des studios insonorisés, offrent une grande qualité d'écoute qui permet de repérer les défauts même minimes de la synthèse.

Chapitre 3

Description des analyses

3.1. Détermination des directions principales

3.1.1. Base des algorithmes

Nous avons développé plusieurs algorithmes répondant aux différents objectifs décrits plus haut. Chacun d'eux possède des particularités leur permettant de répondre à ces objectifs, mais on trouve néanmoins une partie commune. Ainsi, toutes les analyses reposent sur le même modèle d'application d'une ACP par sous-bandes, dont voici le schéma général :

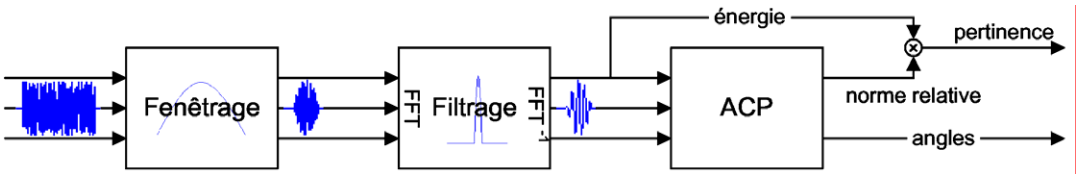

Figure 3.1 - Schéma général d'analyse

Les signaux en entrée sont les composantes HOA W, X et Y dans la plupart des cas, mais peuvent être également des composantes d'ordre plus élevé (voir 1.1). Sur le fenêtrage jouent plusieurs paramètres, comme la taille de la fenêtre, ou encore le pourcentage de recouvrement entre elles. Nous avons choisi une fenêtre en racine de la fenêtre de Hanning, c'est-à-dire une arche de sinusoïde (voir équation (D.7) en Annexe D), avec un recouvrement de 50%, ce qui vérifie le critère de reconstruction parfaite (D.6). Le filtrage est, pour des raisons de rapidité de calcul, réalisé dans le domaine fréquentiel par une simple multiplication, avant de repasser en temporel pour appliquer l'ACP.

Le banc de filtre utilisé est un banc inspiré des filtres ERB. Les filtres ERB (Equivalent Regular Bandwidth) sont des filtres imitant le comportement de séparation fréquentielle de l'oreille humaine, plus précise en basse fréquence. Les filtres utilisés ici ne sont pas directement des filtres ERB : le système n'a pas d'intérêt particulier à essayer de copier le fonctionnement de l'oreille humaine. Ces filtres ont été adaptés au problème, en

s'assurant en particulier de la finesse des bandes permettant la meilleure discrimination des différentes sources, c'est-à-dire autour des fondamentales usuelles (200-1000 Hz) et des harmoniques. Les fréquences aiguës sont bien discriminées par les filtres ERB (relativement à la gamme, c'est-à-dire en échelle logarithmique), le découpage a donc surtout été densifié entre 200 et 800 Hz par rapport aux bandes ERB.

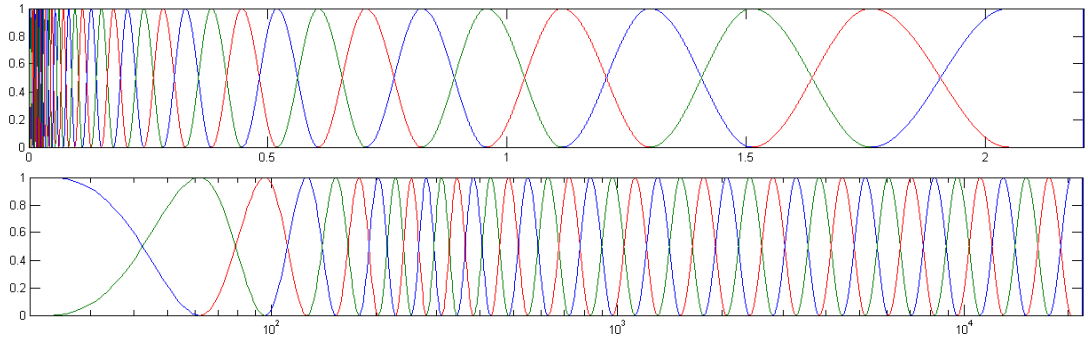

Figure 3.2 - Filtres utilisés, inspirés des filtres ERB, représentés linéairement (en haut) et sur une échelle logarithmique (en bas)

Les résultats de l'ACP sont donc avec cette méthode les angles des sources trouvées dans chaque sous-bande ainsi que pour chacun d'eux un critère de pertinence, calculé en fonction de l'énergie de la sous-bande dans une fenêtre donnée et de la norme relative de la première composante de l'ACP (voir 3.4.2). Les autres vecteurs propres peuvent être utilisés pour extraire davantage d'informations, et même des signaux décorrélés, comme on le verra au paragraphe 3.5, mais l'analyse semble alors moins robuste. S'en tenir aux informations données par la composante principale est un gage de pertinence supplémentaire.

Ces résultats concernent donc les sources prédominantes, ou principales, mais laissent de côté un certain nombre d'événements sonores à considérer.

3.1.2. Traitement du champ résiduel

Le champ résiduel regroupe en quelque sorte tout ce qui n'est pas une source principale. Il inclut donc les réflexions et la réverbération de la pièce, mais aussi les bruits parasites occasionnels ou continus (ventilation, etc.). Ces bruits ou échos n'ont pas de direction de provenance assez distincte pour pouvoir être mesurée précisément. Le champ résiduel est également composé du champ diffus. La notion de champ diffus est couramment utilisée mais rarement précisément définie. On trouve dans la littérature des définitions diverses, l'une d'elles précisant qu'il serait un ensemble d'événements sonores proches en temps et provenant de directions équiprobables.

Les algorithmes proposés traitent correctement les sources principales, mais le champ résiduel est laissé de côté. Il est pourtant très important pour l'immersion du sujet dans la scène : une scène dépourvue de champ sonore sera perçue comme artificielle, ce qui nuit à la perception d'espace. On sera par exemple davantage conscients d'être au centre d'un dispositif expérimental et moins sensibles aux variations de profondeur et de perspectives.

Néanmoins, il faut relativiser l'importance de ce champ résiduel : il n'est pas utile pour la localisation des sources principales et ne caractérise que l'environnement d'enregistrement. En outre, le champ résiduel, qui provient souvent de directions floues, et notre but n'est pas de les préciser artificiellement. Il n'est dès lors généralement pas utile de le transmettre aux ordres élevés.

Ainsi, certains auteurs [Merimaa and Pulkki, 2005] ont choisi de recréer lors d'une synthèse un champ diffus artificiel basé sur une réponse impulsionnelle de salle calculée à l'avance. L'avantage est une immersion, un enveloppement complets, l'inconvénient est que l'ambiance ainsi recréée est fictive et arbitraire, et en outre complètement indépendante de l'ambiance initiale, en laissant notamment de côté les sources secondaires.

Nous proposons dans nos algorithmes de ne pas synthétiser le champ diffus aux ordres élevés, mais de nous contenter de la description du champ résiduel fourni par les ordres transmis (1er ordre dans nos exemples). Nous nous limiterons donc à préciser par la synthèse des ordres supérieurs les positions des sources principales.

3.2. Extension à la 3D

Nous avons pour le moment évoqué le cas de représentations de champ sonore à 2 dimensions, c'est-à-dire issus de sources placées dans le plan horizontal. Mais certaines représentations comportent également des composantes verticales. Pour analyser ces signaux en 3 dimensions, il faut intégrer la composante Z. On effectue alors simplement une ACP sur X, Y, Z et W. Le 1er vecteur propre sera alors bien dirigé dans l'espace (X, Y, Z) dans la direction de la source, l'ambiguïté de ½ espace étant levée par W.

L'algorithme est alors complètement similaire à celui analysant des signaux en 2 dimensions. Des tests ont été réalisés à partir de signaux synthétiques pour lesquels les positions des sources étaient connues ; pour des raisons pratiques, aucune scène sonore réelle en 3 dimensions n'a pu être enregistrée avec une connaissance initiale précise des positions des sources.

Les visualisations et les écoutes étant plus délicates en 3D, et l'extension à la 3D étant immédiate, nous avons donc concentré nos efforts sur l'analyse de scènes en 2D. En outre, en l'absence de dispositif sphérique ou hémisphérique de haut-parleurs, le seul moyen à notre disposition pour écouter une scène 3D est l'écoute binaurale, c'est-à-dire au casque avec des fonctions de transfert de tête (HRTF) relativement bien adaptées à l'auditeur.

3.3. Analyse aux ordres supérieurs

3.3.1. Description

Afin d'améliorer la précision de l'analyse, nous avons essayé d'appliquer une ACP sur des signaux ambisoniques d'ordre plus élevé (signaux HOA), dans le plan horizontal. Deux méthodes ont été essayées.

La première consiste à appliquer une ACP à 3 dimensions successivement sur les composantes X, Y, W (ordre 1 et 0) puis U, V, W (ordre 2 et 0) et ainsi de suite, c'est-à-dire (en utilisant la notation générale) sur les composantes $Y_{m0}^{(1)}$, $Y_{m0}^{(-1)}$ et W (ordre m et 0). Les résultats de ces analyses (les directions des 1$^{\text{ères}}$ composantes) sont ensuite regroupées, l'analyse de l'ordre 1 fournissant des résultats modulo 2π, l'analyse de l'ordre 2 des résultats plus précis mais modulo π, et plus généralement, l'analyse de l'ordre m permettant de trouver des angles modulo $2\pi/m$.

La limite de cette approche se perçoit lors de la fusion des résultats : les angles trouvés à un ordre élevé sont plus précis dans des bandes de fréquences élevées, mais la limitation théorique de ces signaux HOA dans les basses fréquences font que l'analyse de l'ordre 1 reste la plus pertinente dans les premières bandes. Parvenir à déterminer à partir de quelle fréquence on peut faire davantage confiance à l'ordre n+1 qu'à l'ordre n n'est pas aisé, et cette limite provoque des aberrations dans les angles trouvés. Une source unique pourra par exemple être interprétée par l'algorithme des k-moyennes (voir 3.4.1) comme 2 sources, une en hautes fréquences et une en basses fréquences, les 2 étant trouvés à 2 angles différents, chaque angle relevant d'un ordre analysé différent.

La deuxième méthode consiste, pour éviter cette fusion des résultats, à effectuer une seule ACP sur les composantes U, V, X, Y et W. Cette méthode souffre d'une très grande complexité algorithmique, pour une amélioration de la précision peu sensible. En effet, si cette méthode permet d'obtenir des résultats plus précis dans certaines bandes de fréquences, elle est sensible aux défauts et les limites des ordres supérieurs (l'absence de basses fréquences entre autres) perturbent l'analyse en composantes principales.

3.3.2. Illustration sur une scène virtuelle

La scène présentée ici est une scène artificielle, c'est-à-dire composée de sons placés artificiellement à diverses positions. Elle a une durée totale de 13,5 s., et comporte 4 sources positionnées dans le plan horizontal : une guitare placée à 45°, une basse à -70°, une batterie à 110° et une voix féminine qui effectue continûment un tour complet dans le sens indirect en partant de l'arrière (elle passe donc successivement par les positions 180°, 90°, 0°, -90° et -180°). La guitare et la batterie s'arrêtent peu avant la fin de la scène (12$^{\text{ème}}$ seconde, au 9/10$^{\text{ème}}$ du son approximativement).

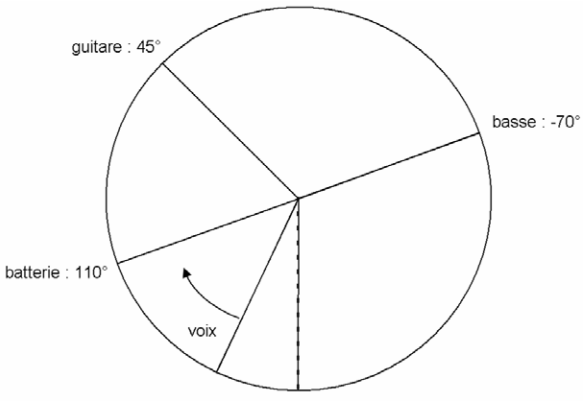

Figure 3.3 - Position des sources de la scène de test à 4 sources dont 1 mobile.

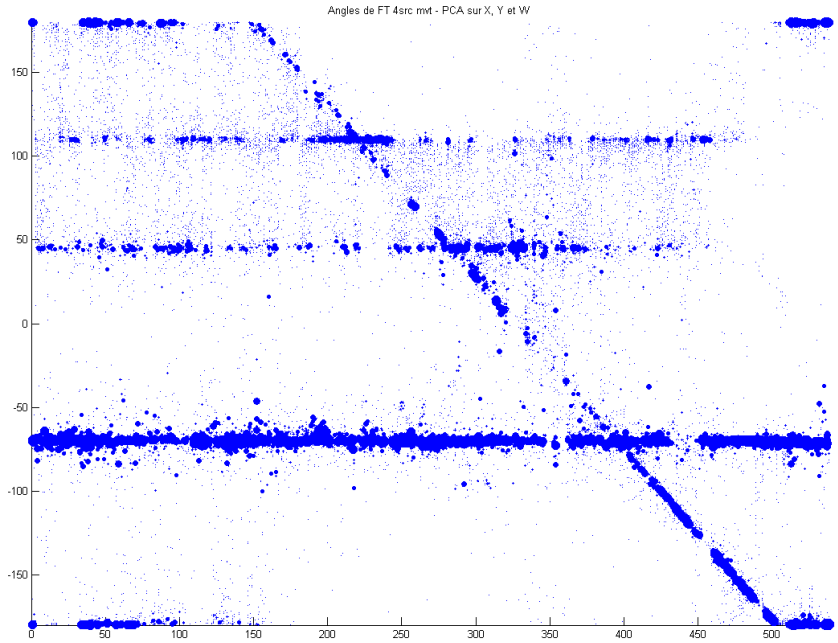

Figure 3.4 - Angles d'une scène virtuelle ayant 1 source mobile et 3 sources fixes (décrite en 3.3.2) trouvés à partir d'une ACP sur les ordres 0 et 1.
Pour chaque trame (en abscisse), 20 angles (1 par bande, en ordonnée et en degrés) sont reportés, le diamètre du point correspondant à un critère de pertinence (voir 3.4.2)

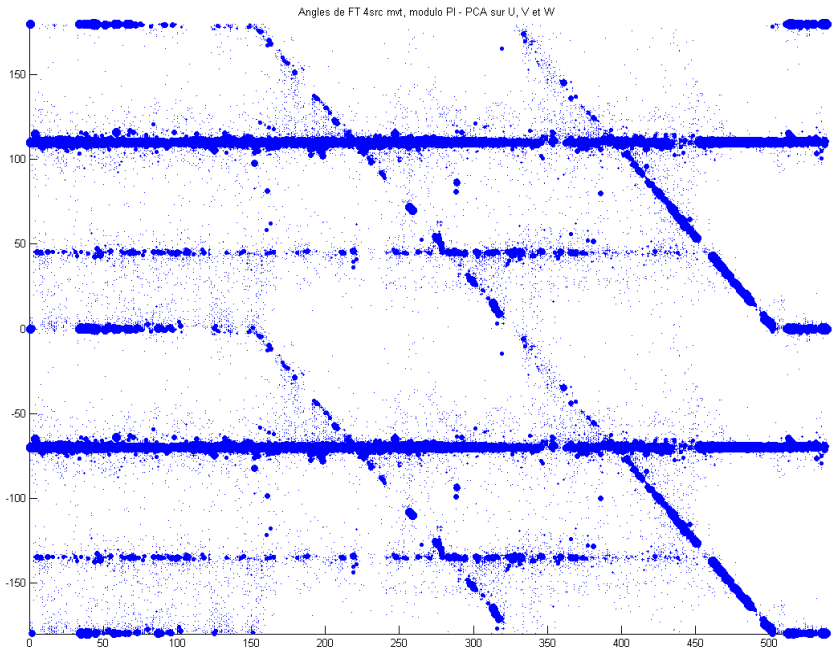

Figure 3.5 - Angles de la même scène trouvés à partir d'une ACP sur les ordres 0 et 2. Résultats modulo π répliqués pour faciliter la comparaison avec la figure précédente. On remarque que les lignes sont plus nettes, qu'il y a moins d'angles aberrants et que les angles corrects sont plus marqués (ils sont déterminés comme plus pertinent)

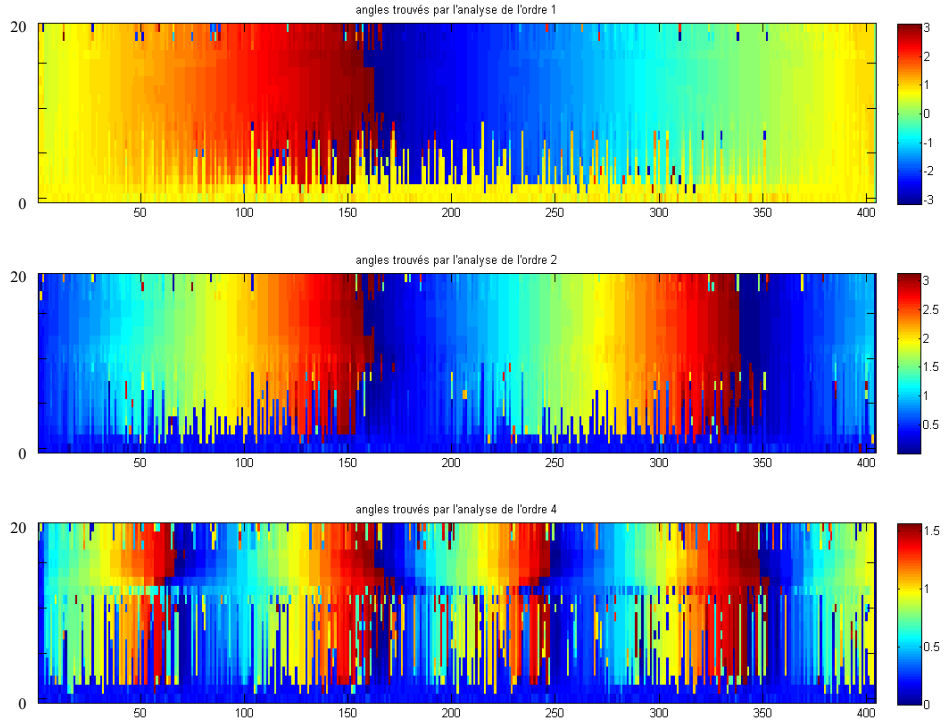

Figure 3.6 - Angles d'une scène réelle constituée d'un bruit parcourant le cercle, trouvés à partir d'une ACP sur les ordres 0 et 1 (en haut), 0 et 2 (au milieu), et 0 et 4 (en bas). Le numéro de trame est reporté en abscisse, et le numéro de bande en ordonnée. On retrouve qu'une limitation des ordres HOA supérieurs issus d'un enregistrement microphonique est la restitution des basses fréquences.

3.4. Traitement des données

3.4.1. Algorithme des k-moyennes

À l'issue de l'analyse en composantes principales, on récupère avec la première des méthodes exposées au chapitre précédent des angles correspondants aux positions des sources prédominantes à raison d'une valeur par fenêtre et par bande de fréquence. Il faut à partir de ces données parvenir à isoler les sources, c'est-à-dire à déterminer quels sont les angles pertinents et examiner quels sont ceux qui, à une certaine tolérance près, sont présents dans plusieurs bandes de fréquence et/ou plusieurs fenêtres temporelles.

Pour pouvoir exploiter ces données spatiales, nous avons utilisé l'algorithme des k-moyennes (ou k-means) [MacQueen, 1967] sur les angles pondérés par des critères de pertinence. Cet algorithme fonctionne de la manière suivante :

> Initialisation : choix de *k* barycentres (les *k* moyennes)
> Répéter :
> Associer chaque élément au barycentre le plus proche
> Recalculer les positions des barycentres
> Jusqu'à ce que les barycentres soient stables (ou au bout de *t* itérations)

L'avantage de cet algorithme est d'être peu glouton puisqu'il converge en $O(tkn)$, où t est le nombre d'itérations, k le nombre de centres et n le nombre de points (généralement, on a $t, k \ll n$). Par contre, il est plutôt sensible aux exceptions : un point aberrant tirera à lui la moyenne auquel il aura été associé. Il faut donc faire en sorte que ces points erronés soient aussi rares que possibles. C'est pourquoi les angles traités sont pondérés par 2 critères de pertinence.

La pondération des angles fournis en entrée de l'algorithme des k-moyennes se fait de manière simple en répétant ces angles un nombre de fois (de 0 à 10) proportionnel à un critère. Ainsi, un angle très pertinent sera dupliqué et sera présent plusieurs fois dans la liste des données traitées, attirant ainsi à lui la position du barycentre le plus proche. Au contraire, un angle trop peu pertinent ne sera pas répété, voir même sera enlevé de la liste.

Le seul problème de cet algorithme est qu'il nécessite la connaissance de k, c'est-à-dire du nombre de sources qui composent la scène. Une détermination automatique se basant sur la répétition de l'algorithme avec un k croissant a été testée, mais le critère d'arrêt, reposant sur la dispersion des points autour des barycentres, n'étant pas assez robuste, cette méthode a provisoirement été mise de côté.

3.4.2. Critères de pertinence

Les 2 critères de pondération choisis sont l'énergie de la sous-bande dans la fenêtre donnée et le rapport de la norme de la 1ère composante (la composante principale) de l'ACP sur la somme des normes de toutes les composantes.

En effet, un angle est trouvé dans chaque sous-bande de chaque fenêtre, même si le signal y est très faible. L'angle trouvé sera alors très peu représentatif de la direction d'une source, puisqu'il indiquera la direction d'un bruit. L'énergie doit donc être prise en compte pour ne conserver que les angles trouvés dans des bandes comportant un signal suffisant.

La norme relative de la 1ère composante de l'ACP rend quant à elle compte du poids relatif de cette composante par rapport aux autres. Ainsi, si 2 sources sont présentes dans une bande donnée, la 1ère composante sera à peine plus marquée que la 2ème. Ceci peut être calculé en prenant simplement le rapport de la 1ère valeur propre sur la somme des valeurs propres.

On peut noter que la norme de la 1ère composante ne se détache nettement que lorsqu'un son est présent dans la bande considérée. On peut alors se demander si l'énergie doit être prise en compte, ou si l'information qu'elle contient n'est comprise dans le 2ème critère. En fait, la norme relative est un critère très bruité, et considérer l'énergie par ailleurs permet de nuancer ces artefacts, et constitue une base solide de pondération.

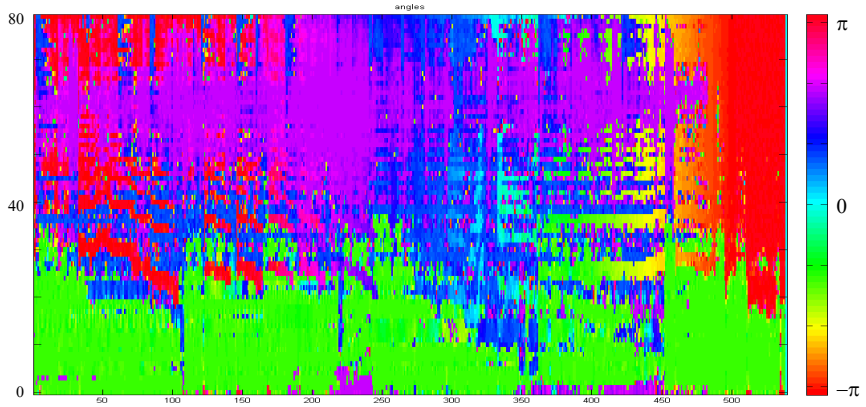

Figure 3.7 - Représentation temps-fréquence des angles du son décrit en 3.3.2. Le numéro de trame est reporté en abscisse, et le numéro de bande en ordonnée. On peut donc voir un angle par trame et par bande, chacun étant le résultat d'une ACP, c'est-à-dire la 1ère direction.

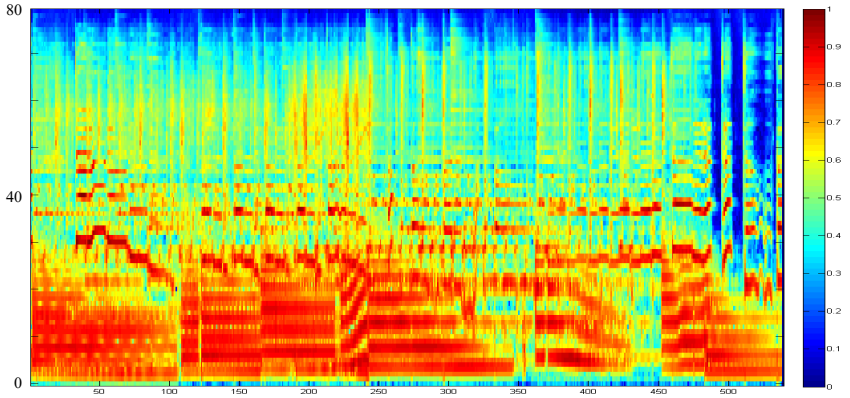

Figure 3.8 - Énergie de ce même son. On peut voir une basse en notes tenues, une voix aigüe, et une batterie qui produit des sons secs dans un spectre plus aigu.

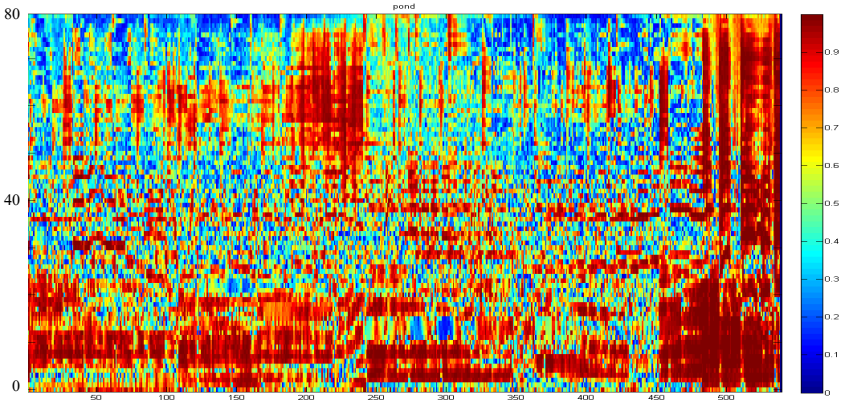

Figure 3.9 - Norme relative de la 1ère composante de ce même son. Les impacts de batterie (dans l'aigu) ressortent davantage.

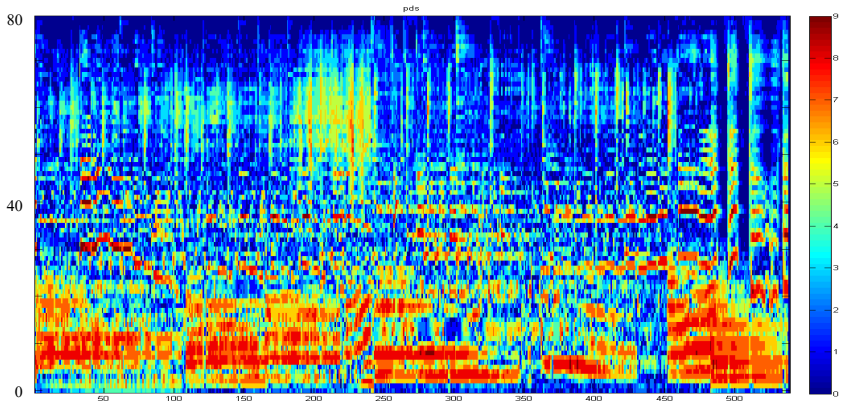

Figure 3.10 - Pondération appliquée à l'entrée des k-moyennes, calculée ici comme le simple produit de l'énergie et de la norme relative

3.4.3. Lissages

Pour faire face au bruit inévitable dans les résultats, il faut lisser les données. Ces lissages, temporels, correspondent à des hypothèses de stabilité des sources, qui sont supposées être fixes ou mobiles relativement lentement (dans une mesure à paramétrer).

Un premier lissage peut être opéré avant le traitement par l'algorithme des k-moyennes. Le filtre choisi est un filtre médian, de manière à éliminer les points aberrants tout en conservant les vraies valeurs des points cohérents. Ce lissage est appliqué sur chaque sous-bande aux angles, et ne tient pas compte des critères de pertinence. Il est donc utilisé avec parcimonie et peut même être évité en réglant de manière adéquate les critères de pondération à l'entrée des k-moyennes.

Le deuxième lissage s'effectue à l'entrée des k-moyennes, en même temps que les critères de pondération. Il consiste à appliquer les k-moyennes non pas sur une fenêtre donnée mais sur plusieurs fenêtres consécutives, antérieures et postérieures, en donnant plus de poids à la fenêtre courante. Dans la pratique, les poids donnés aux différentes fenêtres sont établis par une gaussienne centrée sur la fenêtre courante et de largeur à déterminer (en fonction notamment de la durée d'une fenêtre et du type de scène analysé).

3.5. Exploitation de toutes les composantes de l'ACP

Les analyses présentées précédemment ont toutes exploité uniquement la composante principale extraite par l'ACP. Or cette analyse fournit des informations plus complètes par l'intermédiaire des autres composantes, dont on peut tirer parti pour isoler plusieurs sources au sein d'une même bande. Cette approche est toujours en développement dans ce stage, et n'a pas encore donné pleine mesure des résultats qu'elle laisse entrevoir.

Les composantes de l'ACP indiquent des directions qui correspondent aux différentes sources composant la scène sonore. Les vecteurs propres de la matrice de corrélation donnent ainsi les combinaisons linéaires des signaux HOA permettant de

séparer de la meilleure façon possible ces sons. Ces directions peuvent être alors utilisées pour former des directivités isolant au mieux ces sources :

$$S = B \cdot a \tag{3.1}$$

où la matrice *a* contient les vecteurs propres de la matrice de corrélation (voir Annexe C), le vecteur *B* contient les signaux HOA, et le vecteur *S* contient les sources extraites

$$a = \left[\begin{pmatrix} \\ u_1 \\ \\ \end{pmatrix} \quad \cdots \quad \begin{pmatrix} \\ u_N \\ \\ \end{pmatrix} \right], \quad S = \begin{pmatrix} S_1 \\ \vdots \\ S_L \end{pmatrix}, \quad B = \begin{pmatrix} B_{00}^1 \\ \vdots \\ B_{M0}^1 \end{pmatrix} \tag{3.2}$$

On obtient ainsi des lobes d'autant plus étroits que l'ordre est élevé, chacun dirigé vers une source, et permettant de l'isoler le plus possible des contributions des autres sources. La Figure 3.11 présente les directivités ainsi trouvées, par une ACP en pleine bande. Une analyse en sous-bande permet une plus grande distinction des sources, mais une séparation différente entre les bandes est dangereuse car elle risque de dissocier spatialement plusieurs parties d'un même son.

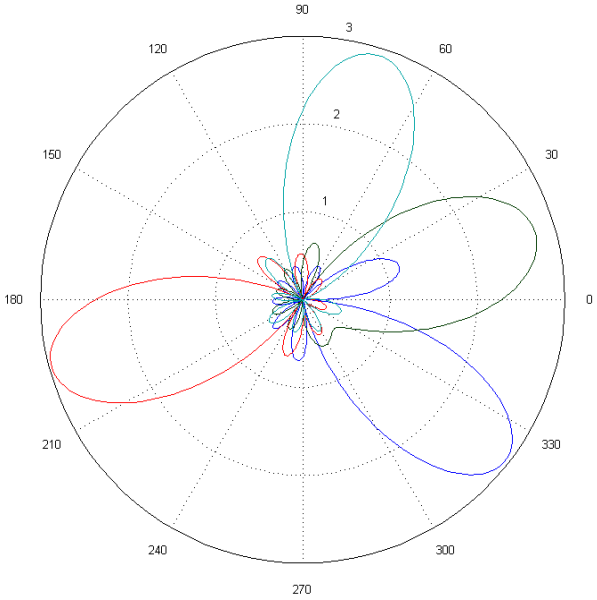

Figure 3.11 - Directivités séparant idéalement 4 sources placées à 17, 73, 195 et -58°, pour une scène virtuelle encodée à l'ordre 4.

Ce principe se heurte néanmoins aux limites de séparation liées à l'ordre d'encodage : en effet, 2 sources ne pourront pas être séparées si elles sont trop proches, comme en témoigne la largeur des lobes principaux. Les lobes secondaires traduisent également un mélange des sources isolées. La figure ci-dessus montre ainsi que la source placée à 17° (en vert) sera présente de manière non négligeable dans le signal sensé contenir la source placée à -58° (en bleu), en raison du lobe secondaire bleu important.

Chapitre 4

Résultats et applications

4.1. Illustration et critique des résultats obtenus

4.1.1. Avant-propos

Sans notification particulière, les résultats présentés ont été obtenus avec un pas de 50 ms, une analyse en 80 bandes, avec une FFT sur 8192 points. Une taille de FTT plus grande et un plus grand nombre de bande augmente la qualité de l'analyse, un pas plus petit améliore sa précision temporelle mais les résultats sont moins fiables et un lissage plus important s'impose alors. Ces paramètres ont été réglés après de multiples essais sur des scènes diverses.

Les résultats sont présentés ici uniquement sous la forme de graphiques, nous ne pouvons donc représenter que les angles trouvés à la sortie de l'algorithme des k-moyennes.

4.1.2. Scène 2D artificielle à 4 sources fixes

La première scène est une scène artificielle composée de 4 sons fixes, placés à -70, -10, 45 et 110°. Ces sons (une basse, une voix, une guitare, et une batterie) ont été choisis pour leur variété spectrale et leur variété d'attaque. Elle a en outre le mérite d'être réaliste et de constituer une séquence musicale (chaque son n'est pas une mélodie indépendante).

On voit sur la Figure 4.1 ci-dessous que les angles trouvés sont proches des angles théoriques. L'un des sons (à 45°) est quasiment nul avant la $100^{ème}$ trame, sa position est alors moins facile à déterminer ; plus précisément, une position approximative est trouvée, et correspondrait exactement à la position du bruit (45°) si l'incertitude de mesure très grande pour un son aussi faible ne rendait la détermination si ardue. On mesure un écart-type de 2.2° sur cet exemple (en excluant le début du son à 45°).

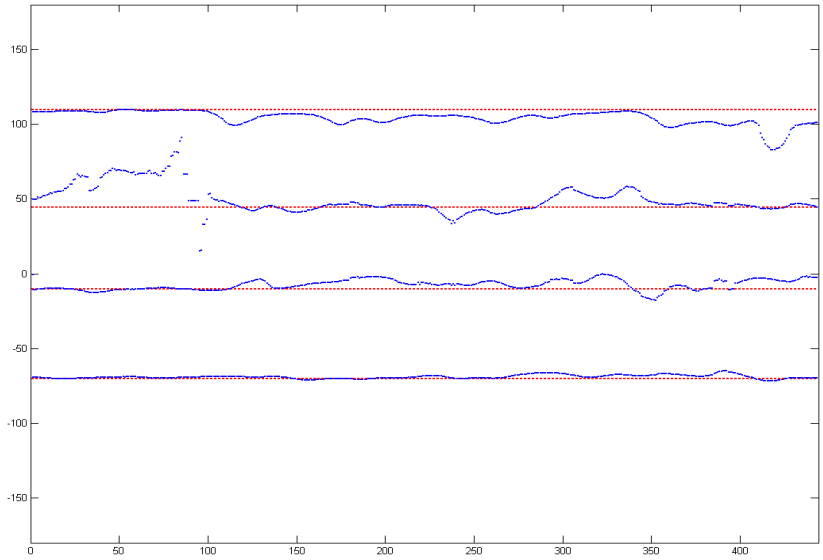

Figure 4.1 - Angles théoriques (en pointillés rouge) et calculés (en trait bleu) pour une scène constituée de 4 sons fixes longue de 11s.

4.1.3. Scène 2D artificielle à 4 sources dont 3 fixes et 1 mobile

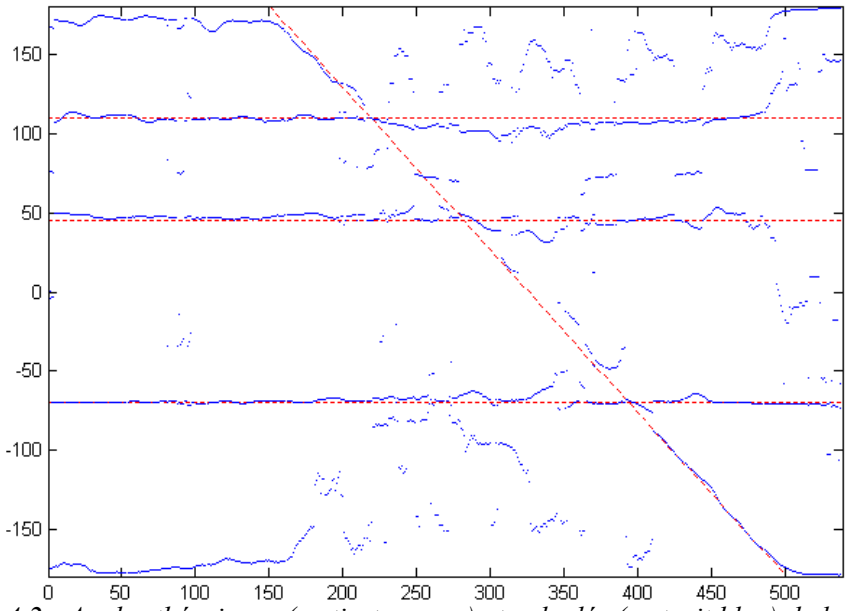

Figure 4.2 - Angles théoriques (en tirets rouge) et calculés (en trait bleu) de la scène à 4 sources dont 1 mobile décrite en 3.3.2.

Cette scène est décrite en 3.3.2.

On a ici demandé aux k-moyennes de renvoyer 5 barycentres par trame alors que seules 4 sources étaient présentes, dans le but de se garantir une marge vis-à-vis des erreurs de détection. Cette 5ème direction n'est pas du tout gênante : elle indique une direction qui ne contient pas de source, et c'est donc un son très faible qui est ainsi pointé. Après une éventuelle extraction, le signal ne sera donc pas du tout gênant.

On peut remarquer que quand le son mobile se rapproche d'un son fixe, la détection est moins bonne ; il y a une fusion entre les 2 sons. Mais encore une fois, lors de l'extraction, l'isolation d'un des sons englobera l'autre son, ce qui ne posera pas de problème pour les objectifs fixés, aucune source n'étant oubliée. En effet, la séparation de sources n'est pas un objectif ici, mais un moyen pour préciser les positions des sources, ce qui sera fait correctement même si 2 sources sont pointées par 1 seule direction.

4.1.4. Scène 3D artificielle à une source et bruit de fond

La scène dont la position d'une source a été extraite ci-dessous est une scène virtuelle en 3D, constitué d'un son (une voix) qu'on a fait se déplacer par paliers tout d'abord vers le haut de 0° à 45° et vers la gauche de 0° à 90°, puis redescendre jusqu'à 35° en continuant son cheminement vers la gauche jusqu'à 120°, et enfin finir la descente de 35° à 0° en la laissant à 120°. À cette source a été superposée une ambiance spatialisée de hall d'aéroport, initialement enregistrée en 2D (au format 5.1) mais transformée pour contenir des composantes verticales. Cette ambiance, décorrelée de la source, avait pour but de tester la robustesse de l'analyse. Elle était à 15 dB en-dessous du niveau de la voix.

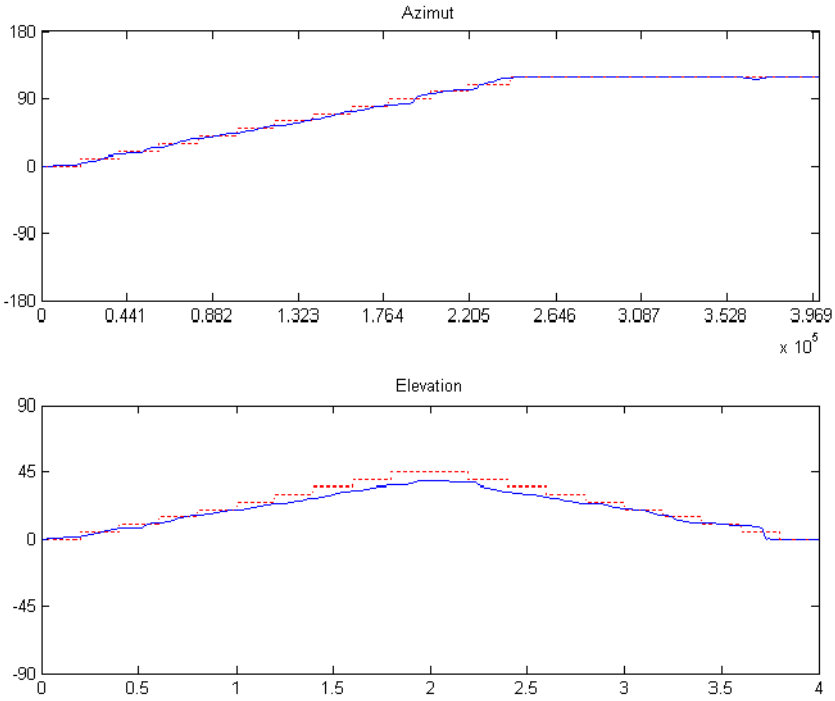

Figure 4.3 - Azimut et élévation théoriques (en tirets rouge) et calculés (en trait bleu) d'un son en 3D artificiel de 9s placé dans une ambiance bruyante.

On remarque que l'analyse est très correcte. Il n'y a pas d'écart important entre les angles théoriques et ceux calculés. On note toutefois que le lissage à l'entrée des k-moyennes a pour effet de gommer les créneaux et d'aplatir quelque peu la courbe d'élévation ; ceci n'est pas vraiment gênant car de tels sauts dans la position d'une source ne sont pas du tout réalistes (ils peuvent caractériser des bruits de pas, mais de tels impacts sont courts, et on peut s'attendre à ce que la courbe suive leur position aux instants où ils sont émis).

4.1.5. Scène 2D naturelle à une source

Les scènes naturelles sont on l'a vu plus délicates à exploiter car on ne peut pas garantir la précision de la position théorique à retrouver par le calcul, et on ne peut donc pas calculer la précision de l'analyse. Seuls les enregistrements utilisant la table tournante (voir 2.3.2) permettent une mesure précise de la position de l'angle. Ces enregistrements on pu confirmer la validité de l'analyse pour des scènes réelles ; on trouve ici un écart-type de 2,5°.

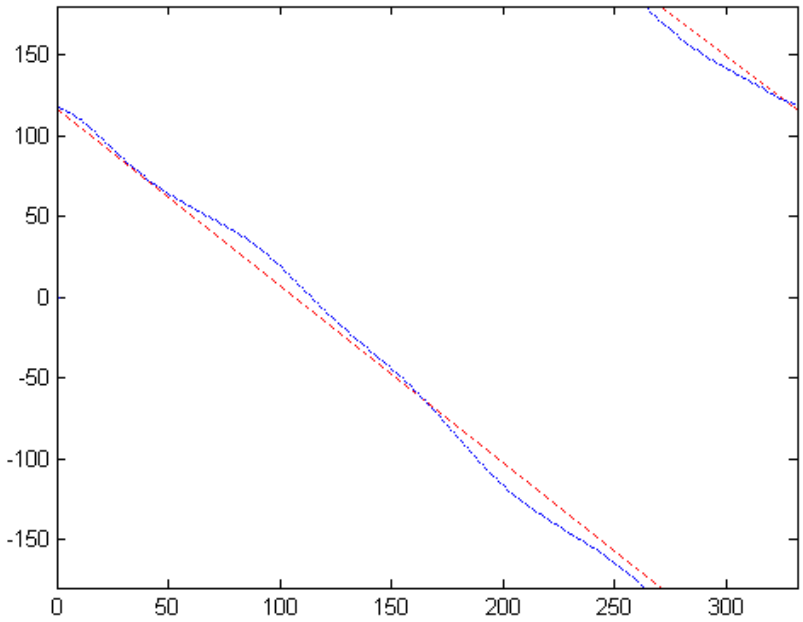

Figure 4.4 - Angles théoriques (en tirets rouge) et calculés (en trait bleu) d'un bruit blanc tournant autour du micro, enregistré en chambre sourde

Cette analyse a été réalisée par sous-bandes mais sur une étendue fréquentielle plus réduite, puisqu'on s'est limité aux possibilités offertes par le microphone. La scène a donc été analysée de 200 à 10 000 Hz.

4.2. Quelques applications

4.2.1. Amélioration de la précision spatiale : synthèse de l'ordre 2 à partir de l'ordre 1

Une fois les positions des sources connues, la porte est ouverte à de nombreuses applications. La connaissance de ces positions permet en effet de synthétiser artificiellement des composantes manquantes ou non transmises. La première synthèse développée est celle des signaux U et V composant la partie horizontale de l'ordre 2 à partir de l'ordre 1 et de paramètres issus de l'analyse pour une source en 2D (des angles dans de cas)

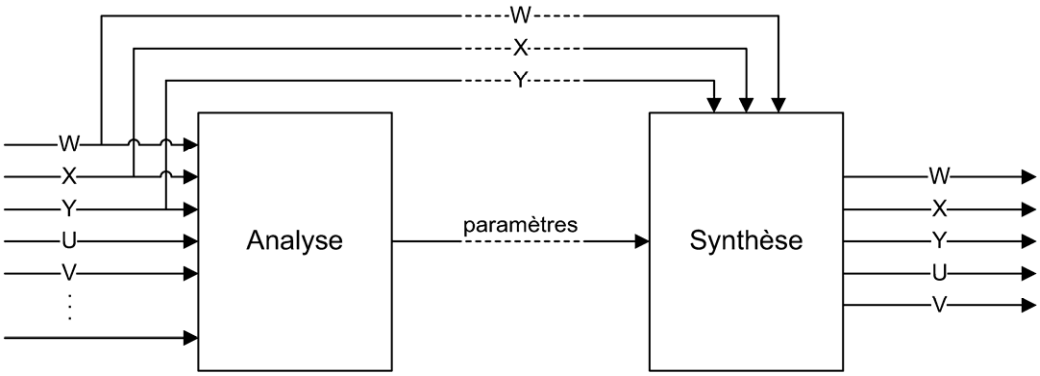

Figure 4.5 - Principe de l'amélioration de la précision spatiale

Le principe de la synthèse est d'extraire les sources originales à partir de W, X, Y et des angles (un angle par trame et par sous-bande). Ceci peut être réalisé de manière approchée en décodant l'ordre 1 dans les directions des sources. Ce décodage revient à calculer les signaux qui devraient alimenter des haut-parleurs qui seraient placés à ces positions dans le cas d'une restitution (voir 1.2.2), c'est-à-dire à former des directivités dans les directions des sources. Dans la plupart des cas, ce décodage, qui fait intervenir une pseudo-inversion de Moore-Penrose (1.11), est imparfait, les sources se retrouvant légèrement mélangées. On peut comparer cette séparation à celle que ferait un ensemble de microphones à directivité hypercardioïde coïncidents.

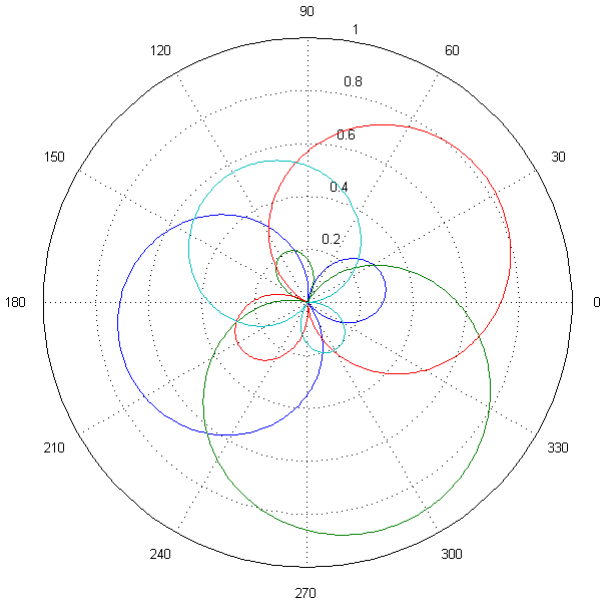

Figure 4.6 - Directivités équivalentes à la séparation de 4 sources à l'ordre 1.

Ces sources séparées vont ensuite constituer, toujours placées aux angles déterminés à l'analyse, une scène virtuelle, qu'on va encoder à l'ordre 2. De cet encodage, on ne conservera que les signaux manquants, dans notre exemple les signaux du 2nd ordre U et V. Ainsi, la scène spatialisée sera au final constituée des signaux W, X et Y d'origine (transmis) et des signaux U et V synthétisés.

4.2.2. Restitution depuis un downmix mono : synthèse de l'ordre 1 à partir de W

Reconstruire une scène sonore spatialisée à partir d'un unique canal mono est alors difficile et le résultat n'est appréciable que pour des scènes simples. Il faut pour cela que les différentes sources composant cette scène aient des supports spectraux aussi distincts que possible. Les différentes sous-bandes sont alors placées à un certain angle, transmis en parallèle, et c'est cette scène reconstituée qui va être encodée au 1er ordre de manière à recréer les signaux X et Y.

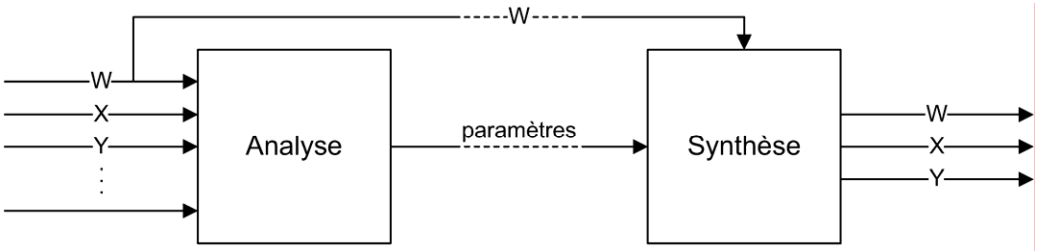

Figure 4.7 - Principe de la restitution de l'ordre 1 à partir du signal mono W

Ce schéma rappelle une méthode développée dans le cas stéréo appelée BCC (Binaural Cue Coding) [Faller and Baumgarte, 2002]. Le principe est d'extraire et coder des informations permettant la reconstruction binaurale d'un signal, à savoir les différences interaurales d'intensité et de temps (Interaural Level/Time Difference, ILD et ITD), qui correspondent dans le cas d'une restitution sur un casque aux différences d'intensité et de temps inter-canal (Inter-Channel Level/Time Difference, ICLD et ICTD) qui sont mesurées à partir d'une analyse en sous-bandes. Les auteurs y ont plus récemment associé un indice de corrélation interaurale, mesuré à partir de la corrélation inter-canal (ICC), afin de caractériser la précision spatiale (la "largeur" d'une source)

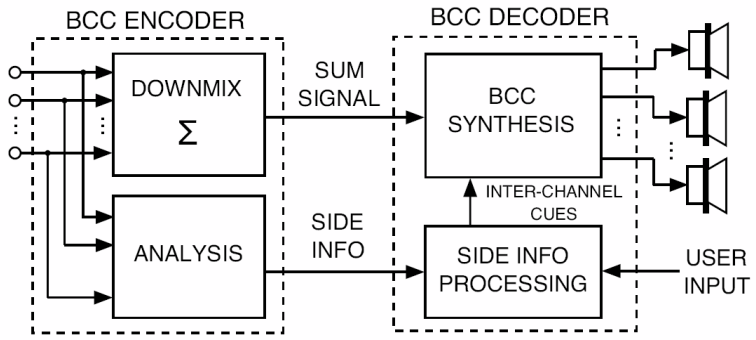

Figure 4.8 - Schéma de principe de la méthode BCC [Faller and Baumgarte, 2003]

On observe en pratique des défauts liés à la brutalité de la procédure : une sous-bande ne peut être que placée à la position indiquée, et ne peut pas rester floue, diffuse ou encore être placée à plusieurs endroits. Pour y remédier, nous avons ajouté 2 traitements à la synthèse. Tour d'abord, nous avons appliqué un léger lissage soulageant les effets désagréables dus au saut du contenu d'une sous-bande d'un endroit à un autre entre de 2 fenêtres consécutives. Les positions varient alors de façon plus souple au sein d'une même sous-bande, traduisant une hypothèse de continuité spatiale (pas de déplacement brusque). La contrepartie est que cette hypothèse est liée à une hypothèse de continuité spectrale.

La 2ème amélioration concerne le positionnement des sources. En incluant dans la transmission un critère de pertinence, on peut savoir s'il est raisonnable de placer une sous-bande à un angle précis ou bien si l'analyse était flou à son sujet, auquel cas un positionnement imprécis serait préférable. Le paramètre choisi est une combinaison des critères évoqués en 3.4.2 ; il nous permet de placer la sous-bande de manière plus ou moins prononcée.

4.3. Comparaison avec une méthode existante

Certains auteurs ont déjà essayé d'analyser un contenu ambisonique pour retrouver les positions des sources. Ville Pulkki a ainsi décrit dans [Pulkki and Faller, 2006] une méthode d'analyse du B-format, basée sur l'étude des amplitudes des signaux W, X, Y et Z.

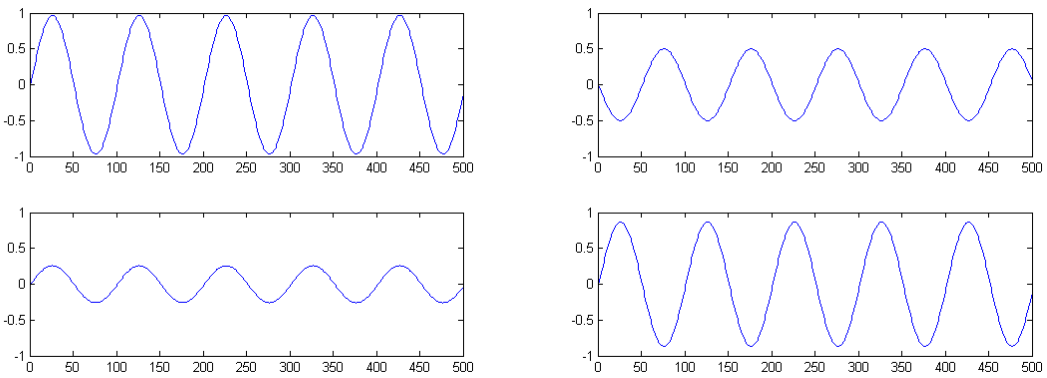

Figure 4.9 - X et Y, dans le cas d'une sinusoïde placée à 15° (à gauche) et 120° (à droite)

Les figures ci-dessus montrent les relations d'amplitudes (algébriques) qui relient X et Y à l'angle de la source. On voit que l'analyse de ces amplitudes respectives permet de retrouver la position de la source : en effet, on a pour une source unique les relations de l'équation (1.1) rappelée ici :

$$\begin{cases} W = S \\ X = S\sqrt{2}\cos\theta\cos\delta \\ Y = S\sqrt{2}\sin\theta\cos\delta \\ Z = S\sqrt{2}\sin\delta \end{cases} \tag{4.1}$$

On peut donc théoriquement retrouver les valeurs de θ et δ par relations trigonométriques. Ainsi, on a :

$$\begin{cases} \tan(\theta) = \dfrac{Y}{X} \\ \sin(\delta) = \dfrac{Z}{W\sqrt{2}} \end{cases} \tag{4.2}$$

On peut donc trouver les angles θ et δ à π près. Pour lever l'ambiguïté, une étude de signe est nécessaire. Mais dans le cas de θ, le signe de X et de Y, à δ fixé, dépend aussi bien de θ que *S*. Il faut s'intéresser aux signes de $\cos\theta$ et $\sin\theta$, c'est-à-dire aux signes de $\frac{X}{W}$ et $\frac{Y}{W}$. Dans la pratique, on s'intéressera aux signes de $W \cdot X$ et $W \cdot Y$ pour se prévenir des divisions par zéros. $W \cdot X$ et $W \cdot Y$ sont proportionnels à S^2, toujours positif.

Les auteurs décrivent également dans leur article un critère appelé *diffuseness* qui peut être comparé à un critère de pertinence inversé. Ce paramètre, calculé par à 1 - *R* où *R* est un rapport de normes, est très proche du critère de norme relative décrit en 3.4.2. Il peut donc être utilisé de la même façon et permettre ainsi une pleine comparaison.

Cette méthode a été implémentée pour pouvoir être comparée avec la méthode par ACP. Tous les paramètres ont scrupuleusement été gardés identiques, et plusieurs réglages ont été testés. Les courbes ci-dessous présentent les résultats des analyses de la scène décrite en 3.3.2 avec un pas de 50ms, un overlap de 50%, un NFFT de 2048 et 40 bandes fréquentielles. Nous avons représenté tous les angles trouvés, c'est-à-dire un par bande et par trame, pour s'affranchir de la sélection des k-moyennes.

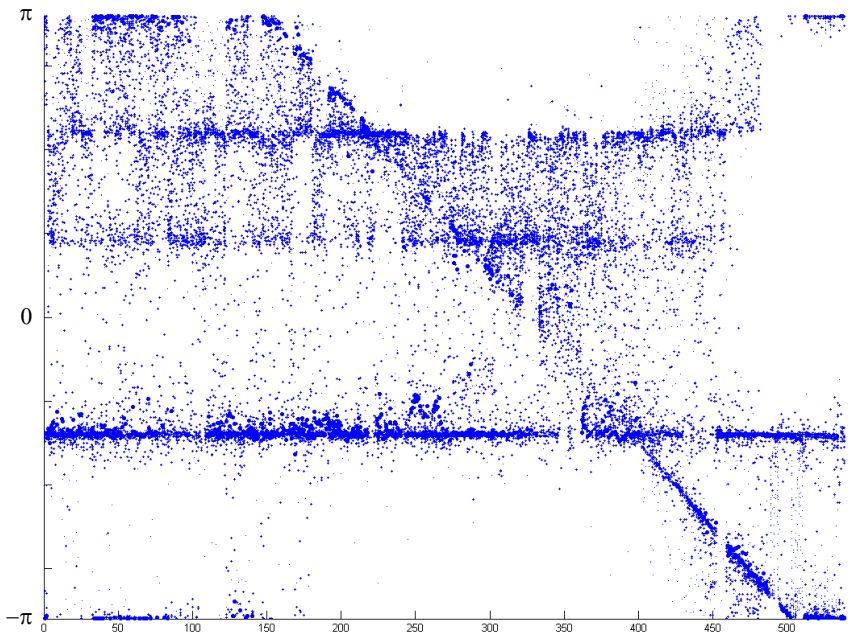

Figure 4.10 - Angles trouvés par l'algorithme de [Pulkki and Faller, 2006]. Le diamètre du point correspondant à la pertinence.

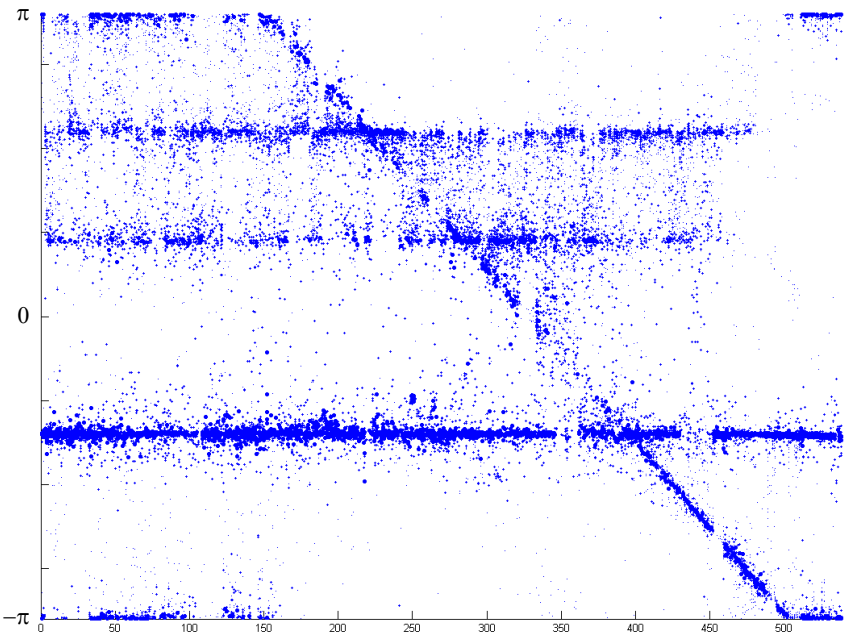

Figure 4.11 - Angles trouvés par une ACP pour la même scène. On remarque que les points excentriques sont moins nombreux et plus rarement trouvés comme pertinents.

4.4. Développement d'une application pour la conférence audio spatialisée

4.4.1. Position du problème

Une application directe de ces analyses a été développée pour la conférence audio spatialisée. La conférence audio utilise en effet la spatialisation pour permettre une meilleure séparation par l'oreille des différents locuteurs. Ainsi, si plusieurs personnes parlent en même temps, il sera très difficile de comprendre chacun individuellement si ces voix sont restituées en mono, mais on pourra focaliser l'attention sur chacune d'entre elles si elles sont séparées spatialement (c'est l'effet *cocktail party* [Cherry, 1953]).

L'objectif était de vérifier qu'il était possible de conserver la spatialisation d'une conférence en n'ayant transmis qu'un downmix mono des signaux. Ce concept semble se rapprocher de celui décrit en 4.2.2, mais une différence fondamentale les sépare : la conférence audio spatialisée n'est pas codée au format ambisonic, mais les signaux traités sont directement les sons provenant des différents locuteurs, c'est-à-dire les sources elles-mêmes. Ainsi, l'analyse ne porte plus sur une scène spatialisée, mais sur des signaux séparés. Il n'y a ainsi plus d'indices spatiaux à trouver.

Avoir accès aux sources permet d'obtenir des informations plus précises pour permettre une séparation après la transmission. Comme nous l'avons vu, l'analyse d'une scène spatialisée ne permet que dans d'ultimes cas à la séparation des sources, ce qui est ici notre point de départ, et sera en fait notre point d'arrivée. En effet, nous tenterons de retrouver les sources séparées après la transmission, de manière à pouvoir par la suite les spatialiser à loisir (voir Figure 4.13).

4.4.2. Solution proposée

Les paramètres extraits ne sont donc plus des positions de sources. Nous avons choisi de mesurer puis transmettre l'énergie de chaque signal dans chaque sous-bande en parallèle du downmix. En effet, une bonne séparation sera possible si l'on sait quelle proportion de chaque source compose chaque sous-bande du signal mono.

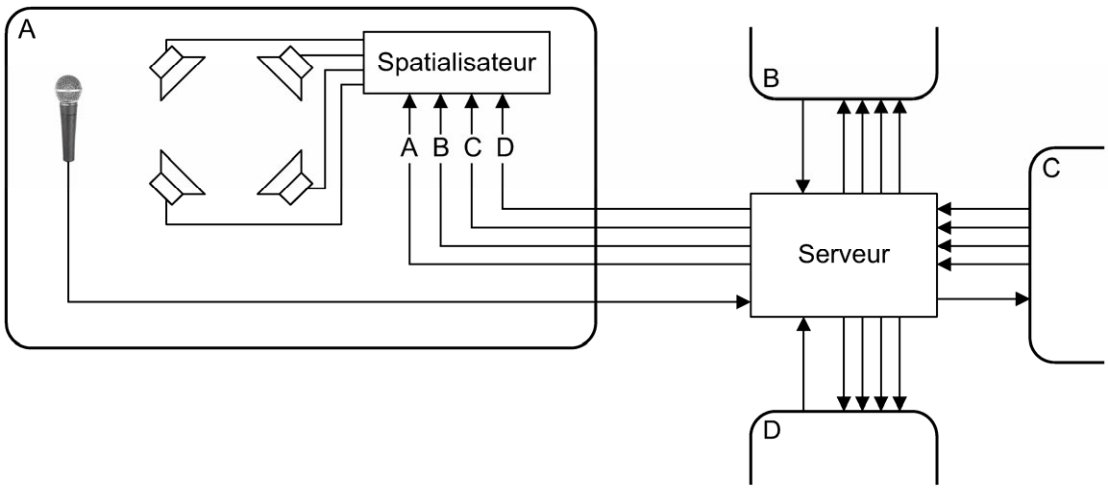

Figure 4.12 - Schéma de principe d'une conférence audio spatialisée

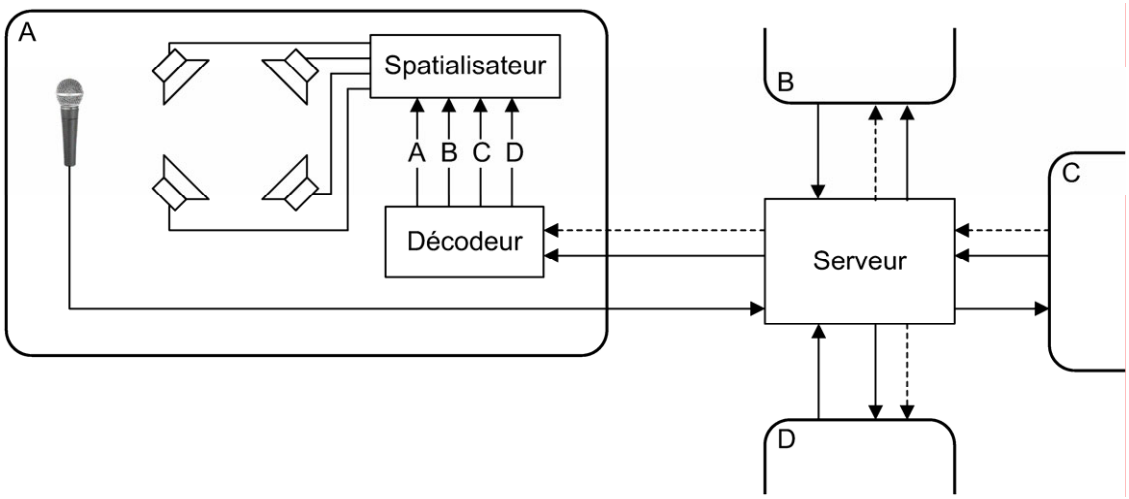

Figure 4.13 - Schéma amélioré (débit réduit)
en pointillés : informations complémentaires (énergie)

L'analyse en sous-bandes a lieu dans le serveur (Figure 4.14). Celui-ci calcule l'énergie de chaque signal dans chaque sous-bande sur une trame donnée. C'est ce paramètre, qui associé au signal mono M, permettra la séparation des sources dans le terminal (Figure 4.15). Ce dernier répartira en effet les différentes sous-bandes vers chaque signal reconstitué proportionnellement à cette information pour reconstruire les 4 sources originales.

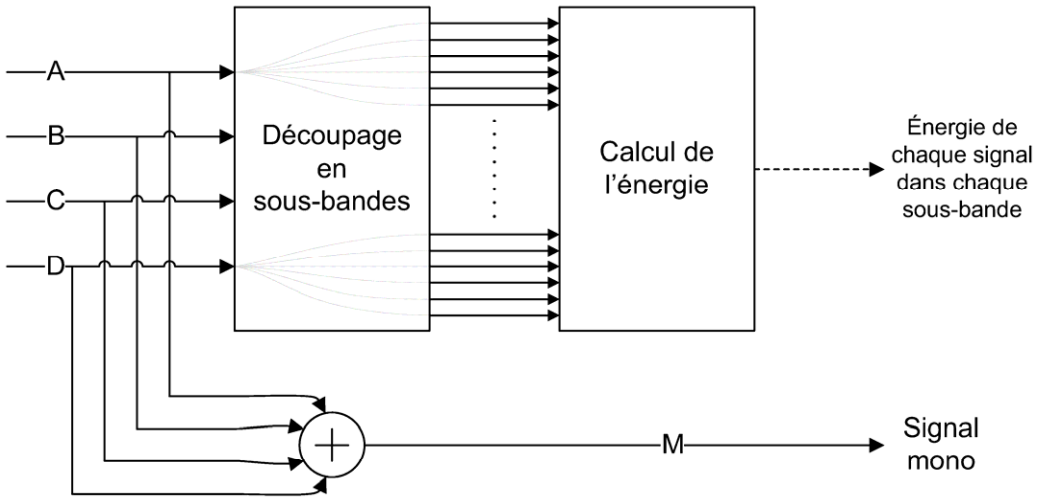

Figure 4.14 - Analyse : principe de l'encodage dans le serveur

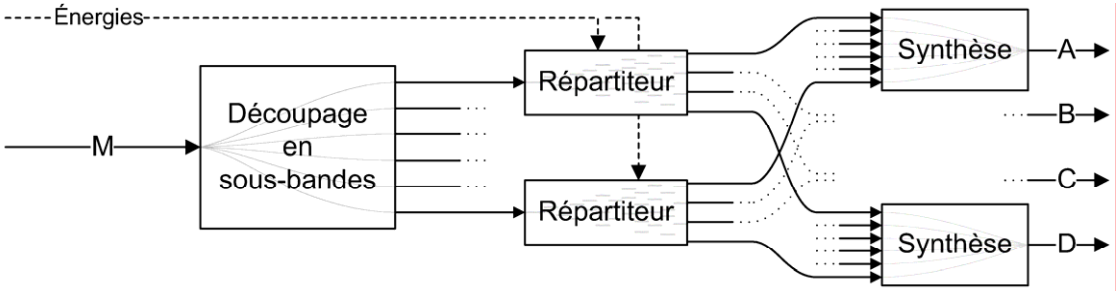

Figure 4.15 - Synthèse : principe du décodage dans le client

Afin de réduire au maximum les informations complémentaires à transmettre, on peut remarquer qu'au cours d'une conférence audio, on a très souvent un seul locuteur, les participants ne parlent que très rarement tous en même temps. Ceci implique que les signaux en entrée sont très souvent tous nuls sauf 1, et appliquer un seuil en entrée de l'analyse permet, outre l'allègement du traitement, d'avoir moins de paramètres à transmettre.

Ce schéma simplifié ne tient pas compte des problèmes de phase : rien n'indique que les énergies des signaux s'ajoutent bien dans chaque sous-bande lors du downmix ; on peut très bien avoir des oppositions de phase. La solution est d'effectuer le downmix (la somme) des signaux dans chaque sous-bande, en prenant soin d'avoir recalé les phases au préalable. Ceci n'a un sens que pour des bandes suffisamment étroites. Les décalages temporels appliqués pour le recalage sont alors des paramètres supplémentaires à transmettre, et sont pris en compte lors de la synthèse au niveau du répartiteur, qui repositionnera la phase de chaque trame pour chaque signal.

4.4.3. Bilan

Le résultat du décodage a pu être écouté de différentes manières : les sources reconstituées individuellement, ensemble en stéréo d'amplitude, et sur un système 5.1. On note que les sources ont été assez bien séparées, mais que des artefacts sont audibles en particulier lorsque le nombre de locuteurs simultanés augmentent. Ces artefacts sont un mélange des voix qui n'est gênant que lorsqu'on les écoute de manière séparée. En effet, l'écoute simultanée de ces signaux synthétisés, même placés à des positions distinctes sur un système de diffusion multicanal, est tout à fait satisfaisante.

Ce procédé, pour pouvoir être appliqué, doit fonctionner en temps-réel. Cette contrainte n'a pas du tout été prise en compte, l'objectif étant de valider la faisabilité d'une reconstruction. Les programmes, sous Matlab, n'ont pas été optimisés et ne respectent pas cette contrainte.

Conclusion

Nous avons présenté dans ce rapport le principe et l'application de plusieurs méthodes d'analyses spatiales de scènes ambisoniques conçues et réalisée pendant le stage. Le format HOA est en effet une représentation du champ acoustique décomposé sur une base d'harmoniques sphériques jusqu'à un certain ordre, à la manière d'un développement limité. Il contient donc de manière indirecte des indications spatiales, qu'on peut extraire avec des analyses comme l'analyse en composantes principales.

Ces analyses nous permettent de déterminer les positions des sources qui composent la scène sonore. Nous avons appliqué les résultats de ces analyses à 2 types de synthèse, l'une destinée aux transmissions bas débit, l'autre à un traitement de restitution de la précision spatiale dans le cas où des ordres supérieurs, dont le but est d'affiner la perception de localisation, n'auraient pas pu être transmis. Ces algorithmes, par le codage d'informations spatiales qu'ils permettent, pourront trouver bien d'autres applications dans la transmission et le codage, où la description du contenu est de plus en plus utilisée (comme par exemple dans la norme MPEG).

Des améliorations sont possibles sur plusieurs points : tout d'abord, l'exploitation de toutes les composantes de l'ACP est à l'étude et semble pouvoir donner des résultats intéressants. D'autre part, on a vu que le résultat final est très dépendant de la bonne convergence de l'algorithme des k-moyennes, utilisé pour traiter les données brutes de l'analyse en sous-bandes. Or cet algorithme nécessite des réglages très fins, et une connaissance *a priori* du nombre de sources est presque indispensable. D'autres algorithmes plus complexes de suivi de formes pourraient être avantageusement utilisés pour détecter dans tous les angles trouvés lesquels sont susceptibles de former une trajectoire.

Annexe A

Développement du champ acoustique en harmoniques sphériques

A.1. Équation d'onde en coordonnées sphériques

On se place dans l'espace euclidien à 3 dimensions muni d'un système de coordonnées sphériques. Un point *M* y est repéré par son rayon *r*, son azimut θ et son élévation δ tels que représentés sur la figure ci-dessous :

- *r* est la norme du vecteur position $\vec{r}$, c'est-à-dire la distance *OM*
- $\theta \in [0;2\pi[$ est l'angle formé par le projeté du point sur le plan *xOy* et l'axe *Ox*.
- $\delta \in [-\frac{\pi}{2};\frac{\pi}{2}]$ est l'angle formé par le plan *xOy* et le vecteur position[2].

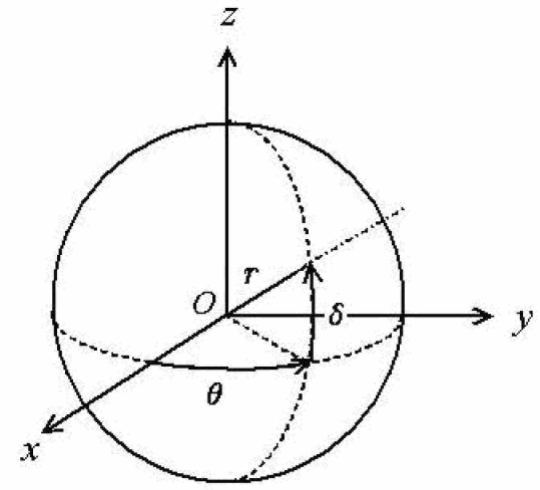

Figure A.1 - Système de coordonnées sphériques

[2] Cet angle diffère de l'angle polaire usuel $\varphi = \frac{\pi}{2} - \delta$, habituellement pris entre l'axe *Oz* et le vecteur position.

Les relations entre les coordonnées sphériques et les coordonnées cartésiennes sont :

$$\begin{cases} x = r\cos\theta\cos\delta \\ y = r\sin\theta\cos\delta \\ z = \sin\delta \end{cases} \quad \text{(A.1)}$$

On s'intéresse dans ce repère à l'équation de propagation des ondes acoustiques de pression. On se place dans les conditions classiques d'acoustique linéaire (les variations de *p* sont petites par rapport à la pression atmosphérique), en supposant que l'air est un fluide parfait (on y négligera les phénomènes dissipatifs). Dans une région de l'espace ne comprenant pas de sources, l'équation d'onde s'écrit alors :

$$\left(\Delta - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right) p(\vec{r},t) = 0 \quad \text{(A.2)}$$

p étant la pression et *c* la vitesse de propagation du son dans l'air.

Cette équation devient, dans le domaine fréquentiel, l'équation de Helmholtz :

$$\left(\Delta + k^2\right) p(\vec{r},\omega) = 0 \quad \text{(A.3)}$$

$k = \frac{\omega}{c}$ étant le nombre d'onde.

En coordonnées sphériques, cette équation s'écrit :

$$\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{r^2\sin\theta}\frac{\partial^2}{\partial\delta^2} + k^2\right] p = 0 \quad \text{(A.4)}$$

A.2. Solutions de l'équation d'onde en coordonnées sphériques

D'après [Bruneau, 1983], elle admet des solutions à variables séparées de la forme

$$p = R(r)\Theta(\theta)\Phi(\delta) \quad \text{(A.5)}$$

Ces solutions sont respectivement :

- pour la partie radiale : une combinaison des fonctions de Bessel sphériques de première et seconde espèce[3] j_n et n_n

$$R(r) = Aj_n(kr) + Bn_n(kr) \quad \text{(A.6)}$$

[3] Les fonctions de Bessel sphériques de seconde espèce sont également appelées fonctions de Neumann

Ces fonctions sont définies comme suit :

$$\begin{cases} j_n(x) &= \sqrt{\frac{\pi}{2x}} J_{n+1/2}(x) = (-1)^n x^n \left(\frac{d}{x\,dx}\right)^n \frac{\sin x}{x} \\ j_0(x) &= \frac{\sin x}{x} \\ j_1(x) &= \frac{\sin x}{x^2} - \frac{\cos x}{x} \\ j_2(x) &= (\frac{3}{x^3} - \frac{1}{x})\sin x - \frac{3}{x^2}\cos x \\ n_n(x) &= (-1)^{n+1} \sqrt{\frac{\pi}{2x}} J_{-n-1/2}(x) = (-1)^{n+1} x^n \left(\frac{d}{x\,dx}\right)^n \frac{\cos x}{x} \\ n_0(x) &= -\frac{\cos x}{x} \\ n_1(x) &= -\frac{\cos x}{x^2} - \frac{\sin x}{x} \\ n_2(x) &= -(\frac{3}{x^3} - \frac{1}{x})\cos x - \frac{3}{x^2}\sin x \end{cases} \tag{A.7}$$

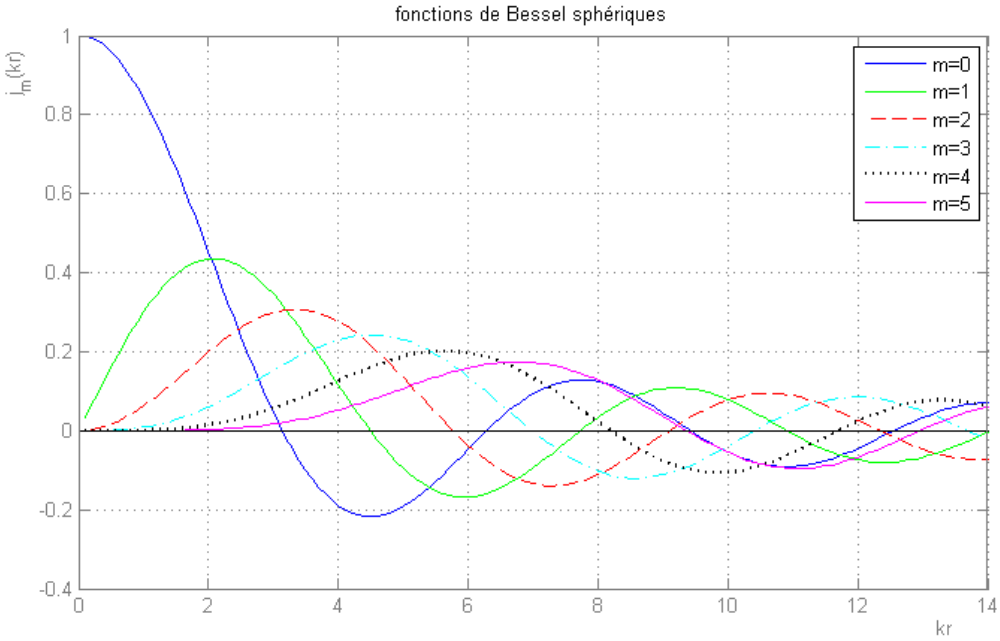

Figure A.2 - Fonctions de Bessel sphériques $j_m(kr)$ pour les ordres 0 à 5

- pour la partie azimutale : les fonctions $\Theta(\theta)$

$$\Theta(\theta) = c_1 \cos(n\theta) + c_2 \sin(n\theta) \tag{A.8}$$

Étant donné la définition de l'angle Θ, celui-ci doit nécessairement être de période 2π $(\Theta(\theta) = \Theta(\theta + 2\pi))$, ce qui impose à n d'être entier.

- pour la partie polaire : les fonctions de Legendre associées $P_{mn}(\sin\delta)$

$$P_{mn}(\sin\delta) = \sin^n\delta \frac{d^n P_m(\sin\delta)}{d(\sin\delta)^n}, \qquad \text{pour} \begin{cases} m,n = 1,2,3... \\ avec\ m > n \end{cases} \tag{A.9}$$

où les $P_n(\sin\delta)$ sont les polynômes de Legendre de degré n en $\sin\delta$, définis par récurrence :

$$\begin{aligned} &P_0(\sin\delta) = 0 \\ &P_1(\sin\delta) = \sin\delta \\ &P_2(\sin\delta) = \frac{1}{2}\left(3\cos^2\varphi - 1\right) \\ &et \quad (m+1)P_{m+1}(\sin\delta) = (2m+1)\sin\delta\, P_m - mP_{m-1} \end{aligned} \tag{A.10}$$

et de formule générale :

$$P_m(\sin\delta) = \frac{1}{2^m m!}\frac{d^m}{d(\sin\delta)^m}(\sin^2\delta - 1)^m \tag{A.11}$$

On peut appliquer aux fonctions $P_{mn}(\sin\delta)$ la semi-normalisation de Schmidt :

$$\widetilde{P}_{mn}(\sin\delta) = \sqrt{\varepsilon_n \frac{(m-n)!}{(m+n)!}}\, P_{mn}(\sin\delta), \qquad \text{où } \varepsilon_0 = 1 \text{ et } \varepsilon_n = 2 \text{ pour } n \geq 1 \tag{A.12}$$

En appariant les fonctions polaires et azimutales, on définit les harmoniques sphériques $Y_{mn}^{(1)}$ de degré $m \geq 0$ et d'ordre[4] $n, 0 \leq n \leq m$, et $Y_{mn}^{(-1)}$ de degré $m \geq 1$ et d'ordre $n, 1 \leq n \leq m$ (voir Figure A.3) :

$$\begin{aligned} Y_{mn}^{(1)}(\theta,\delta) &= \cos(n\theta)\widetilde{P}_{mn}(\sin\delta) \\ Y_{mn}^{(-1)}(\theta,\delta) &= \sin(n\theta)\widetilde{P}_{mn}(\sin\delta) \end{aligned} \tag{A.13}$$

[4] En général, lorsque le degré et l'ordre ne sont pas évoqués conjointement, on parle par abus de langage d'harmoniques d'ordre *m* plutôt que de degré *m*

soit

$$Y_{m1}^{(1)}(\theta,\delta) = \cos\theta\sqrt{(2m+1)\frac{(m-1)!}{(m+1)!}}P_{m1}(\sin\delta) \quad pour\ n=1$$

$$\begin{cases} Y_{mn}^{(1)}(\theta,\delta) = \cos n\theta\sqrt{2(2m+1)\dfrac{(m-n)!}{(m+n)!}}P_{mn}(\sin\delta) \\ Y_{mn}^{(-1)}(\theta,\delta) = \sin n\theta\sqrt{2(2m+1)\dfrac{(m-n)!}{(m+n)!}}P_{mn}(\sin\delta) \end{cases} \quad pour\ n>1 \tag{A.14}$$

La solution angulaire générale s'écrit alors sous la forme d'un développement sur la base d'harmoniques sphériques :

$$\left(\Theta(\theta)\right)_n = \sum_{m=1}^{n-1}\left(B_{mn}Y_{mn}^{(1)} + B_{mn}Y_{mn}^{(-1)}\right) \tag{A.15}$$

La solution générale est alors la série de Fourier-Bessel :

$$p(kr,\theta,\delta) = \sum_{m=0}^{+\infty} i^m j_m(kr)\sum_{n=0}^{m}\sum_{\sigma=\pm1} B_{mn}^{(\sigma)}Y_{mn}^{(\sigma)}(\theta,\delta) \tag{A.16}$$

où

$$B_{mn}^{(\pm1)} = \frac{1}{i^m j_m(kr)} \iint_{S\left\{\begin{array}{l}\theta\in[0,2\pi[\\ \delta\in[-\frac{\pi}{2},\frac{\pi}{2}]\end{array}\right.} p(1,\theta,\delta)Y_{mn}^{(\pm1)}(\theta,\delta)dS \tag{A.17}$$

		Normalisation $\sqrt{\varepsilon_0 (2m+1)\frac{(m-n)!}{(m+n)!}}$		Fonctions de Legendre associées $P_{mn}(\sin\delta)$	
Ordre m = 0 (1 composante)	$Y_{00}^{1}(\theta,\delta) =$	1	$\times$	1	
Ordre m = 1 (3 composantes)	$Y_{11}^{1}(\theta,\delta) =$	$\sqrt{3}$	$\times$	$\cos\delta$	$\times \cos\theta$
	$Y_{11}^{-1}(\theta,\delta) =$	$\sqrt{3}$	$\times$	$\cos\delta$	$\times \sin\theta$
	$Y_{10}^{1}(\theta,\delta) =$	$\sqrt{3}$	$\times$	$\sin\delta$	
Ordre m = 2 (5 composantes)	$Y_{22}^{1}(\theta,\delta) =$	$\sqrt{\frac{5}{12}}$	$\times$	$3\cos^2\delta$	$\times \cos(2\theta)$
	$Y_{22}^{-1}(\theta,\delta) =$	$\sqrt{\frac{5}{12}}$	$\times$	$3\cos^2\delta$	$\times \sin(2\theta)$
	$Y_{21}^{1}(\theta,\delta) =$	$\sqrt{\frac{5}{3}}$	$\times$	$3\cos\delta\sin\delta$	$\times \cos\theta$
	$Y_{21}^{-1}(\theta,\delta) =$	$\sqrt{\frac{5}{3}}$	$\times$	$3\cos\delta\sin\delta$	$\times \sin\theta$
	$Y_{20}^{1}(\theta,\delta) =$	$\sqrt{5}$	$\times$	$\left(\frac{3\sin^2\delta - 1}{2}\right)$	
Ordre m = 3 (7 composantes)	$Y_{33}^{1}(\theta,\delta) =$	$\sqrt{\frac{7}{360}}$	$\times$	$15\cos^3\delta$	$\times \cos(3\theta)$
	$Y_{33}^{-1}(\theta,\delta) =$	$\sqrt{\frac{7}{360}}$	$\times$	$15\cos^3\delta$	$\times \sin(3\theta)$
	$Y_{32}^{1}(\theta,\delta) =$	$\sqrt{\frac{7}{60}}$	$\times$	$15\cos^2\delta\sin\delta$	$\times \cos(2\theta)$
	$Y_{32}^{-1}(\theta,\delta) =$	$\sqrt{\frac{7}{60}}$	$\times$	$15\cos^2\delta\sin\delta$	$\times \sin(2\theta)$
	$Y_{31}^{1}(\theta,\delta) =$	$\sqrt{\frac{7}{6}}$	$\times$	$\left(\frac{3\cos\delta\cdot\left(5\sin^2\delta - 1\right)}{2}\right)$	$\times \cos\theta$
	$Y_{31}^{-1}(\theta,\delta) =$	$\sqrt{\frac{7}{6}}$	$\times$	$\left(\frac{3\cos\delta\cdot\left(5\sin^2\delta - 1\right)}{2}\right)$	$\times \sin\theta$
	$Y_{30}^{1}(\theta,\delta) =$	$\sqrt{7}$	$\times$	$\sin\delta\cdot\frac{\left(5\sin^2\delta - 3\right)}{2}$	

Tableau A.1 - Expressions analytiques des harmoniques sphériques pour les ordres 0 à 3

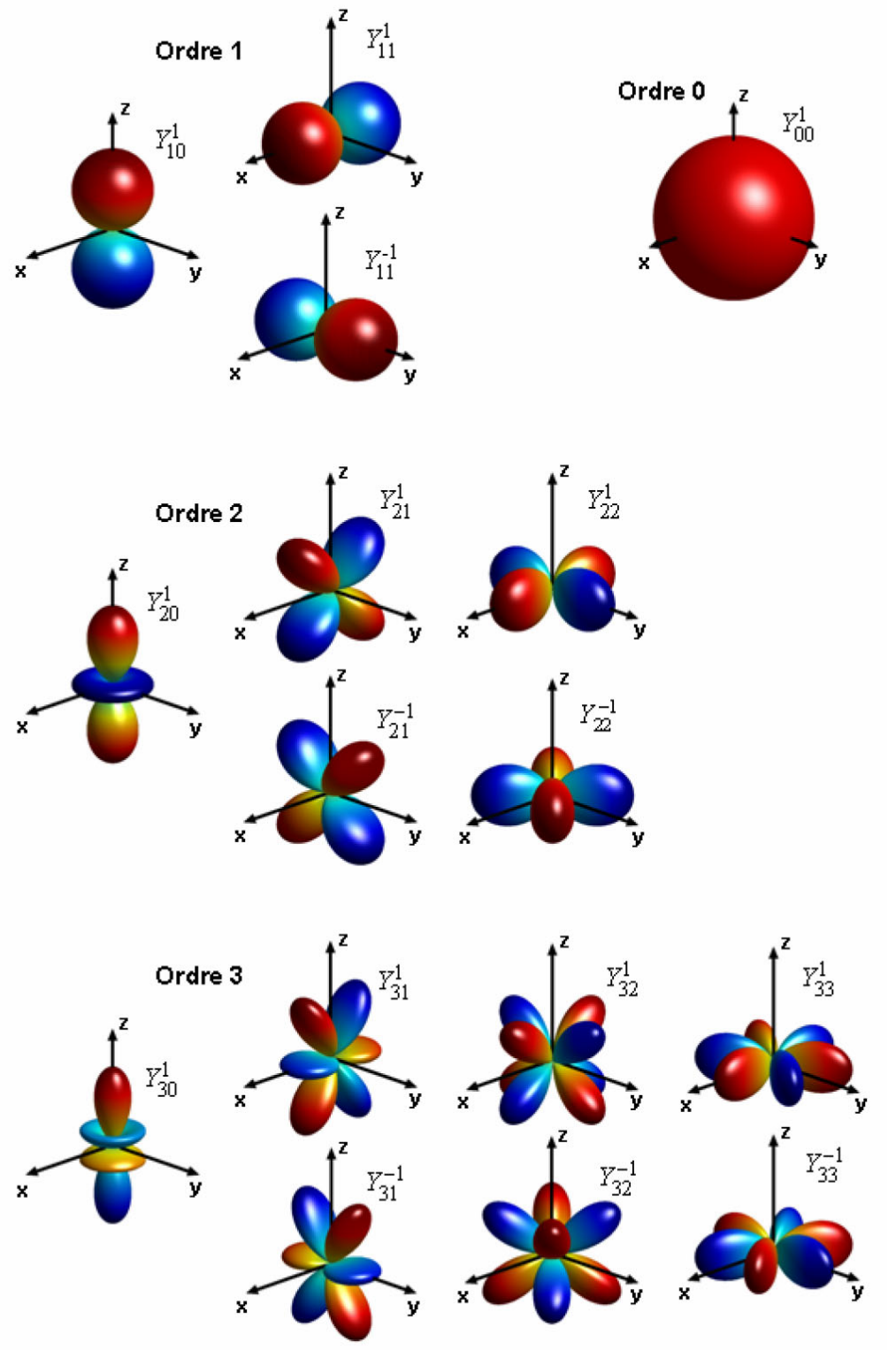

Figure A.3 - Représentation des harmoniques sphériques pour les ordres m = 0 à 3. Les couleurs rouge et bleu indiquent respectivement des valeurs positives et négatives.

Annexe B

Compléments théoriques sur le décodage HOA

B.1. Vecteur vélocité et vecteur énergie

Le vecteur vélocité est un indice de la direction apparente d'une source, introduit par Gerzon [Gerzon, 1992]. Il se calcule comme le rapport de la vitesse particulaire sur la pression acoustique :

$$\vec{V} = -\frac{1}{\rho c}\frac{\vec{v}(r,\theta,\delta)}{p(r,\theta,\delta)} = -\frac{i}{k}\frac{\vec{\nabla}p(r,\theta,\delta)}{p(r,\theta,\delta)} \tag{B.1}$$

Dans le cas d'une onde plane, le vecteur vélocité $\vec{V}$ est ainsi directement un vecteur unitaire orienté vers la direction de provenance de l'onde. En généralisant à une superposition d'ondes planes, en phase ou opposition de phase, émises par L haut-parleurs à équidistance du point d'observation, on obtient la formule suivante :

$$\vec{V} = \frac{\sum_{l=1}^{L} g_l \vec{u}_l}{\sum_{l=1}^{L} g_l} = r_V \vec{u}_V \tag{B.2}$$

où $\vec{u}_l$ désigne la direction du $l^{ème}$ haut-parleur, auquel est associé le gain g_l. r_V et $\vec{u}_V$ correspondent respectivement à la norme et à la direction du vecteur vélocité.

Le vecteur énergie est un autre indice de direction. Il est définit par :

$$\vec{E} = \frac{\sum_{l=1}^{L} |g_l|^2 \vec{u}_l}{\sum_{l=1}^{L} |g_l|^2} = r_E \vec{u}_E \tag{B.3}$$

Il peut s'interpréter comme une indication de la concentration de l'énergie (r_E) et de sa direction moyenne de provenance ($\vec{u}_E$). On a toujours $r_E < 1$ sauf dans le cas où un seul haut-parleur contribue à la reconstruction (dans ce cas $r_E = 1$). r_E peut être considéré comme un indicateur de précision de la localisation.

B.2. Décodages optimisés

La restitution la plus confortable d'un champ sonore veillera donc à maximiser r_E tout en essayant de garder r_V le plus proche possible de 1. L'optimisation du décodage consiste alors à multiplier chaque coefficient de la matrice de décodage D (voir (1.11) par des gains calculés d'après des critères d'optimisation.

Ainsi, le décodage appelé *max-*r_E maximise la norme du vecteur énergie, en sacrifiant celle du vecteur vélocité puisqu'on a dans ce décodage (pour des configurations régulières de haut-parleurs) $r_V = r_E$. Ce décodage est donc avantageux par rapport au décodage basique pour les hautes fréquences ou dans le cas d'un auditoire élargi. Le décodage *in-phase* va encore plus loin, en cherchant à annuler complètement les contributions des haut-parleurs placés à l'opposé de la direction de provenance de l'onde reproduite. Ce critère est utile dans le cas d'une écoute dans une position très excentrée, à laquelle on pourrait percevoir de façon prédominante la contribution en opposition de phase d'un tel haut-parleur.

Ces différents décodages ont chacun des avantages et inconvénients. Ainsi, seul le décodage basique garantit $r_V = 1$, le décodage *max-*r_E permet une précision meilleure notamment en haute fréquence, et le décodage *in-phase* est plus adapté à un auditoire plus large. Jérôme Daniel [Daniel, 2000] propose ainsi de changer le décodage en fonction de la fréquence et de l'étendue de la zone d'écoute désirée.

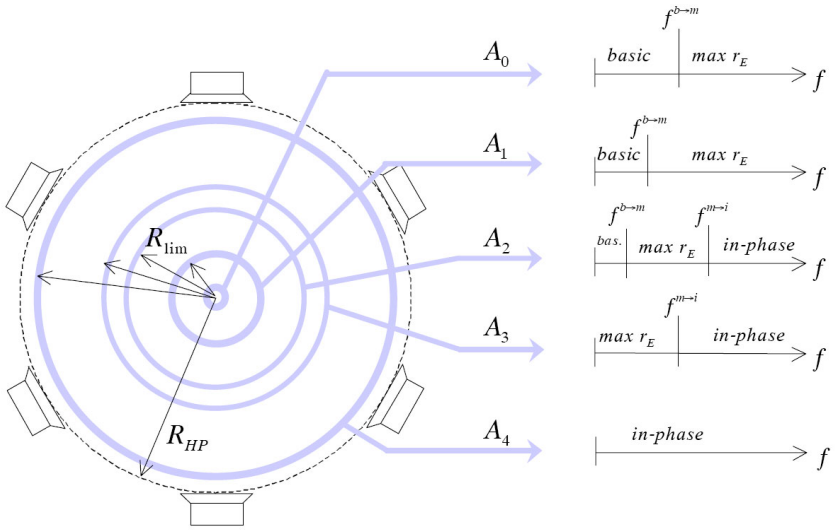

Figure B.1 - Décodages les plus appropriés par sous-bande et en fonction de l'étendue de la zone d'écoute [Daniel, 2000]

Annexe C

Analyse en composantes principales

C.1. Espace des données

Également souvent appelée transformée de Karhunen-Loève[5] (TKL), l'analyse en composantes principales (ACP) est une technique statistique de représentation d'information. Introduite sur 2 variables par Pearson [Pearson, 1900], elle a été étendue par Hotelling [Hotelling, 1933]. Elle a pour but de caractériser des données d'un espace à *N* dimensions dans un sous-espace plus petit en minimisant les pertes d'information due à la projection, c'est-à-dire en maximisant la variance projetée. Les données ainsi réorganisées deviennent indépendantes. Les variables, corrélées, sont remplacées par de nouvelles variables décorrélées et de variance maximale, combinaisons linéaires des variables initiales.

Les données sont constituées de *n* variables aléatoires (dans notre cas des canaux) et *p* réalisations ou individus (des échantillons) :

$$M = \begin{bmatrix} x_1^1 & \cdots & \cdots & \cdots & x_n^1 \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \vdots & \cdots & x_i^j & \cdots & \vdots \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_1^p & \cdots & \cdots & \cdots & x_n^p \end{bmatrix} \qquad \text{(C.1)}$$

Ces données sont généralement centrées et peuvent aussi être réduites[4] (on note $\bar{x}_1$ la moyenne de la variable x_1 et $\sigma(x_1)$ sa variance). Réduire ou non le nuage de points (i.e. les *p* réalisations de la variable aléatoire) est un choix de modèle :

---

[5] Quelques rares sources mentionnent une différence entre ACP et TKL, en plaçant celle-ci au niveau du choix de la réduction des données : dans le cas d'une ACP, les données seraient centrées et réduites, alors qu'elles seraient uniquement centrées pour une TKL.

- si on ne réduit pas le nuage, une variable à forte variance pourrait tirer tout l'effet de l'ACP à elle
- si on réduit le nuage, une variable qui ne serait qu'un bruit se retrouverait avec une variance apparente égale à une variable informative.

$\hat{M}$ est centrée :

$$\hat{M} = \begin{bmatrix} x_1^1 - \bar{x}_1 & \cdots & x_n^1 - \bar{x}_n \\ \vdots & x_i^j - \bar{x}_i & \vdots \\ x_1^p - \bar{x}_1 & \cdots & x_n^p - \bar{x}_n \end{bmatrix} \quad \text{(C.2)}$$

$\tilde{M}$ est centrée et réduite :

$$\tilde{M} = \begin{bmatrix} \frac{x_1^1 - \bar{x}_1}{\sigma(x_1)} & \cdots & \frac{x_n^1 - \bar{x}_n}{\sigma(x_n)} \\ \vdots & \frac{x_i^j - \bar{x}_i}{\sigma(x_i)} & \vdots \\ \frac{x_1^p - \bar{x}_1}{\sigma(x_1)} & \cdots & \frac{x_n^p - \bar{x}_n}{\sigma(x_n)} \end{bmatrix} \quad \text{(C.3)}$$

On multiplie ensuite la matrice obtenue par sa transposée pour obtenir

- La matrice de covariance $C_v = \hat{M}^t \cdot \hat{M}$
- La matrice de corrélation $C_r = \tilde{M}^t \cdot \tilde{M}$

Ces matrices sont carrées (N x N), symétriques et réelles. Elles sont donc diagonalisables dans une base orthonormée.[6]

Le principe de l'ACP est alors de trouver un axe *u* issu d'une combinaison linéaire des x_n tel que le critère d'inertie du nuage autour de cet axe soit maximal. Le critère d'inertie couramment utilisé est la variance. Il faut donc trouver l'axe qui porte le maximum de variance, c'est-à-dire l'axe *u* tel que la projection du nuage sur cet axe ait une variance maximale.

[6] Nos données étant ici des valeurs d'échantillons, elles sont par nature centrées et réduites. On formera donc naturellement la matrice de corrélation.

C.2. Diagonalisation

Pour alléger l'écriture de la suite de cette partie, nous considérerons que les variables x_i ont été centrées et si besoin réduites, et nous noterons x_i ces variables transformées (ainsi, $x_i \leftarrow x_i - \bar{x}_i$ ou bien $x_i \leftarrow \frac{x_i - \bar{x}_i}{\sigma(x_i)}$). On fera de même pour M ($M \leftarrow \hat{M}$ ou bien $M \leftarrow \tilde{M}$) et C ($C \leftarrow C_v$ ou bien $C \leftarrow C_r$).

La projection d'un échantillon x_i sur un axe u s'écrit :

$$\pi_u(x_i) = M \cdot u \tag{C.4}$$

La variance empirique de $\pi_u(x_i)$ vaut donc :

$$\pi_u(x_i)^t \cdot \pi_u(x_i) = u^t \cdot M^t \cdot M \cdot u \tag{C.5}$$

Or on a vu que $C = M^t M$ est diagonalisable dans une base orthonormée, donc on peut noter P le changement de base associé et Δ la matrice diagonale formée de son spectre $(\lambda_i)_1^n$: $C = M^t M = P^t \Delta P$. Ainsi :

$$\pi_u(x_i)^t \cdot \pi_u(x_i) = u^t \cdot P^t \Delta P \cdot u = (Pu)^t \cdot \Delta \cdot (Pu) \tag{C.6}$$

On pose $v = Pu$. Chercher l'axe u qui maximise $\pi_u(x_i)$ revient donc à chercher le vecteur v qui maximise $v^t \Delta v$, où $\Delta = diag(\lambda_1, \ldots, \lambda_n)$ avec les λ_i rangés par ordre décroissant. Ce vecteur est trivialement le premier vecteur unitaire, et on a alors $v^t \Delta v = \lambda_1$.

Ainsi, le vecteur qui porte la plus grande part d'inertie du nuage est le premier vecteur propre de C (premier par norme décroissante). De même, le deuxième vecteur qui explique la plus grande< part de l'inertie restante est le deuxième vecteur propre de C. Finalement, la question de l'ACP se ramène à un problème de diagonalisation de la matrice de corrélation.

C.3. Code Matlab

```
function [vep, vap] = pca(f, N)        % f : données, N : nombre de composantes à garder
    cor = f * f';                      % formation de la matrice de corrélation
    [vep, vap] = eig(cor);             % calcul des vecteurs (vep) et valeurs (vap) propres
    [vap, ind] = sort(diag(vap));      % tri des valeurs propres par ordre croissant
    vap = flipud(vap);                 % disposition inverse, par ordre d'importance
    vep = vep(:, flipud(ind));         % classement en conséquence des vecteurs propres
    vep = vep(:, 1:N);                 % on ne garde que les N premiers vecteurs propres
```

Annexe D

Analyse en sous-bandes

D.1. La transformée de Fourier à court terme (TFCT)

Pour affiner l'analyse, les signaux sont donc découpés en bandes de fréquence par TFCT. La TFCT consiste à effectuer des transformées de Fourier discrètes sur des portions de signal fenêtrées régulièrement espacées.

On rappelle que la transformée de Fourier d'un signal $x(t)$ est définie (lorsque l'intégrale existe) par :

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-2i\pi ft} dt \tag{D.1}$$

La transformée de Fourier à temps discret (TFTD) d'un signal $x(t)$ s'écrit :

$$X(f) = \sum_{n=-\infty}^{+\infty} x(n) e^{-2i\pi fn} \tag{D.2}$$

La transformée de Fourier discrète (TFD) d'un signal discret $x(k)$ est définie par :

$$X(k) = \sum_{n=-\infty}^{+\infty} x(n) e^{-2i\pi \frac{kn}{N}} \tag{D.3}$$

La TFCT discrète est une TFD observée sur une fenêtre w_a de taille N :

$$X(\tau_a, k) = \sum_{n=0}^{N-1} x(n + \tau_a(t)) w_a(n) e^{-2i\pi \frac{nk}{N}} \tag{D.4}$$

où $\tau_a(t)$ est l'instant d'analyse.

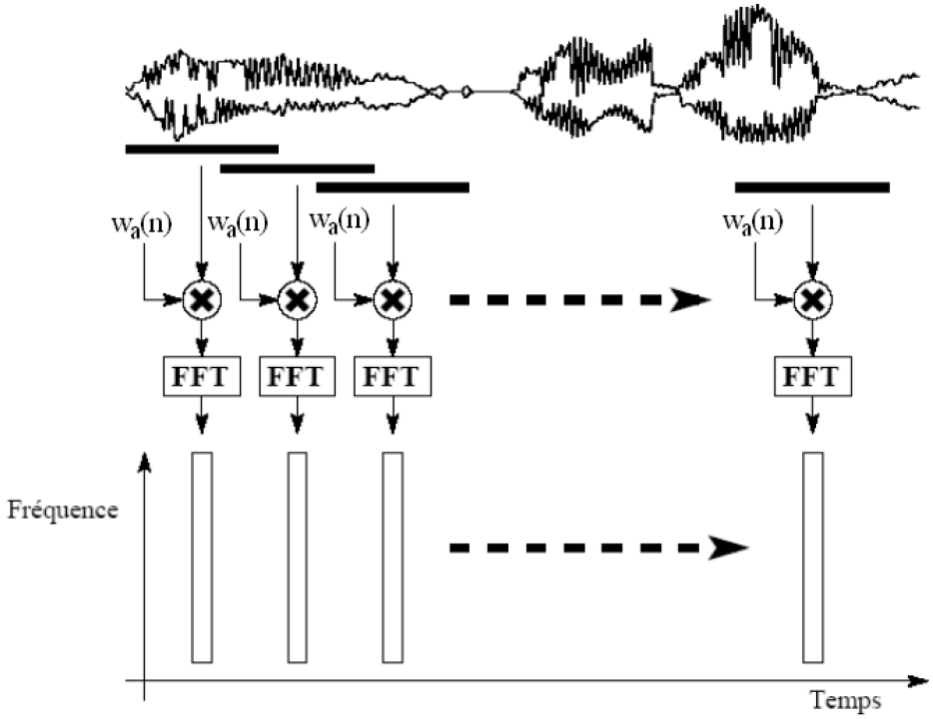

Figure D.1 - Schéma de principe de l'analyse par TFCT [Laroche, 1995]

D.2. Synthèse par addition-recouvrement

Lors de la synthèse, après le découpage et l'analyse des fenêtres, on cherche à reconstruire le signal temporel à partir de sa représentation fréquentielle. Pour cela, on utilise la méthode OLA (overlap and add, addition-recouvrement) qui consiste à prendre la TFCT inverse des spectres X_t, puis à additionner les signaux obtenus après les avoir multipliés par une fenêtre de pondération en les faisant se recouvrir partiellement.

Le résultat est donné par :

$$\tilde{x}(n) = \sum_{t=1}^{T} w_s(n - \tau_s(t))\tilde{x}_{\tau_s}(n - \tau_s(t)) \qquad \text{(D.5)}$$

avec $\tau_s(t) = \tau_a(t)$ et $\tilde{x}_{\tau_s}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}(\tau_s(t), k) e^{2i\pi\frac{kn}{N}}$

Lors de la mise en œuvre de ce traitement, il faut s'assurer d'être en mesure de reconstruire parfaitement un signal lorsqu'aucune transformation ne lui est appliquée. Le critère de reconstruction parfaite, condition suffisante, est alors :

$$\sum_{t=1}^{T} w_a(n - \tau_a(t)) w_s(n - \tau_a(t)) = 1 \qquad \text{quelque soit } n \qquad \text{(D.6)}$$

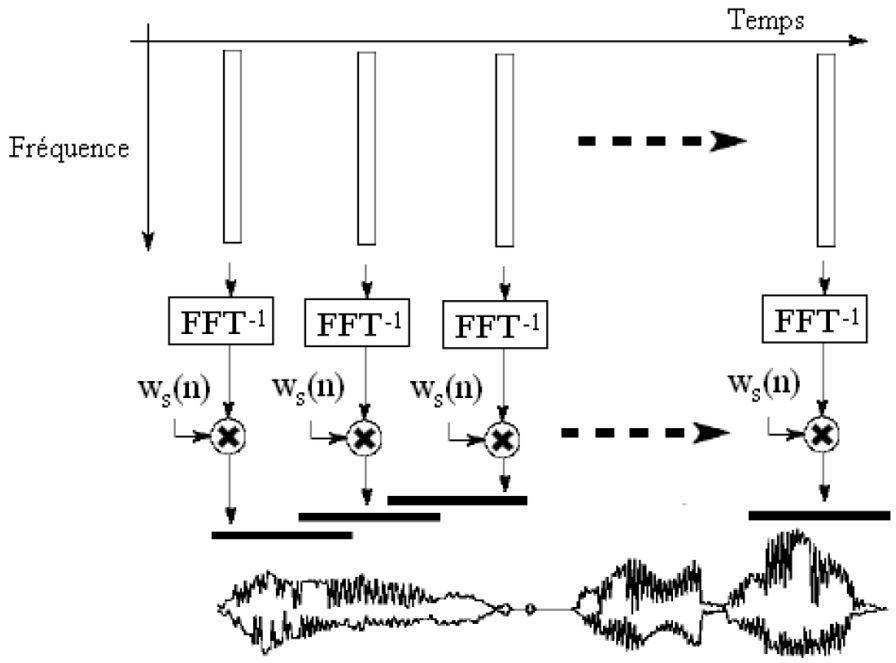

Figure D.2 - Schéma de principe de la synthèse par la méthode overlap-add [Laroche, 1995]

Plusieurs paramètres sont à choisir avec soin. Le premier d'entre eux est le type de fenêtre à utiliser. Les différentes options qui se présentent à nous sont :

- la fenêtre rectangulaire, qui présente dans son spectre le lobe principal le plus étroit mais des lobes secondaires non négligeables
- la fenêtre de Hamming, au lobe principal un peu plus large mais pour laquelle le 1er lobe secondaire se situe à -43 dB
- la fenêtre de Hann (ou cosinus surélevé), qui ressemble à la précédente avec un 1er lobe secondaire plus fort mais une pente de 18 dB par octave pour les lobes suivants
- les fenêtres de Blackman-Harris, aux lobes secondaires très bas mais au lobe principal plus large.

Fenêtre	Lobe 2aire (dB)	Pente (dB/oct)	Bande passante à 6 dB (bins)
Rectangulaire	-13	-6	1.21
Triangulaire	-27	-12	1.78
Hann	-32	-18	2.00
Hamming	-43	-6	1.81
Blackman-Harris 4	-92	-6	2.72

Tableau D.1 - Caractéristiques des principales fenêtres d'analyse [Harris, 1978]

Le meilleur compromis est donc dans notre contexte la fenêtre de Hann ou celle de Hamming.

Il nous faut ensuite choisir le pourcentage de recouvrement des fenêtres d'analyse. Un recouvrement de 50% permet un bon compromis entre précision temporelle de l'analyse et lourdeur de traitement.

Pour satisfaire le critère de reconstruction parfaite (D.6), nous prendrons donc un recouvrement de 50% avec la racine de la fenêtre de Hann, c'est-à-dire une arche de sinusoïde, à l'analyse comme à la synthèse :

$$w_a(n) = w_s(n) = \sin\left(\frac{\pi n}{N}\right) \tag{D.7}$$

La taille de la transformée de Fourier (NFFT) et le nombre de bandes fréquentielles sont également des paramètres cruciaux à déterminer, de même que la longueur de la fenêtre de signal analysé à chaque pas. Des tests ont été réalisés avec 1, 4, 20, 40 et 80 bandes inspirées des bandes ERB (voir Figure 3.2) pour différentes valeurs de NFFT (de 512 à 8192) ainsi qu'avec un grand nombre de longueurs de fenêtre différentes. Les largeurs des filtres de lissage (voir 3.4.3) sont aussi à régler. Tous ces paramètres doivent être choisis en fonction de la scène sonore étudiée, en prenant en compte les objectifs de précision et les moyens à disposition pour les obtenir, notamment en termes de capacité de calcul.

Bibliographie

[Bruneau, 1983] Bruneau, M. *Introduction aux théories de l'acoustique*. Le Mans, Université du Maine. (1983)

[Cherry, 1953] Cherry, E.C. *Some experiments on the recognition of speech with one and with two ears* Journal of the Acoustical Society of America, Vol. **25**, pp.975-979. (1953)

[Daniel, 2000] Daniel, J. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Thèse de Doctorat. Paris, Université Pierre et Marie Curie (Paris VI). (2000) http://gyronymo.free.fr

[Daniel et al., 2003] Daniel, J., Nicol, R. et Moreau, S. *Further investigations of High Order Ambisonics and Wave Field Synthesis for holophonic sound imaging*. 114th AES Convention, Amsterdam. (2003)

[Faller and Baumgarte, 2002] Faller, C. et Baumgarte, F. *Binaural Cue Coding: A novel and efficient representation of spatial audio*. Proc. ICASSP, Orlando, Floride. (2002)

[Faller and Baumgarte, 2003] Faller, C. et Baumgarte, F. *Binaural Cue Coding - Part II: Schemes and applications* IEEE Trans. on Speech and Audio Proc., Vol. **11** (n°6), pp.520-531. (2003)

[Gerzon, 1973] Gerzon, M.A. *Periphony: With-Height Sound Reproduction* Journal of the Audio Engineering Society, Vol. **21** (n°1), pp.2-10. (1973)

[Gerzon, 1985] Gerzon, M.A. *Ambisonics in Multichannel Broadcasting and Video* Journal of the Audio Engineering Society, Vol. **33** (n°11), pp.859-871. (1985)

[Gerzon, 1992] Gerzon, M.A. *General Metatheory of Auditory Localisation*. 92nd AES Convention. (1992)

[Harris, 1978] Harris, F.J. *On the use of windows for harmonic analysis with the discrete fourier transform* Proceedings of the IEEE, Vol. **66** (n°1), pp.51-82. (1978)

[Hotelling, 1933] Hotelling, H. *Analysis of a complex of statistical variables into principal components.* Journal of Educational Psychology, Vol. **24**, pp.417-441, 498-520. (1933)

[Laroche, 1995] Laroche, J. *Traitement des signaux audio-fréquences,* notes de cours. Paris, ENST. (1995)

[MacQueen, 1967] MacQueen, J. *Some methods for classification and analysis of multivariate observations.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, University of California Press. Vol.**1**, pp.281-297. (1967)

[Merimaa and Pulkki, 2005] Merimaa, J. et Pulkki, V. *Spatial impulse response rendering I: analysis ans synthesis* Journal of the Audio Engineering Society, Vol. **53** (n°12). (2005)

[Moreau, 2006] Moreau, S. *Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance*, Thèse de Doctorat. Le Mans, Université du Maine. (2006)

[Moreau et al., 2006] Moreau, S., Daniel, J. et Bertet, S. *3D Sound Field Recording with Higher Order Ambisonics – Objective Measurements and Validation of a 4th Order Spherical Microphone*. 120th AES Convention, Paris. (2006)

[Nicol, 1999] Nicol, R. *Restitution sonore spatialisée sur une zone étendu : application à la téléprésence*, Thèse de Doctorat. Le Mans, Université du Maine. (1999)

[Pearson, 1900] Pearson, K. *On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling.* Philosophical Magazine, Vol. **50**, pp.157-175. (1900)

[Pulkki and Faller, 2006] Pulkki, V. et Faller, C. *Directional audio coding: filterbank and STFT-based Design*. 120th AES Convention, Paris. (2006)