UNIVERSITE BRETAGNE LOIRE

Université du Maine

Thèse de Doctorat

Mathieu BAQUE

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université du Maine
sous le sceau de l'Université Bretagne Loire

École doctorale : SPIGA

Discipline : 60
Spécialité : Acoustique
Unité de recherche : Laboratoire d'Acoustique de l'Université du Maine — UMR CNRS 6613

Soutenue le 09 juin 2017
Thèse N° : 2017LEMA1013

Analyse de scène sonore multi-capteurs

Un front-end temps-réel pour la manipulation de scène

JURY

Rapporteurs :	**Régine LE BOUQUIN-JEANNES**, Professeure à l'université Rennes 1, LTSI
	Emmanuel VINCENT, Directeur de recherche, INRIA Nancy
Examinateurs :	**Rozenn NICOL**, Ingénieure de recherche, Orange Labs
	Rémi GRIBONVAL, Directeur de recherche, INRIA Rennes
Directeur de thèse :	**Manuel MELON**, Professeur à l'université du Maine, LAUM
Co-directeur de Thèse :	**Alexandre GUERIN**, Ingénieur de recherche, Orange Labs

Résumé

La thèse s'inscrit dans un contexte d'essor de l'audio spatialisé (contenus 5.1, Dolby Atmos...). Comme un prolongement des formats 2D existants, l'audio 3D qui intègre une dimension supplémentaire - l'élévation - promet d'améliorer encore l'immersion sonore de l'auditeur. Parmi les formats audio 3D existants, l'ambisonie permet une représentation spatiale homogène du champ sonore et se prête naturellement à des manipulations comme des rotations ou des distorsions du champ sonore.
L'objectif de la thèse a été de fournir un outil d'analyse et de manipulation de contenus audio captés au format ambisonique, essentiellement vocaux. La nécessité d'un fonctionnement en temps-réel, donc causal et à faible latence, et d'une utilisation dans des conditions acoustiques réelles ont été les principales contraintes à respecter. L'outil mis au point est basé sur une décomposition de la scène en objets sonores (source). Pour ce faire, un algorithme d'analyse en composantes indépendantes (ACI) est appliqué trame à trame, permettant de décomposer le champ sonore en un ensemble de contributions acoustiques, certaines correspondant aux champs directs des sources et d'autres à de la réverbération résiduelle. Une étape de classification bayésienne, appliquée aux composantes extraites, permet alors l'identification et le dénombrement des sources sonores réelles présentes dans la scène au cours du temps. Par ailleurs, le formalisme ambisonique permet d'extraire à partir de la matrice de mélange identifiée par ACI la localisation de chacune des sources identifiées et donc de fournir une cartographie de la scène sonore en temps-réel.
Une étude exhaustive des performances a été menée sur des scènes sonores réelles en fonction de plusieurs paramètres, comme le nombre de sources présentes dans le contenu, l'environnement acoustique, la longueur des trames d'analyse, ou l'ordre ambisonique utilisé. Des résultats fiables en terme de localisation et de comptage de sources ont été obtenus pour des trames d'analyse de quelques centaines de millisecondes. L'algorithme, exploité comme prétraitement dans un prototype d'assistant vocal domestique, a permis d'améliorer significativement les performances de reconnaissance, notamment en prise de son lointaine et en présence de sources interférentes.

Abstract

The context of this thesis is the development of spatialized audio (5.1 contents, Dolby Atmos...) and particularly of 3D audio. Among the existing 3D audio formats, Ambisonics and Higher Order Ambisonics (HOA) allow an homogeneous spatial representation of a sound field and allows basics manipulations, like rotations or distorsions.
The aim of the thesis is to provides efficient tools for ambisonics and HOA sound scenes analysis and manipulations. A real-time implementation and robustness to reverberation are the main constraints to deal with. The implemented algorithm is based on a frame-by-frame Independent Component Analysis (ICA), with decomposes the sound field into a set of acoustic contributions. Then a bayesian classification step is applied to the extracted components to identify the real sources and the residual reverberation. Direction of arrival of the sources are extracted from the mixing matrix estimated by ICA, according to the ambisonic formalism , and a real-time cartography of the sound scene is obtained. Performances have been evaluated in different acoustic environnements in view of several parameters such as the ambisonic order, the frame length or the number of sources. Accurate results in terms of source localization and source counting have been obtained for frame lengths of a few hundred milliseconds. The algorithm is exploited as a preprocessing for a speech recognition prototype and allows a significant increasing of the recognition results, in far-field conditions and in the presence of noise and interferent sources.

Remerciements

Je tiens avant toute chose à remercier les personnes qui m'ont permis de mener à bien ce travail de thèse.

Je remercie en premier lieu mon directeur de thèse M. Alexandre Guérin, avec qui j'ai passé trois années riches à tout point de vue. Merci pour tes conseils et le temps passé à superviser mon travail.

Je remercie mon directeur de thèse académique M. Manuel Melon, merci Manu pour tes relectures avisées et pour avoir toujours su trouver du temps pour moi.

Je remercie les membres de mon jury de thèse : Mme Régine Le Bouquin-Jeannès et M. Emmanuel Vincent, rapporteurs, pour leur relecture attentive de mon manuscrit et leurs remarques précieuses ; Mme Rozenn Nicol et M. Rémi Gribonval, examinateurs, pour avoir accepté d'évaluer mes travaux.

Je remercie M. Marc Emerit, M.Rémi Gribonval et Mme Rozenn Nicol d'avoir bien voulu participer à mon comité de suivi de thèse et m'avoir ainsi aidé à orienter mes recherches.

J'ai eu la chance de trouver, au sein de l'équipe TPS, des collègues dont la compagnie au quotidien a été plus qu'agréable. Merci donc à Arnaud, Gregory, Marc et Lauréline, pour vos conseils et votre aide lorsque j'en ai eu besoin, pour votre bonne humeur et votre camaraderie, pour vos plaisanteries, et j'en passe...

Je remercie également les collègues de Lannion : merci Jérôme pour ton aide précieuse sur le HOA, merci Rozenn pour tes relectures et tes encouragements.

Je tiens à remercier également mes amis et ma famille, en particulier mes parents, qui m'ont toujours soutenu durant ces longues études.

Table des matières

Table des figures

Acronymes

ACI Analyse en Composantes Indépendantes

ACP Analyse en Composantes Principales

AVI Analyse en Vecteurs Indépendants

BSS Blind Source Separation

dirAC directional Audio Coding

DEMIX Direction Estimation of Mixing matrIX

DNN Deep Neural Network

DOA Direction of Arrival

DUET Degenerate Unmixing Estimation Technique

DVD Digital Versatile Disc

EBM Entropy Bound Minimization

EFICA Efficient FastICA

ERB Equivalent Regular Bandwidth

ERBM Entropy Rate Bound Minimization

ERM-ARG Entropy Rate Minimisation - AutoRegressive Generalized gaussian distribution

FOBI Fourth Order Blind Identification

HARPEX High Angular Resolution Planewave Expansion

HMM Hidden Markov Model

HRTF Head-Related Transfer Function

HOA High Order Ambisonics

JADE Joint Approximate Diagonalization of Eigen-matrices

JBSS Joint Blind Source Separation

MCCA Multiset Canonical Correlation Analysis

RMS Root Mean Square

SAD Sound Activity Detection

SAR Signal-to-Artifacts Ratio

SAS Séparation Aveugle de Sources

SDR Signal-to-Distorsion Ratio

SIR Signal-to-Interference Ratio
SIRR Spatial Impulse Response Rendering
SNR Signal-to-Noise Ratio
SoC System on Chip
SRIR Spatial Room Impulse Response
STFT Short-Time Fourier Transform
SVD Singular Value Decomposition
SVM Support Vector Machine
TF Transformée de Fourier
TR Temps de réverbération
VBAP Vector Based Amplitude Panning
WASOBI Weights-Adjusted Second-Order Blind Identification

Introduction

Ce rapport expose les travaux réalisés durant ma thèse CIFRE effectuée au sein d'Orange Labs en partenariat avec le Laboratoire d'Acoustique de l'Université du Maine. J'ai intégré l'équipe de recherche TPS - Traitement de la Parole et du Son -, dont les principaux axes de recherche sont le codage de la parole, l'amélioration de la qualité de service (débruitage, annulation d'écho...) et le son immersif, à travers des travaux portant sur les formats audio 3D ambisonique et binaural.
Ma thèse a eu pour objet l'analyse et la manipulation temps-réel de contenus audio 3D, à partir d'une captation microphonique réelle de type ambisonique. Plus précisément, l'approche développée a consisté en une décomposition de la scène captée en sources, appelées aussi objets sonores, via une étape de localisation et de séparation des principales sources d'intérêt au cours du temps.
L'analyse et le traitement de contenus audio 3D, notamment au format ambisonique, trouvent une utilité à travers différents champs d'applications que nous détaillons brièvement dans cette introduction.

Utilisation du son 3D dans un contexte multimédia Ces dernières années ont vu l'avènement des contenus audiovisuels en trois dimensions, qui améliorent l'immersivité du spectacle et placent le spectateur au cœur de l'action. La vidéo 3D est maintenant solidement implantée, avec notamment un grand nombre de productions cinématographiques récentes filmées ou réalisées en 3D. L'immersion sonore est également en plein essor, à travers la musique ou le cinéma, même si celle-ci est pour l'instant souvent restreinte à une représentation en deux dimensions (horizontale) de la scène sonore. Les formats multicanal 2D 5.1 ou 7.1. sont parmi les plus répandus et sont utilisés depuis longtemps dans le cadre d'un usage professionnel (salles de cinéma ou salles de spectacle). Ceux-ci font maintenant leur apparition chez les particuliers avec la démocratisation des installations de type home-cinéma : les disques DVD ou Blu-ray supportent ces types de formats audio spatialisés, et certaines chaînes de télévision proposent maintenant des films ou émissions au format 5.1.
Par rapport aux formats 2D, l'audio 3D intègre l'élévation comme dimension supplémentaire et promet ainsi d'améliorer l'immersion sonore des auditeurs. Au cinéma, le son 3D est en passe de se généraliser grâce notamment à la technologie ATMOS de Dolby proposant une solution intégrée pour la génération et le rendu de contenus audio 3D. Par ailleurs, des produits intégrant de l'audio 3D pour le grand public ont déjà vu le jour, comme par exemple le site web Nouvoson [1] de Radio France qui propose l'écoute au casque de contenus spatialisés grâce à la technologie binaurale. La réalité virtuelle est également un domaine où le son 3D est amené à prendre une place importante. Les

casques de réalité virtuelle sont maintenant légion : aux côtés du précurseur Oculus Rift, on peut citer le tout dernier HTC Vive ou encore le PlayStation VR de Sony. Ces casques sont par essence basés sur une représentation 3D de contenus immersifs, même si celle-ci se résume encore bien souvent à l'aspect visuel, l'audio restant le parent pauvre en termes de contenus générés. En cela, l'acteur Youtube et son lecteur de vidéos 360° semble être le plus avancé en ce qui concerne l'immersion sonore, avec l'intégration native du format audio 3D ambisonique.
D'un point de vue technique, la généralisation de l'audio 3D et le foisonnement des formats de représentation utilisés soulèvent des problèmes pour les éditeurs de contenus ou les fournisseurs de services audio immersifs. Pour la restitution sonore, des problèmes d'interopérabilité apparaissent entre les formats de captations et ceux de diffusion dont la compatibilité n'est pas garantie, le rendu sonore peut alors se retrouver fortement dégradé. Ces problématiques existaient depuis l'avènement de la stéréo et du 5.1 mais elles ont étés mises en exergue avec le passage à la 3D.
Pour un ingénieur du son, la manipulation et le mixage de formats différents deviennent rapidement insolubles, en particulier pour travailler à partir de prises de son réelles où plusieurs types de captations sont utilisés conjointement, chacune d'elles possédant des propriétés différentes en termes de directivité ou de timbre. Il manque en fait actuellement des outils permettant de travailler facilement et intuitivement avec des contenus spatialisés et de gérer simultanément les différents formats.

La prise de son multicanal pour la domotique L'audio multicanal n'est pas l'apanage de la création de contenus et trouve également des applications dans le secteur des télécommunications et de la téléphonie, ainsi que plus récemment pour la domotique via l'assistance vocale.
Jusqu'alors cantonnées à des interfaces tactiles, les applications de domotique s'enrichissent de l'interface vocale grâce aux progrès immenses des moteurs de transcription automatique, comme ceux de l'assistant vocal Siri d'Apple ou de Google Voice, portés notamment par le développement des réseaux de neurones profonds (DNN, *Deep Neural Networks*).
Si les outils de reconnaissance vocale présentent actuellement d'excellentes performances pour des prises de son de type champ proche, il en est en général autrement lorsqu'il s'agit d'une captation sonore en champ lointain, en présence de réverbération ou de bruit. L'enjeu technique est alors de pouvoir maintenir une interaction naturelle et fiable lorsque le locuteur s'éloigne du capteur ou lorsque celui-ci se trouve dans des conditions acoustiques défavorables.
L'étude récente de Kinoshita et al. [2] dans le cadre du projet REVERB a mis en avant l'apport d'une prise de son multicanal par rapport à une prise de son monophonique pour de la reconnaissance vocale en milieu réverbérant. Amazon a été un des premiers acteurs à tirer parti du multicanal avec son assistant vocal Echo, grâce à un réseau circulaire de 7 capsules microphoniques, qui promet grâce à du traitement d'antenne une interaction naturelle même en champ lointain. Dernièrement, les constructeurs de *chipsets* ont commencé à proposer des SoC (*System on Chip*) dédiés à ces types d'applications. Ceux-ci

intègrent une antenne de plusieurs microphones et un logiciel couplant un moteur de reconnaissance vocale avec des algorithmes de focalisation et de réduction de bruit. XMOS propose ainsi une antenne circulaire de 6 microphones [3] couplée à une suite logicielle de traitement audio.
En résumé, l'intérêt d'un système multi-capteurs est actuellement reconnu par la majorité des acteurs du domaine qui souhaitent rendre leur solution de reconnaissance vocale la plus robuste possible dans des scénarios d'utilisation variés.

Le format ambisonique Pour toutes les applications citées précédemment, le format audio multicanal ambisonique apparaît comme une solution adaptée pour capter et manipuler aisément des contenus sonores spatialisés. Introduit par Gerzon dans les années 70 [4] [5], l'ambisonie présente ainsi des propriétés particulièrement intéressantes.
Basée sur un modèle physique, celle-ci consiste en une projection du champ sonore sur une base de fonctions harmoniques sphériques autour d'un point, assimilable au point de vue de l'auditeur. En pratique, chaque canal audio correspond à un encodage directif du champ sonore suivant une fonction harmonique sphérique. L'ambisonie permet ainsi une représentation spatialement homogène de la scène sonore - aucune direction de l'espace n'est *a priori* privilégiée - et se prête naturellement à des manipulations simples, telles que des rotations du champ ou des focalisations. C'est d'ailleurs pour cela qu'elle a été choisie par Youtube comme format audio pour ses contenus 360° en ligne. Elle est de plus multi-échelle, et permet donc de modifier la précision spatiale de la scène en modulant le nombre de canaux utilisés tout en conservant une représentation cohérente. La représentation ambisonique la plus réduite, appelée ambisonie d'ordre 1, est constituée de quatre canaux, soit une composante omnidirectionnelle et trois composantes bidirectionnelles suivant les axes $\vec{x}$, $\vec{y}$ et $\vec{z}$. Au delà de quatre canaux, des directivités plus sélectives sont ajoutées, et forment les ordre ambisoniques supérieurs, allant de l'ordre 2 jusqu'à théoriquement un ordre infini. On parle alors de format HOA pour *Higher Order Ambisonics* [6].
Pour capter des scènes sonores réelles au format ambisonique, il existe actuellement plusieurs microphones sur le marché, permettant de générer des contenus à différents ordres, allant principalement de l'ordre 1 (microphone SoundField) jusqu'à l'ordre 4 (microphone Eigenmike). Par ailleurs, il existe des systèmes de captation expérimentaux permettant de monter théoriquement jusqu'à l'ordre 6. Ces dispositifs sont conçus à des fins de création de contenus et sont sollicités par exemple dans le cadre du projet 4Ever2 [7], auquel Orange participe, qui porte sur les nouveaux formats audiovisuels immersifs. Dans un contexte de domotique par commande vocale, une captation ambisonique peut potentiellement être mise à profit pour effectuer du réhaussement de la parole et améliorer la reconnaissance comme évoqué dans le paragraphe précédent, grâce notamment à son formalisme se prêtant naturellement à la formation de voies. L'homogénéité spatiale d'une captation ambisonique est également un atout pour la reconnaissance vocale, car le locuteur peut potentiellement se trouver à n'importe quelle position autour du microphone sans que la qualité de la prise de son ne soit altérée.

Enjeux de la thèse L'objectif du travail de thèse a été de mettre au point un outil pour l'analyse et la manipulation avancée de contenus ambisoniques issus d'une captation réelle. Les tâches ciblées sont principalement la localisation, le déplacement, le réhaussement ou l'atténuation d'une source sonore au sein d'une scène complexe en temps-réel, ainsi que la déréverbération de la scène.
Ces tâches couvrent l'ensemble des cas d'usage évoqués précédemment, de l'édition de contenu pour l'ingénieur du son jusqu'à l'assistant vocal basé sur le multicanal. Quelle que soit la finalité de l'outil mis au point, la réalisation de ces tâches passenécessairement par une étape préalable d'analyse de la scène sonore, afin de pouvoir identifier, localiser et isoler les principales sources d'intérêt et de bruit. La précision de l'analyse est primordiale pour espérer réaliser les manipulations souhaitées, c'est pourquoi l'étape de cartographie de la scène sonore a été au cœur du travail de thèse, de même que la séparation des différentes sources sonores identifiées.
Le dénombrement des sources sonores actives en présence de réverbération est une problématique centrale pour la manipulation de contenus. Pour de la commande vocale par exemple, l'identification des sources actives permet de réduire le coût algorithmique en restreignant les traitements aux seules sources réelles identifiées. Dans un contexte multimédia, il est également nécessaire d'identifier la contribution de chaque source, afin de pouvoir appliquer les manipulations désirées tout en conservant un rendu cohérent du point de vue de l'auditeur.
La latence de l'outil d'analyse est un point qui a été pris en compte tout au long des expérimentations menées. En effet, pour un certain nombre d'applications, il est nécessaire de prévoir un traitement en temps-réel des données, impliquant une approche causale et à faible latence. Pour des applications comme les télécommunications, la réduction de la latence de traitement est un enjeu fort afin de pouvoir conserver des interactions fluides dans un contexte conversationnel. Pour de la commande vocale, quelques centaines de millisecondes de latence sont acceptables et permettent donc de relâcher un peu la contrainte de causalité. A l'opposé, l'aspect temps-réel n'est pas essentiel pour un outil de post-traitement pour l'ingénieur du son, dans la mesure toutefois d'un temps de traitement raisonnable.
La capacité à traiter des contenus microphoniques **réels** est également une caractéristique essentielle de l'outil d'analyse, qui nécessite donc de prendre en compte les imperfections du système de prise de son et les artefacts introduits par l'environnement acoustique (bruit de fond, réverbération).
Le dernier point concerne la qualité sonore des signaux manipulés. Une attention particulière a été apportée aux traitements appliqués, de façon à dégrader le moins possible le contenu spectral des sources.

Approche mise en œuvre Le principe de l'approche mise en place est une décomposition de la scène sonore en objets, représentant les contributions des différentes sources, auxquelles sont associés des indicateurs de localisation.
La séparation des sources sonores dans un contenu ambisonique est une problématique déjà abordée dans la littérature, avec des solutions à la fois temps-fréquence comme HAR-

PEX [8] et dirAC [9] destinées à l'*upmix*, ou basées sur des techniques de régularisation parcimonieuse ([10], [11]) ou encore sur des méthodes de séparation aveugle de sources ([12]).
L'algorithme d'analyse développé durant la thèse est articulé autour de deux modules principaux :

— Une première étape d'analyse en composantes indépendantes (ACI), appliquée dans le domaine temporel, permet de décomposer le champ acoustique en un ensemble d'évènements sonores. L'ACI permet à la fois d'identifier des évènements sonores mais également des coefficients de mélange pour chacun d'entre eux, pouvant être facilement convertis en informations de localisation grâce au formalisme de l'encodage ambisonique.

— En conditions réelles, certaines des composantes sonores extraites par ACI vont correspondre à la contribution principale d'une source, tandis que d'autres seront uniquement des résidus de réverbération causés par l'effet de salle. C'est pourquoi une étape additionnelle de classification supervisée a été mise en œuvre, pour déterminer si chaque composante extraite relève d'une source sonore réelle ou bien s'il s'agit de réverbération résiduelle.

A l'issue de ces traitements que l'on applique sur de courtes trames temporelles, une cartographie de la scène sonore est obtenue, permettant de connaître au cours du temps à la fois le nombre de sources sonores actives et leur position dans la scène.
La séparation des sources identifiées peut ensuite être effectuée grâce à un procédé de formation de voies sous contraintes.

Plan du document Ce rapport de thèse est constitué de six parties, les deux premières constituant l'état de l'art sur l'ambisonie et l'analyse de scène multi-capteurs, les quatre suivantes décrivant les techniques mises en œuvre et les résultats expérimentaux obtenus.
Le premier chapitre est consacré à l'acoustique et expose la décomposition du champ sonore sur une base de fonctions harmoniques sphériques, qui est à la base de l'ambisonie. Les équations de l'encodage ambisonique sont décrites de même que le système de captation utilisé durant la thèse et ses limitations physiques. Quelques considérations sur la restitution d'un contenu ambisonique et sur l'effet de salle viennent compléter cette première partie.
Le second chapitre aborde l'état de l'art concernant l'analyse de scène et la séparation de sources multicanal. L'accent est mis sur les méthodes de séparation spatiale et sur celles spécialement dédiées au format ambisonique.
Dans un troisième chapitre, une évaluation des principales méthodes d'analyse de scène est effectuée sur des contenus ambisoniques synthétiques - donc respectant parfaitement le formalisme de l'ambisonie -, avec tout d'abord un encodage parfait instantané des sources sonores, puis en présence d'une réverbération simulée.
L'analyse de scène et la séparation de sources appliquées à des contenus réels fait l'objet du chapitre 4, avec la description de la solution mise au point et articulée autour d'un algorithme d'ACI nommé ERBM. Une étude exhaustive des résultats obtenus en présence de deux ou trois sources sonores est réalisée, en fonction de certains paramètres comme

l'ordre ambisonique, le niveau de réverbération ou la longueur des trames d'analyse.
Le cinquième chapitre décrit le procédé permettant de résoudre la problématique du dénombrement des sources actives dans une scène réelle, qui est basé sur une approche de type classification bayésienne appliquée aux composantes extraites par ACI.
Dans le dernier chapitre, l'outil complet d'analyse de scène est appliqué à un cas d'usage concret, à savoir une application de commande vocale. L'étape de dénombrement des sources est tout d'abord évaluée, puis des scores comparatifs de reconnaissance vocale viennent valider l'intérêt de l'outil mis au point pour du pré-traitement appliqué à la transcription automatique dans un environnement bruité, par rapport à une simple prise de son omnidirectionnelle.

1. Format ambisonique et environnement acoustique

Les travaux effectués durant la thèse ont eu pour objet l'analyse de contenus sonores au format ambisonique. Ce premier chapitre pose les bases théoriques de ce formalisme. Les équations de l'acoustique linéaire sont dérivées jusqu'au développement en harmoniques sphériques du champ acoustique, qui est à la base de l'ambisonie. Les limitations théoriques de cette représentation sont ensuite évoquées, ainsi que les limites physiques propres aux systèmes de prise de son ambisonique. Le système de captation utilisé durant cette thèse est également caractérisé. Enfin, pour prendre en compte des considérations pratiques, la dernière partie traite de l'influence de l'environnement acoustique (effet de salle) sur la captation.

1.1. Représentation du champ acoustique

Les équations développées dans cette section aboutissent à l'expression du champ sonore à l'aide d'une décomposition sur une base d'harmoniques sphériques, ce qui constitue le fondement de l'ambisonie. On peut trouver de tels développements dans la plupart des ouvrages de référence en acoustique (par exemple [13] [14] et [15]). On peut noter que suivant les auteurs les formulations et les notations diffèrent, concernant notamment l'expression des séries de Fourier-Bessel sphériques.

1.1.1. Equation des ondes

En combinant les équations linéarisées de la mécanique des fluides, on établit l'équation des ondes, dite équation de d'Alembert :

$$\triangle p - \frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{1.1}$$

où p est la pression acoustique et c_0 la célérité de l'onde acoustique. Cette équation est valable dans un fluide parfait au repos et homogène, dans une zone exempte de source, sous l'hypothèse de petites perturbations. C'est la configuration que nous adopterons par la suite. Pour les développements suivants, l'indice a est sous-entendu, considérant que l'on ne s'intéresse ici qu'aux variations de pression générées par le champ acoustique.

1.1.2. Représentation de Fourier

Fourier énonce que tout signal périodique de période T peut se décomposer en une somme pondérée de sinusoïdes, oscillant à des fréquences multiples de $f_1 = \frac{1}{T}$, qui est

appelée fréquence fondamentale. Par extension aux signaux temporels non périodiques (vus comme des signaux périodiques de période infinie), on peut décomposer tout signal temporel comme une somme de signaux sinusoïdaux. On appelle cette décomposition transformée de Fourier (TF) qui se formule ainsi :

$$P(\omega) = \mathrm{TF}\left[p(t)\right] = \int_{-\infty}^{+\infty} p(t)e^{-i\omega t}dt \tag{1.2}$$

où $\omega = 2\pi f$ représente la fréquence angulaire ou pulsation.
Réciproquement, on définit l'opération de transformée de Fourier inverse permettant de représenter le signal fréquentiel sous forme temporelle :

$$p(t) = \mathrm{TF}^{-1}\left[P(\omega)\right] = \frac{1}{2\pi}\int_{-\infty}^{+\infty} P(\omega)e^{i\omega t}d\omega \tag{1.3}$$

1.1.3. Equation de Helmoltz

En se basant sur la représentation de Fourier, on peut ensuite reformuler l'équation des ondes pour un signal harmonique. Soit $p(t) = Pe^{i\omega t}$, alors $\frac{\partial^2 p}{\partial t^2} = -\omega^2 p$, avec ω la pulsation propre du signal.
On obtient alors l'équation de Helmoltz :

$$\triangle p + k^2 p = 0 \tag{1.4}$$

grâce à la relation de dispersion $k = \frac{\omega}{c_0}$ faisant apparaître $k = \frac{2\pi}{\lambda}$ le nombre d'ondes. L'équation de Helmoltz correspond à la représentation fréquentielle de l'équation des ondes.

1.1.4. Solutions en coordonnées sphériques

Il est parfois pratique de se placer en coordonnées sphériques pour décrire un phénomène acoustique. En particulier dans le cadre applicatif visé ici, à savoir l'audio immersif, un repère sphérique centré sur l'auditeur semble être la façon la plus naturelle de décrire la scène sonore. Le Laplacien s'exprime alors sous la forme suivante :

$$\triangle = \frac{1}{r}\frac{\partial}{\partial r}\Big(r^2\frac{\partial}{\partial r}\Big) + \frac{1}{r^2\sin\theta}\Big(\sin\theta\frac{\partial}{\partial\theta}\Big) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} \tag{1.5}$$

avec θ l'angle azimutal, ϕ l'angle polaire et r la distance du centre du repère.
Les développements suivants sont exprimés pour un signal harmonique de pulsation ω et généralisables à un signal quelconque grâce au formalisme de Fourier.

1.1.4.1. Solution générale

Lorsque le champ acoustique rayonné par une source dépend des trois composantes r, θ et ϕ, l'équation de Helmoltz :

$$[\frac{1}{r}\frac{\partial}{\partial r}\Big(r^2\frac{\partial}{\partial r}\Big) + \frac{1}{r^2\sin\theta}\Big(\sin\theta\frac{\partial}{\partial\theta}\Big) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2} + k^2]p = 0 \tag{1.6}$$

admet un ensemble de solutions se présentant sous une forme à variables séparées :

$$p(r,\theta,\phi,\omega) = R(r)\Theta(\theta)\Phi(\phi)e^{i\omega t} \tag{1.7}$$

avec le système d'équations suivant :

$$\left\{\begin{array}{rcl} R(r,k,m) & = & Ah_m^1(kr) + Bh_m^2(kr) \\ \Theta(\theta,n) & = & C\cos n\theta + D\sin n\theta \\ \Phi(\phi,m,n) & = & P_{mn}(\cos\phi) \end{array}\right.$$

où h_m^1 et h_m^2 désignent respectivement les fonctions de Hankel d'ordre m du premier et du second type pouvant s'exprimer comme des combinaisons linéaires de fonctions de Bessel j_m et de Neumann n_m sphériques :

$$\left\{\begin{array}{l} h_m^1(kr) = j_m(kr) + in_m(kr) \\ h_m^2(kr) = j_m(kr) - in_m(kr) \end{array}\right.$$

$P_{mn}(\cos\phi)$ est une fonction polaire impliquant le polynôme de Legendre :

$$P_{mn}(\eta) = (-1)^n(1-\eta^2)^{\frac{n}{2}}\frac{d^n}{d\eta^n}P_m(\eta) \text{ où } 0 \leqslant n \leqslant m \tag{1.8}$$

et

$$P_m(\eta) = \frac{1}{2^m \cdot m!}\frac{d^n}{d\eta^n}(\eta^2-1)^m \tag{1.9}$$

Pour alléger les notations, on applique à P_{mn} la semi-normalisation de Schmidt :

$$\tilde{P}_{mn}(\eta) = \sqrt{\epsilon_n\frac{(m-n)!}{(m+n)!}}P_{mn}(\eta) \text{ avec } \epsilon_0 = 1 \text{ et } \epsilon_n = 2 \text{ pour } n \geqslant 1 \tag{1.10}$$

On exprime alors les fonctions dites harmoniques sphériques $Y_{mn}^{\sigma}(\theta,\phi)$ de degré $m \geqslant 0$ et d'ordre $0 \leqslant n \leqslant m$ avec $\sigma = \pm 1$ en regroupant les fonctions polaires et azimutales :

$$Y_{mn}^{\sigma}(\theta,\phi) = \tilde{P}_{mn}(\cos\phi) \cdot \left\{\begin{array}{ll} \cos n\theta & \text{si } \sigma = 1 \\ \sin n\theta & \text{si } \sigma = -1 \text{ et } n \geqslant 1 \end{array}\right.$$

Les harmoniques sphériques forment une base orthogonale au sens du produit scalaire, soit

$$\langle Y_{mn}^{\sigma} | Y_{m'n'}^{\sigma'} \rangle = \frac{1}{2m+1}\delta_{mm'}\delta_{nn'}\delta_{\sigma\sigma'} \tag{1.11}$$

La figure 1.1 illustre les fonctions harmoniques sphériques de degrés 0 à 3. Ainsi, la première harmonique est omnidirectionnelle (degré 0) tandis que le degré 1 est constitué de trois fonctions bidirectionnelles suivant les axes $\vec{x}$, $\vec{y}$ et $\vec{z}$.
Les fonctions peuvent être normalisées suivant l'ordre auquel elles appartiennent pour simplifier l'expression du champ acoustique, soit :

$$\tilde{Y}_{mn}^{\sigma} = \sqrt{2m+1} \cdot Y_{mn}^{\sigma} \tag{1.12}$$

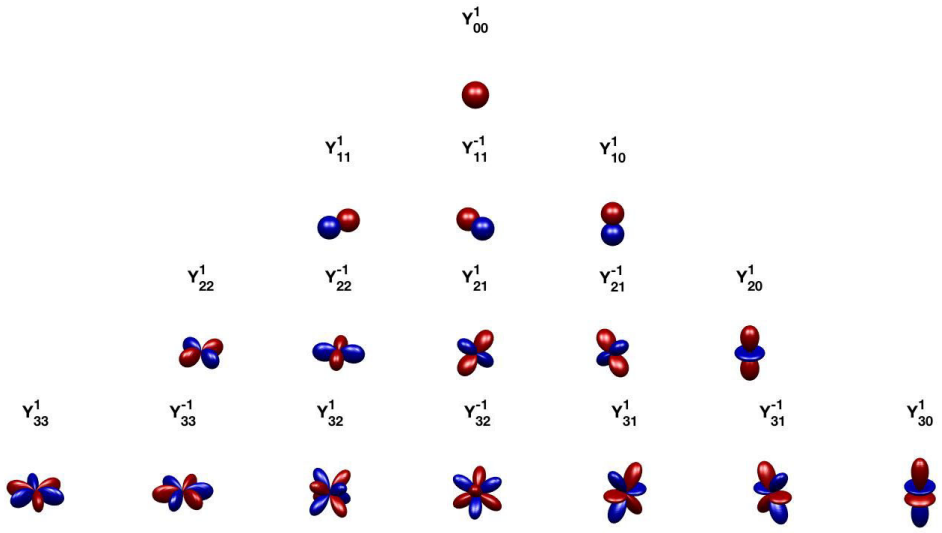

FIGURE 1.1. – Fonctions harmoniques sphériques de degré 0 (première ligne) à 3 (dernière ligne). En rouge : valeurs positives, en bleu : valeurs négatives.

On obtient ainsi une représentation générale du champ de pression :

$$p(r,\theta,\phi,\omega,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{m}\sum_{\sigma=\pm 1} \tilde{Y}_{mn}^{\sigma}(\theta,\phi) i^m (A_{mn}^{\sigma} j_m(kr) + iB_{mn}^{\sigma} n_m(kr)) e^{i\omega t} \tag{1.13}$$

où i est un nombre imaginaire dont le carré vaut -1.
La propagation en champ libre d'un signal provenant de l'infini se traduit par une expression où le coefficient B_{mn}^{σ} est nul, soit :

$$p(r,\theta,\phi,\omega,t) = \sum_{m=0}^{\infty}\sum_{n=0}^{m}\sum_{\sigma=\pm 1} A_{mn}^{\sigma} \tilde{Y}_{mn}^{\sigma}(\theta,\phi) i^m j_m(kr) e^{i\omega t} \tag{1.14}$$

Dans le cas d'un champ acoustique généré par une onde plane d'amplitude A et d'incidence (θ_p, ϕ_p), on montre dans [13] et [15] que A_{mn}^{σ} s'exprime de la façon suivante :

$$A_{mn}^{\sigma} = A\ \tilde{Y}_{mn}^{\sigma}(\theta_p, \phi_p) \tag{1.15}$$

On note que, dans le cas d'une onde sphérique, l'expression sera légèrement différente.
On peut interpréter le terme A_{mn}^{σ} comme une captation microphonique du signal A avec une directivité égale à $\tilde{Y}_{mn}^{\sigma}(\theta_p, \phi_p)$. Grâce à l'équation 1.14 et au principe de superposition, on peut ainsi théoriquement décrire, en tout point d'un espace exempt de sources, un champ acoustique complexe dès lors que l'on possède une infinité d'observations coïncidentes ayant pour directivités les fonctions harmoniques sphériques.

1.1.4.2. Solution approchée - troncature à l'ordre m

Il est d'usage, lorsque l'ordre et le degré ne sont pas évoquées simultanément, de parler d'harmoniques d'ordre m plutôt que de degré m [6], c'est la convention que nous allons à présent adopter.
La représentation du champ acoustique à l'aide de fonctions harmoniques sphériques est exacte mais n'est pas utilisable en l'état, car elle nécessite une infinité de termes. On peut néanmoins approcher celle-ci en utilisant seulement une partie des harmoniques sphériques, c'est-à-dire en effectuant une troncature du champ acoustique à un ordre m donné. L'expression approchée du champ est alors :

$$p(r,\theta,\phi,\omega,t) = \sum_{m'=0}^{m} \sum_{n=0}^{m'} \sum_{\sigma=\pm 1} A^{\sigma}_{m'n} \tilde{Y}^{\sigma}_{m'n}(\theta,\phi) i^{m'} j_{m'}(kr) e^{i\omega t} \tag{1.16}$$

L'utilisation d'un nombre plus important de composantes spatiales permet une représentation précise du champ sonore sur une zone plus étendue autour du centre du repère, comme illustré par la figure 1.2 pour une onde plane. L'étendue de cette zone est fonction à la fois de l'ordre de la troncature et de la longueur d'onde. Ainsi, pour un ordre donné, l'erreur de reconstruction sera plus importante en hautes fréquences qu'en basses fréquences. On peut trouver dans [6] une caractérisation de l'erreur théorique commise lors de l'estimation du champ de pression, en fonction de la distance du centre du repère et de l'ordre auquel la troncature est appliquée.

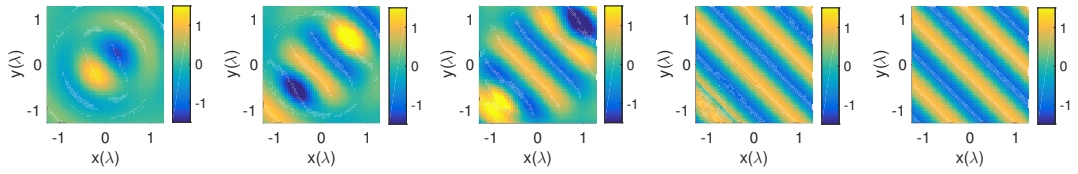

FIGURE 1.2. – Représentation tronquée à différents ordres d'une onde plane harmonique d'incidence $(45°, 0°)$ (coupe dans le plan $(\vec{x},\vec{y},z=0)$). De gauche à droite : approximations aux ordres 1, 3, 5, 10 et référence. Unité : longueur d'onde.

Cette représentation du champ sonore à l'aide d'un nombre fini de composantes spatiales constitue le fondement de l'ambisonie, qui est l'objet de la partie suivante.

1.2. Le format ambisonique

Michael Gerzon introduit dans les années 70 l'approche ambisonique pour la prise de son spatialisée [4] [5]. Celle-ci est basée sur la décomposition du champ acoustique sur une base de fonctions harmoniques sphériques comme vu précédemment (eq. 1.16). Concrètement il s'agit de synthétiser des directivités microphoniques correspondant aux premières harmoniques sphériques, par une combinaison linéaire de capteurs quasi- coïncidents.
Initialement limité à une représentation d'ordre 1 (utilisation des fonctions harmoniques sphériques jusqu'à l'ordre 1), le formalisme ambisonique a par la suite été étendu aux

ordres supérieurs notamment par les travaux de Jérôme Daniel [6]. Cette représentation avec un nombre de composantes plus important est couramment nommée HOA pour *Higher Order Ambisonics*.

1.2.1. Intérêt du format ambisonique

La représentation ambisonique présente plusieurs avantages. Le premier est de contenir directement l'information directionnelle du champ acoustique capté au point de mesure. En outre, on relie les coefficients d'encodage à l'ordre 1 d'une onde plane arrivant avec une incidence donnée aux coordonnées de celle-ci exprimée sous la forme $(\vec{x},\vec{y},\vec{z})$ (voir eq. 1.17). Le second intérêt est l'homogénéité de la captation spatiale, qui ne privilégie aucune direction de l'espace et autorise aisément des manipulations simples du contenu sonore, comme par exemple des rotations ou des focalisations. Enfin, le format ambisonique est indépendant du système de restitution et permet de restituer la scène sonore dans un grand nombre de configurations, allant de l'écoute au casque à la diffusion sur un ensemble de haut-parleurs, et ce en conservant une représentation spatiale cohérente. Un décodage sur un ensemble de haut-parleurs, de préférence régulièrement disposés, permet de restituer un champ sonore physiquement très proche du champ réel capté, dans une zone s'étalant autour du centre de la sphère d'écoute. Cette zone d'écoute optimale est appelée *sweet spot* et sa taille va dépendre de l'ordre ambisonique utilisé (figure 1.2).

1.2.2. Ambisonie à l'ordre 1 - le format B

La représentation de Gerzon se base sur les ordres ambisoniques 0 et 1 et est appelée format B. Les quatre composantes (ou canaux) qui constituent ce format sont habituellement nommées W pour l'ordre 0 et X, Y et Z pour les composantes bidirectionnelles de l'ordre 1. La composante W est une captation omnidirectionnelle du champ sonore tandis que les composantes X, Y et Z sont assimilables à des gradients de pression orientés suivant les trois dimensions de l'espace. On retrouve une illustration des directivités associées sur la figure 1.3. Une onde plane s d'incidence (θ,ϕ) génère ainsi le format B suivant :

$$\begin{cases} W &= Y_{00}^{1}(\theta,\phi)s &= s \\ X &= Y_{11}^{1}(\theta,\phi)s &= \sqrt{3}\cos\theta\cos\phi s \\ Y &= Y_{11}^{-1}(\theta,\phi)s &= \sqrt{3}\sin\theta\cos\phi s \\ Z &= Y_{10}^{1}(\theta,\phi)s &= \sqrt{3}\sin\phi s \end{cases} \tag{1.17}$$

Historiquement, le premier système de captation dédié au format B a été mis au point par Craven et Gerzon [5] sous le nom de microphone Soundfield. Celui-ci se présente sous la forme d'un tétraèdre régulier (figure 1.4), avec au centre de chaque face une capsule microphonique de type cardioïde. Le signal multicanal issu des capsules est nommé format A, la combinaison des canaux à l'aide d'une matrice de filtres permet ensuite de former les directivités ambisoniques et de retrouver le format B.
Bien que pratique d'utilisation, le format B souffre de limitations dues au faible nombre de composantes utilisées pour décrire le champ sonore, ce qui engendre une faible précision du rendu spatial. Dans le cadre d'une diffusion du format B sur des haut-parleurs, le

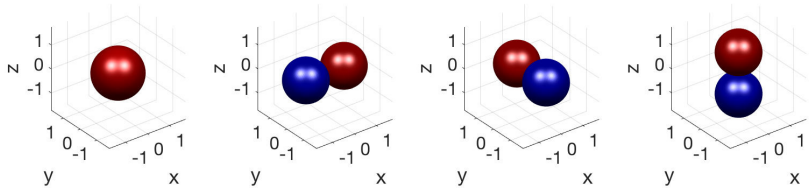

FIGURE 1.3. – Directivité des composantes du format B. De gauche à droite, composantes W, X, Y et Z.

FIGURE 1.4. – Photographie du microphone Soundfield, constitué de capsules cardioïdes positionnées sur les faces d'un tétraèdre virtuel [16].

.

manque de précision est à mettre en relation avec la taille du *sweet spot*, zone d'écoute dans laquelle le rendu sonore sera considéré comme satisfaisant. Pour avoir un ordre de grandeur, au-delà de 500 Hz, la zone d'écoute dans laquelle l'erreur de reconstruction est inférieure à 20 % devient plus petite que la dimension caractéristique d'une tête humaine [17] [6], ne permettant pas une restitution fidèle au niveau des oreilles de l'auditeur même si celui-ci se positionne de façon optimale.

1.2.3. Ambisonie aux ordres supérieurs - le format HOA

Afin d'augmenter la précision et la robustesse de la restitution sonore, l'ambisonie a par la suite été généralisée aux ordres supérieurs, on parle alors de format HOA. Le format HOA pallie en partie les défauts du format B. Il permet une meilleure précision de l'image sonore et un *sweet spot* élargi. D'un point de vue perceptif, les avantages de l'utilisation des ordres supérieurs pour la restitution d'une scène sonore immersive sont mis en évidence dans la thèse de S. Berthet [18].
Un contenu HOA d'ordre m est constitué de $(m+1)^2$ canaux, les ordres élevés faisant appel à des fonctions de directivité plus fines. Celles-ci sont illustrées jusqu'à l'ordre 3 par la figure 1.1 et on retrouve leur expression mathématique dans le tableau 1.1.
Il existe différentes conventions de normalisation des composantes ambisoniques, pour des raisons pratiques (optimisation de la dynamique dans les ordres supérieurs [19]) ou des raisons de formalisme. Une liste des différentes conventions et des facteurs de conversion associés est disponible dans [6]. Les plus utilisées sont les conventions SN3D et N3D, que

l'on retrouve dans le tableau 1.1. Ces deux conventions sont reliées par l'équation :

$$Y^{\sigma}_{mn,\mathrm{N3D}} = \sqrt{2m+1}\; Y^{\sigma}_{mn,\mathrm{SN3D}} \tag{1.18}$$

On remarque ici l'analogie avec l'équation 1.12 énoncée plus haut. La convention N3D permet d'obtenir des canaux de même énergie quel que soit l'ordre lors de la captation d'un champ sonore complètement diffus, alors qu'en utilisant la convention SN3D, c'est la somme des composantes de chaque ordre qui est de même énergie. Par la suite, nous adoptons la convention N3D.

Ordre	Nom	Y^{σ}_{mn}	Y_{SN3D}	Y_{N3D}
0	W	Y^{1}_{00}	1	1
1	X	Y^{1}_{11}	$\cos\theta\cos\phi$	$\sqrt{3}\cos\theta\cos\phi$
1	Y	Y^{-1}_{11}	$\sin\theta\cos\phi$	$\sqrt{3}\sin\theta\cos\phi$
1	Z	Y^{1}_{10}	$\sin\phi$	$\sqrt{3}\sin\phi$
2	U	Y^{1}_{22}	$\frac{\sqrt{3}}{2}\cos 2\theta\cos^2\phi$	$\sqrt{5}Y_{\mathrm{SN3D}}$
2	V	Y^{-1}_{22}	$\frac{\sqrt{3}}{2}\sin 2\theta\cos^2\phi$	$\sqrt{5}Y_{\mathrm{SN3D}}$
2	S	Y^{1}_{21}	$\frac{\sqrt{3}}{2}\cos\theta\sin 2\phi$	$\sqrt{5}Y_{\mathrm{SN3D}}$
2	T	Y^{-1}_{21}	$\frac{\sqrt{3}}{2}\sin\theta\sin 2\phi$	$\sqrt{5}Y_{\mathrm{SN3D}}$
2	R	Y^{1}_{20}	$\frac{(3\sin^2\phi-1)}{2}$	$\sqrt{5}Y_{\mathrm{SN3D}}$
3	-	Y^{1}_{33}	$\sqrt{\frac{5}{8}}\cos 3\theta\cos^3\phi$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{-1}_{33}	$\sqrt{\frac{5}{8}}\sin 3\theta\cos^3\phi$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{1}_{32}	$\frac{\sqrt{15}}{2}\cos 2\theta\sin\phi\cos^2\phi$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{-1}_{32}	$\frac{\sqrt{15}}{2}\sin 2\theta\sin\phi\cos^2\phi$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{1}_{31}	$\sqrt{\frac{3}{8}}\cos\theta\cos\phi(5\sin^2\phi-1)$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{-1}_{31}	$\sqrt{\frac{3}{8}}\sin\theta\cos\phi(5\sin^2\phi-1)$	$\sqrt{7}Y_{\mathrm{SN3D}}$
3	-	Y^{1}_{30}	$\frac{\sin\phi(5\sin^2\phi-3)}{2}$	$\sqrt{7}Y_{\mathrm{SN3D}}$

TABLEAU 1.1. – Fonctions de directivité ambisoniques d'ordres 0 à 3 en fonction de la convention d'encodage SN3D ou N3D. Les lettres capitales sont les noms usuels donnés aux canaux ambisoniques correspondants.

1.2.4. Système de prise de son HOA

Ce travail de thèse a eu pour objet la mise au point d'outils d'analyse et de manipulation de contenus ambisoniques réels ou simulés, indépendamment du système de captation utilisé à partir du moment où le signal encodé respecte le formalisme évoqué précédemment. Néanmoins, le formalisme ambisonique se heurte en pratique aux limites physiques des microphones et il est important de prendre celles-ci en compte pour l'évaluation des

traitements mis en œuvre. La plupart des systèmes de prise de son HOA sont généralement constitués d'un réseau de capteurs disposés à la surface d'une sphère. En effet, il est physiquement impossible de mettre au point un réseau de capteurs microphoniques coïncidents tel que chaque capteur possèderait des propriétés de directivité équivalentes à une composante de la représentation HOA. On passe ainsi par une représentation intermédiaire, le format capsules ou format A, qui est ensuite converti en format ambisonique par une matrice de filtres (figure 1.5). Pour ce faire, l'antenne doit posséder au moins autant de capsules que de canaux HOA à synthétiser.

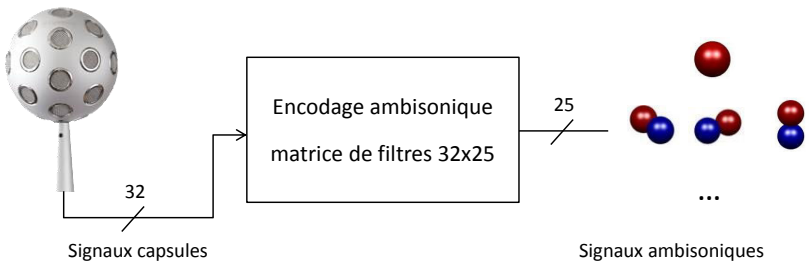

FIGURE 1.5. – Schéma de l'encodage microphonique HOA à l'ordre 4 (25 canaux) à partir d'un microphone sphérique 32 capsules.

1.2.4.1. Limitations théoriques de l'encodage microphonique

Les contraintes physiques du système de captation (taille de l'antenne, nombre de capteurs) impliquent des limites à la fois spatiales (directivité) et fréquentielles (bande passante) à la synthèse des composantes ambisoniques.
Tout d'abord, l'utilisation d'un nombre fini de capsules réparties sur une sphère revient à effectuer un échantillonnage spatial du champ sonore, de la même façon que le signal temporel est discrétisé lors de la conversion analogique/numérique. Cela implique un phénomène de repliement spatial en hautes fréquences, lorsque la longueur d'onde λ devient inférieure à deux fois la distance caractéristique inter-capsule d. À ce moment là, l'encodage ambisonique n'est plus valide. Lors de la conception d'une antenne, il est donc judicieux de maintenir d le plus faible possible, soit en réduisant le diamètre de l'antenne, soit en augmentant le nombre de capsules, afin de profiter de l'encodage spatial en hautes fréquences. Les équations liées à l'échantillonnage spatial du champ sonore et l'erreur introduite par celui-ci sont détaillées dans la thèse de S. Moreau [20].
À l'autre extrémité du spectre, le fait d'avoir $\lambda \gg d$ rend problématique l'encodage des basses fréquences. Hormis la composante omnidirectionnelle, les canaux ambisoniques sont des gradients de pression, ne pouvant être calculés précisément qu'en présence d'une différence de marche entre les capsules. Lorsque cette différence n'est plus significative, l'erreur d'estimation du gradient dégrade le rapport signal-sur-bruit. De plus, les capsules microphoniques tendent à devenir omnidirectionnelles en basses fréquences, la différence entre les pressions mesurées tend à devenir nulle. Un encodage correct en basses fré-

quences nécessiterait donc un augmentation du rayon de l'antenne ainsi qu'un nombre important de capsules pour améliorer le rapport signal-à-bruit.
En résumé, il est presque impossible d'obtenir un encodage correct à la fois des hautes et basses fréquences, car le rayon de l'antenne doit être dans un cas minimisé et dans l'autre maximisé. Il existe à l'état de prototypes des microphones essayant de contourner ces limitations, constitués par exemple de double sphères imbriquées l'une dans l'autre : la sphère intérieure pour les hautes fréquences, et la sphère de diamètre plus important pour les plus basses fréquences. La figure 1.6 représente le microphone développé par A. Party et al. à l'université de Sidney [21]. Les deux sphères de 32 capsules chacune permettent un encodage ambisonique d'ordre 4 avec des directivités constantes entre 900 Hz et 16 kHz. Néanmoins, ce genre de procédé reste coûteux et compliqué à mettre en place en raison du nombre de capteurs et de la géométrie de l'antenne.

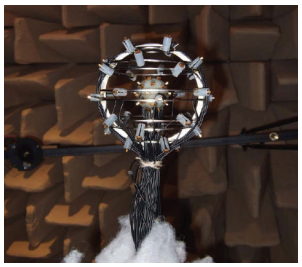

FIGURE 1.6. – Microphone ambisonique d'ordre 4 composé de 32 capteurs sur une sphère rigide intérieure et 32 capteurs sur une structure sphérique de rayon plus grand [21].

1.2.4.2. Description du microphone utilisé

Le système de captation utilisé durant la thèse pour générer les contenus ambisoniques est un modèle commercial mis au point par la société Mh Acoustics et nommé Eigenmike. Celui-ci est constitué d'une sphère rigide de 8.4 cm de diamètre à la surface de laquelle sont disposées 32 capsules à électret (figure 1.7), permettant de générer un contenu ambisonique allant jusqu'à l'ordre 4 (25 canaux). L'emplacement des capsules correspond au centre des faces d'un polyèdre semi-régulier à 32 faces. Les caractéristiques techniques fournies par le constructeur sont les suivantes :

— sensibilité des capsules : 30 mV/Pa,

— niveau maximum supporté par les capsules : 130 dB SPL,

— niveau de bruit des capsules : 17 dBA,

— bande-passante (théorique) du microphone : 0-20 kHz.

Les signaux capsules sont convertis au format HOA par une matrice de filtres élaborée au sein d'Orange Labs par J. Daniel et S. Moreau.

FIGURE 1.7. – Photographie du microphone Eigenmike de Mh acoustics [22], constitué de 32 capsules et permettant une captation HOA jusqu'à l'ordre 4.

1.2.4.3. Caractérisation du système de captation ambisonique

L'encodage spatial obtenu à partir du microphone a été caractérisé à partir de réponses impulsionnelles mesurées en chambre anéchoïque par des chercheurs de l'IRCAM (figure 1.8). Ainsi, 1717 mesures ont été réalisées sur un maillage de l'espace avec une résolution angulaire d'environ 6° en azimuth et 6° en élévation. Pour chaque point de maillage, la procédure est la suivante :

1. Un sinus glissant entre 0.1 et 24 kHz d'une durée de 1.37 s est diffusé par un haut-parleur ELAC 301 situé à 2 m du microphone.
2. Les 32 signaux capsules sont convolués par le sinus inverse pour obtenir la réponse impulsionnelle du format A.
3. Le format A est convolué par la matrice de filtres permettant le passage au format HOA 25 canaux.

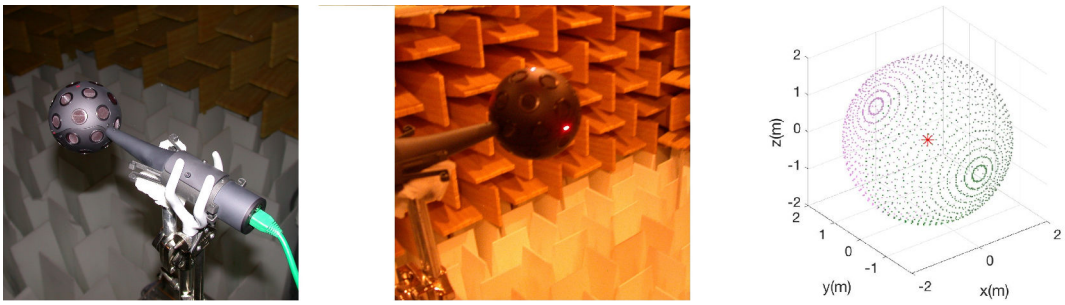

FIGURE 1.8. – g, c : Microphone en chambre anéchoique à l'IRCAM. d : Maillage des points de mesures (microphone au centre).

La réponse en fréquence du canal omnidirectionnel pour les directions d'arrivée (0°, 90°), (0°, 0°) et (0°, -120°) est décrite par la figure 1.9. Les réponses des différentes directions apparaissent parfaitement en phase. De plus, la phase est linéaire sur l'ensemble du spectre, à partir de 200 Hz. Cette fréquence correspond également à la fréquence de coupure basse à -3 dB du spectre d'amplitude par rapport à sa valeur moyenne. En hautes fréquences, la captation cesse d'être omnidirectionnelle aux alentours de 8 kHz, qui

correspond à la valeur théorique pour laquelle l'échantillonnage spatial devient insuffisant.

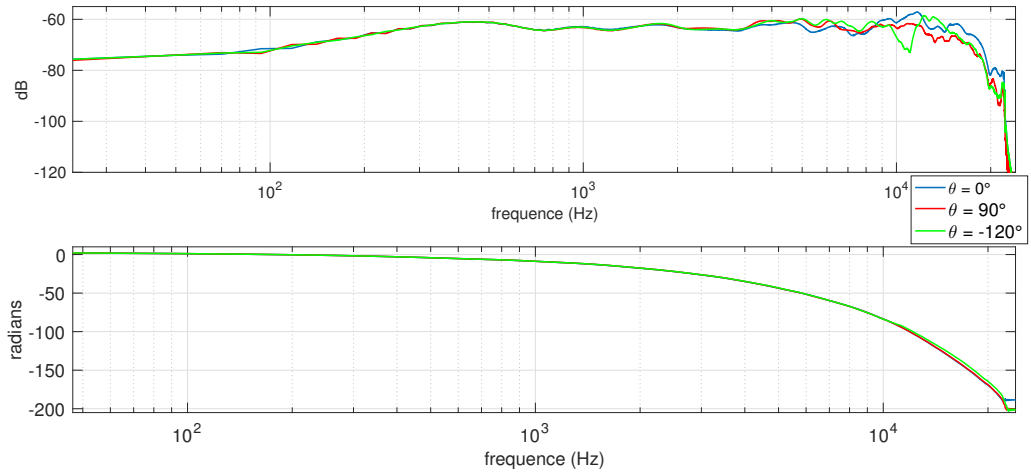

FIGURE 1.9. – Réponse en fréquence du canal omni-directionnel pour des sources positionnées à respectivement (0°, 90°), (0°, 0°) et (0°, -120°). En haut : gain. En bas : phase.

La figure 1.10 offre une représentation de la directivité synthétisée (dans le plan horizontal) de la première composante de chaque ordre ambisonique, en fonction de la fréquence. La première ligne correspond au canal W omnidirectionnel. Les limitations de l'encodage en basses fréquences (à gauche) sont visibles en dessous d'environ 200 Hz, tandis que le repliement spatial apparaît nettement au dessus de 8 kHz (au centre et à droite). A noter que pour chaque figure la référence en dB a été choisie de façon à optimiser l'affichage polaire, seules les différences entre les courbes du même graphique sont donc significatives. De façon générale, on peut déjà faire à la vue de ces graphiques deux principales observations :

— La limite haute de l'encodage valide est à peu près constante quel que soit l'ordre (8 kHz), et correspond à la fréquence à partir de laquelle le repliement spatial dû à l'espacement inter-capsules apparaît.

— En basses fréquences, la limite de validité de l'encodage est de plus en plus haute lorsque l'on monte en ordre. La raison est que la synthèse de directivités de plus en plus complexes avec le même nombre de capteurs entraîne mécaniquement une baisse du rapport signal à bruit, en particulier lorsque la longueur d'onde devient importante devant la dimension de l'antenne [23], [20].

On constate également une dynamique assez faible entre les maxima et les minima de directivité, ainsi qu'une légère déformation des lobes, et ce particulièrement dans les ordres élevés. Cela est dû en grande partie au nombre limité de points de mesures disponibles, ne permettant pas de représenter précisément des variations rapides de directivité. On utilise en effet ici 59 points de mesures dans le plan azimutal, permettant donc une résolution angulaire de 6°.

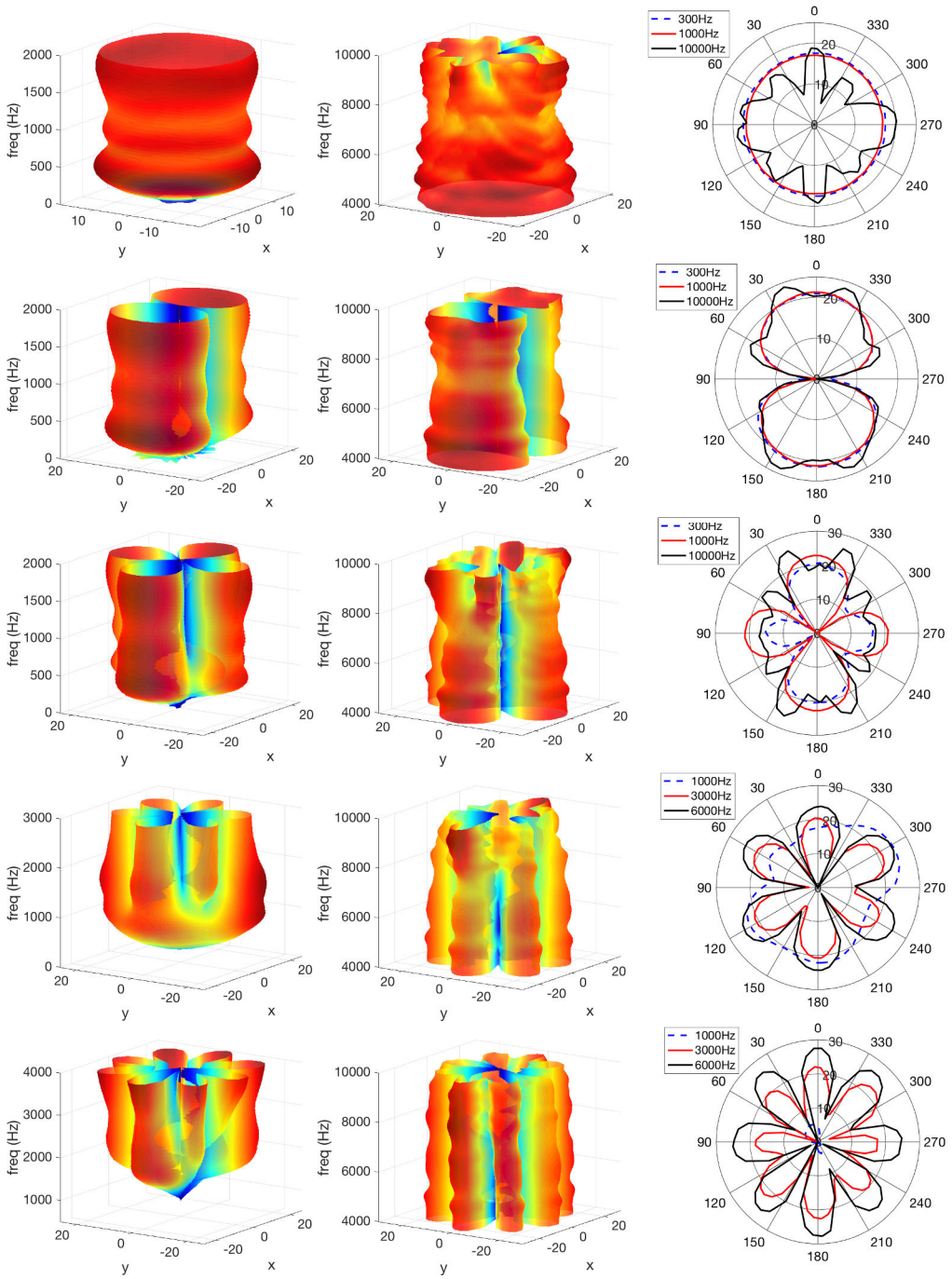

FIGURE 1.10. – Directivité réelle en fonction de la fréquence des composantes ambisoniques synthétisées (plan horizontal, dB) pour le microphone Eigenmike. De haut en bas : premières composantes des ordres 0 à 4. Gauche/centre : basses/hautes fréquences. Droite : Coupes transversales.

Deux critères objectifs sont utilisés pour évaluer la validité de la représentation ambisonique, en fonction de l'ordre et de la fréquence. La corrélation spatiale entre la directivité réelle et la directivité théorique est utilisé dans [23] pour caractériser l'encodage de chaque composante, soit :

$$C_{mn}^{\sigma}(f) = \frac{\sum_{k=1}^{K} \mathcal{D}_{mn}^{\sigma}(f,\theta_k,\phi_k) Y_{mn}^{\sigma}(\theta_k,\phi_k)}{\| \mathcal{D}(f,\theta_k,\phi_k) \| \| Y_{mn}^{\sigma}(\theta_k,\phi_k) \|} \tag{1.19}$$

où Y_{mn}^{σ} est la fonction harmonique sphérique décrivant la directivité théorique de la composante étudiée, $\mathcal{D}_{mn}^{\sigma}$ la directivité réelle, (θ_k, ϕ_k) la position associée au point de mesure k et K est le nombre total de mesures. On définit ici la norme $\| \cdot \|$ telle que :

$$\| \mathcal{D} \| = \sqrt{< \mathcal{D}, \mathcal{D} >} \tag{1.20}$$

avec l'opération $< ., . >$ désignant le produit scalaire.
On peut naturellement en déduire un estimateur de la validité de l'encodage d'ordre m comme étant la moyenne des scores des composantes de cet ordre, soit :

$$C_m(f) = \frac{1}{2m+1} \sum_{n\sigma} C_{mn}^{\sigma}(f) \tag{1.21}$$

La figure 1.11 présente les scores de corrélation spatiale sur le graphique du haut, tandis que le graphique du bas indique le niveau énergétique moyen, en prenant comme référence l'énergie du canal omnidirectionnel à 1 kHz. L'objectif est alors d'obtenir une directivité proche de la référence, tout en conservant un spectre suffisant plat. Les mesures valident les observations précédentes :

- L'encodage spatial se dégrade en basses fréquences et est couplé à une baisse générale du niveau d'énergie en fonction de l'ordre. L'amplification observée sur les ordres 0, 1 et 2 entre 200 et 600 Hz correspond à un *bass boost*, fonction de l'ordre, appliqué lors de l'encodage pour éviter une chute trop rapide du niveau énergétique.
- En hautes fréquences, le repliement spatial dégrade l'encodage au delà d'environ 8 kHz, avec une chute significative de l'énergie au-delà de 12 kHz. On remarque également que la limite haute de l'encodage diminue légèrement pour les ordres supérieurs. Cela est dû au fait qu'il est difficile en hautes fréquences de synthétiser des directivités complexes avec un nombre de capteurs limité. Plus de détails sur ce point sont disponibles dans [23] et [20].

Le tableau 1.2 récapitule les scores obtenus en termes de précision de l'encodage directionnel et de niveaux énergétiques ainsi que les bandes de fréquences que l'on a retenues pour l'exploitation du formalisme ambisonique avec ce microphone. Malgré des scores corrects jusqu'à 10 kHz, les ordres 1 et 2 ne sont conservés que jusqu'à 8 kHz pour éviter tout phénomène de repliement.

1.2.5. Restitution des contenus ambisoniques

La problématique de la diffusion des contenus HOA n'est pas au centre de la thèse, aussi cette section aborde seulement brièvement les deux principaux modes de restitution

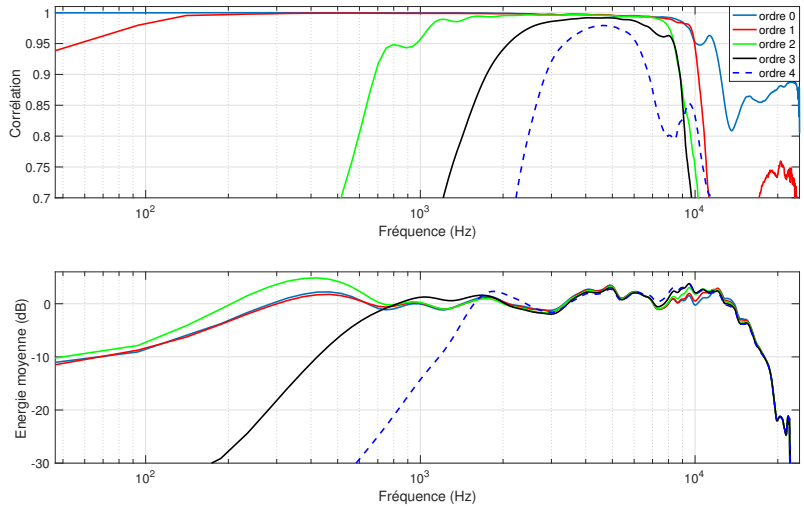

FIGURE 1.11. – En haut : Corrélation de la directivité ambisonique synthétisée avec la directivité théorique, en fonction de l'ordre ambisonique. En bas : Energie moyenne (dB, ref : canal omni à 1 kHz).

Ordre	Corr. > 0.95		Gain > -3 dB		Bande de validité	
	Freq. min	Freq. max	Freq. min	Freq. max	Freq. min	Freq. max
0	0	10 000	200	14 000	**200**	**8 000**
1	20	10 000	200	14 000	**200**	**8 000**
2	800	8 500	150	14 000	**800**	**8 000**
3	2 100	8 500	610	14 000	**2 100**	**8 000**
4	3 400	6200	1 400	14 000	**3 400**	**6 200**

TABLEAU 1.2. – Synthèse de la caractérisation de l'encodage ambisonique réalisé. Gauche : corrélation entre directivité réelles et théorique > 0.95. Centre : énergie moyenne > -3 dB (référence : omni à 1 kHz). Droite : plage de validité retenue pour l'analyse de scène ambisonique.

spatialisée : la restitution sur haut-parleurs et la restitution binaurale.

1.2.5.1. Diffusion sur haut-parleurs

L'équation 1.16 nous renseigne sur la possibilité de générer, à partir d'un contenu ambisonique et sur une certaine étendue (*sweet spot*), une scène sonore approchant le champ acoustique d'origine. En pratique, cela est envisageable en utilisant un jeu de haut-parleurs régulièrement répartis autour de l'auditeur. Celui-ci peut alors se déplacer dans la scène sonore en conservant un rendu spatial cohérent, dans les limites du *sweet spot*. La figure 1.12 représente la salle d'écoute mise en place à Orange Labs dans le cadre de la thèse. Celle-ci est constituée de 12 enceintes dans le plan horizontal régulièrement

espacées de 30°, ainsi que de 4 enceintes situées à 30° d'élévation et espacées de 45° en azimuth. De manière générale, les conditions d'une restitution spatiale cohérente sont :

- un nombre de haut-parleurs au moins égal au nombre de composantes HOA décodées,
- une répartition des haut-parleurs la plus régulière possible,
- le positionnement de l'auditeur au *sweet spot*.

FIGURE 1.12. – Salle de restitution 3D à Orange Labs, 16 haut-parleurs (configuration 16.0).

Le décodage sur haut-parleurs d'un contenu HOA peut faire appel à différentes stratégies, dépendant notamment de la disposition plus ou moins régulière du système de diffusion et de la qualité de l'encodage microphonique. On nomme ici $\mathbf{D}$ la matrice de décodage dont chaque terme D_{ij} contient le gain appliqué à la i^{e} composante ambisonique diffusée sur le haut-parleur j. Dans le cas d'un contenu ambisonique $\mathbf{s}(t)$ encodé parfaitement sur N canaux et d'un système de diffusion régulier de N haut-parleurs disposés sur une sphère de rayon suffisamment important, $\mathbf{D}$ de dimensions $N \times N$ est calculée telle que le signal multicanal $\mathbf{s_d}(t)$ diffusé par les enceintes puis enregistré au centre du dispositif avec un microphone ambisonique virtuel serait identique au signal initial [24]. En considérant le signal émis par chaque haut-parleur i placé à la position (θ_i, ϕ_i) comme une onde plane, on doit obtenir au centre du dispositif une captation virtuelle telle que :

$$\mathbf{A}\mathbf{s_d}(t) = \mathbf{s}(t) \tag{1.22}$$

avec $\mathbf{A}$ une matrice de coefficients harmoniques sphériques de dimensions $N \times N$ dont la j^{e} colonne correspond aux coefficients d'encodage $[Y_{00}^1(\theta_j, \phi_j), \cdots, Y_{mn}^{\sigma}(\theta_j, \phi_j)]^T$ du signal provenant du j^{e} haut-parleur. En exprimant le signal diffusé comme étant le produit de la matrice de décodage et du contenu initial, on obtient alors :

$$\mathbf{A}\mathbf{D}\mathbf{s}(t) = \mathbf{s}(t) \tag{1.23}$$

On a alors l'expression de la matrice décodage :

$$\mathbf{D} = \mathbf{A}^{-1} \tag{1.24}$$

Dans un cas idéal, la matrice de décodage d'un contenu ambisonique est donc la matrice inverse de la matrice des coefficients harmoniques sphériques relatifs à la position des haut-parleurs. Lorsque la disposition des haut-parleurs n'est pas régulière, ou que le nombre de haut-parleurs diffère du nombre de composantes HOA, le calcul d'une matrice de décodage se fait généralement en optimisant des paramètres physiques, comme le vecteur vélocité, ou encore des paramètres psycho-acoustiques. Plus de considérations sur ce sujet sont développées dans [23] et [24].

1.2.5.2. Diffusion binaurale

L'écoute au casque est un autre moyen de profiter d'un contenu sonore immersif, grâce à l'utilisation de filtres binauraux (HRTFs pour *head-related transfer functions*). Les HRTFs reproduisent, pour un signal sonore ayant une incidence donnée, les modifications du champ sonore perçues au niveau des oreilles de l'auditeur induites par l'obstacle que représentent la tête et le buste de celui-ci. Ces fonctions contiennent ainsi les indices de localisation que sont le retard interaural (différence de temps de propagation entre les deux oreilles), la différence de niveau interaurale et la coloration spectrale du signal. Celles-ci sont différentes pour chaque individu, le rendu spatial sera donc différent suivant que l'auditeur ait accès à ses propres filtres binauraux mesurés en chambre sourde ou bien qu'on utilise des HRTFs génériques ou appartenant à un autre individu.
Une façon d'écouter un contenu ambisonique au format binaural est de faire un décodage sur des haut-parleurs virtuels (comme décrit dans la section précédente), puis de procéder à un filtrage du signal de chaque haut-parleur virtuel par les HRTFs associées à la position de celui-ci.

1.3. Environnement acoustique

L'hypothèse de propagation du son en champ libre a été faite jusqu'à présent car elle permet une expression relativement simple des phénomènes acoustiques. Cependant, celle-ci est rapidement mise en défaut dans un cas réel. En effet, même dans un environnement ouvert, le sol reste souvent une surface réfléchissante, de même que la présence d'objets ou de personnes vient perturber la propagation de l'onde sonore. En environnement fermé, l'interaction de l'onde sonore avec les parois de la pièce et avec les éléments présents à l'intérieur est nommé effet de salle, plus ou moins important en fonction de la géométrie et des matériaux qui composent celle-ci. La modélisation, l'identification et la manipulation de l'effet de salle sont des problématiques mises en avant depuis de nombreuses années et toujours d'actualité. Les travaux réalisés durant la thèse ne se sont pas focalisés sur l'identification de l'effet de salle à partir d'un enregistrement ambisonique, cependant c'est un paramètre qui influe sur les performances de l'analyse de scène et qu'il est essentiel de prendre en compte pour évaluer la robustesse et les limites des algorithmes mis au point.

1.3.1. Modélisation d'une réponse de salle

On caractérise un effet de salle en identifiant une réponse impulsionnelle entre une source sonore et un point de mesure. On peut mesurer une réponse omnidirectionnelle (RIR pour *room impulse response*), ou bien une réponse spatiale (SRIR pour *spatial room impulse response*) dans le cas où la réponse dépend de la directivité du capteur (ce qui est le cas avec un microphone ambisonique). On décompose classiquement une réponse impulsionnelle de salle comme un signal comportant trois parties :

— Le champ direct, qui correspond au premier front d'onde empruntant le trajet direct source/microphone,

— Les premières réflexions qui correspondent à des sources-images issues de la réflexion du son sur des parois,

— La réverbération tardive, souvent modélisée comme un champ sonore diffus, se traduisant par une densité d'énergie acoustique statistiquement uniforme dans l'ensemble du volume.

La figure 1.13 illustre une réponse impulsionnelle de salle sur laquelle des premières réflexions sont clairement identifiables, tandis que la figure 1.14 représente la réponse impulsionnelle spatiale du format B (ordre 1), enregistrée avec l'Eigenmike pour une source positionnée à $(90°, 0°)$ par rapport à l'axe du microphone. Le champ direct intervient aux alentours de 23 ms, avec une amplitude maximale de 0 dB pour la composante Y qui a son maximum de directivité dans l'axe de la source (valeur prise comme référence pour l'affichage en dB), alors que celle-ci est d'environ -20 dB pour les canaux X et Z qui possèdent des minima de directivité dans cette direction. Les différence énergétiques entre les différents canaux tendent ensuite à s'atténuer, au fur et à mesure que la contribution tardive de l'effet de salle devient omnidirectionnelle et diffuse.

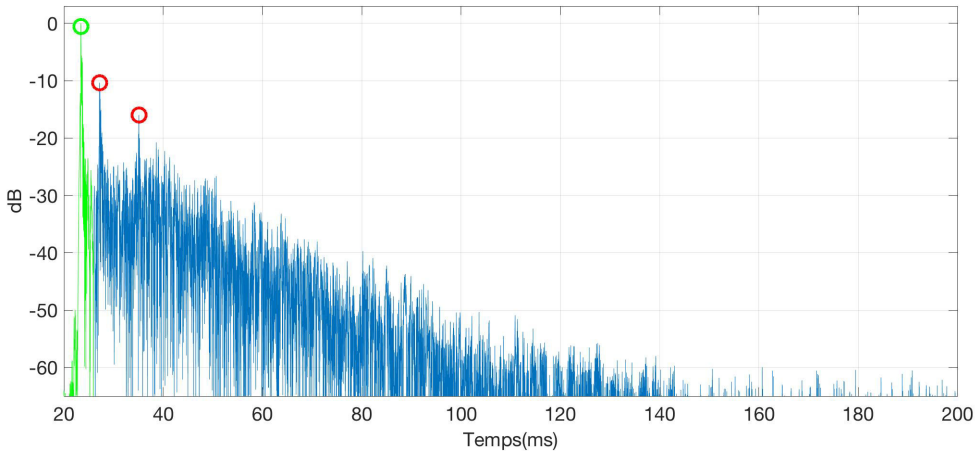

FIGURE 1.13. – Réponse impulsionnelle de salle omnidirectionnelle mesurée avec le microphone Eigenmike. En vert : champ direct. En bleu : champ réverbéré. Cercle vert : premier front d'onde. Cercles rouges : premières réflexions.

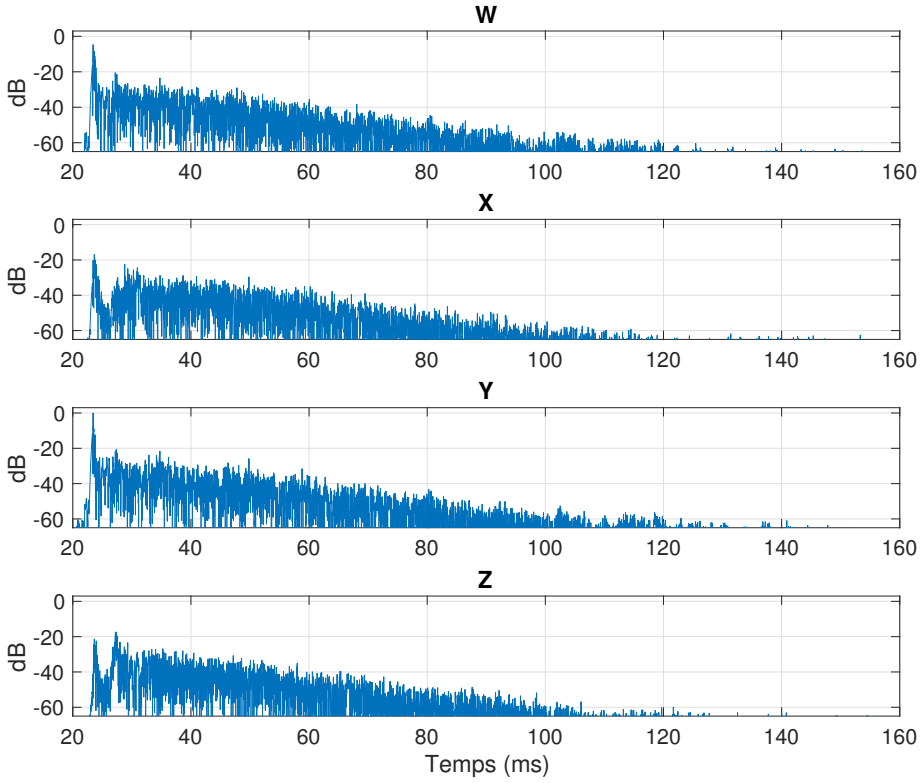

FIGURE 1.14. – SRIR ambisonique d'ordre 1 mesurée avec le microphone Eigenmike pour une source positionnée à $(90°, 0°)$ par rapport au microphone (amplitude normalisée, en dB)

.

1.3.2. Indices de caractérisation de l'effet de salle

Il existe différents indicateurs permettant de caractériser l'effet de salle, suivant que l'on s'intéresse à la durée de celui-ci (temps de réverbération), à l'énergie des premières réflexions, à l'intelligibilité de la parole, etc. On utilisera par la suite le temps de réverbération TR_{60} qui correspond au temps que met l'énergie acoustique pour décroître de 60 dB dans une salle en l'absence d'excitation. Cet indicateur est directement lié au volume de la salle via la loi de Sabine [14]. La décroissance du niveau sonore en dB au cours du temps est considérée comme linéaire pour la réverbération tardive, aussi le TR_{60} est le plus souvent obtenu en extrapolant le TR_{20} ou le TR_{30} car il est rare d'atteindre 60 dB de rapport signal à bruit sur la mesure de la réponse de salle. La procédure pour mesurer le temps de réverbération est encadrée par les normes ISO 3382-1 :2009 et ISO 3382-2 :2010.

1.3.3. Intensité acoustique

La direction de propagation de l'énergie acoustique est modifiée par l'effet de salle, de par les multiples réflexions qui se superposent et la réverbération tardive qui noie l'information spatiale, comme illustré par la figure 1.14. Comme nous le verrons dans le chapitre suivant, la direction de propagation de l'énergie est un élément utilisé couramment pour déterminer la direction d'incidence d'une source sonore.
On pose ici l'équation de bilan de l'énergie acoustique :

$$\frac{\partial e}{\partial t} + \mathrm{div}(\vec{I}) = 0 \tag{1.25}$$

avec $e = \frac{\rho_0}{2}\vec{v}^2 + \frac{1}{2\rho_0 c_0^2}p^2$ l'énergie acoustique par unité de volume et $\vec{I} = p\vec{v_a}$ l'intensité acoustique instantanée.
L'intégration de cette expression sur un volume V fait apparaître la formule suivante :

$$\int\int\int_V \frac{\partial e}{\partial t} dV + \int\int\int_V \mathrm{div}(\vec{I}) dV = 0 \tag{1.26}$$

dont le second membre peut s'écrire sous la forme d'une intégrale de surface, soit :

$$\frac{\partial}{\partial t}\int\int\int_V e dV + \int\int_{\partial V} \vec{I}\vec{dS} = 0 \tag{1.27}$$

où $\vec{dS}$ est le vecteur de la surface élémentaire dS pointant vers la normale extérieure du volume $\vec{n}_{ext}$.On établit ainsi que l'intensité instantanée correspond au flux d'énergie acoustique par unité de surface.
En notation complexe, l'intensité moyenne dite aussi intensité active est égale à [14] :

$$\vec{I} = \frac{1}{2}\Re\{p\vec{v}*\} = \frac{1}{2}|p||\vec{v}|\cos\phi_p - \phi_v\vec{n}_{ext} \tag{1.28}$$

avec ϕ_p et ϕ_v les phases respectives de la pression et de la vitesse particulaire. On lui associe un terme appelé intensité réactive, valant :

$$\vec{J} = \frac{1}{2}\Im\{p\vec{v}*\} = \frac{1}{2}|p||\vec{v}|\sin\phi_p - \phi_v \cdot \vec{n}_{ext} \tag{1.29}$$

pour former un vecteur d'intensité complexe :

$$\vec{\prod} = \vec{I} + i\vec{J} \tag{1.30}$$

L'intensité active décrit alors le flux d'énergie acoustique se propageant à travers une surface élémentaire, alors que l'intensité réactive reflète un échange d'énergie local non propagatif. Pour une onde sonore se propageant en champ libre, l'intensité réactive est nulle en champ lointain, alors qu'à l'inverse c'est l'intensité active qui est nulle en présence d'un phénomène stationnaire (résonance).

1.4. Conclusion

Ce chapitre a permis d'expliciter le développement en harmoniques sphériques du champ acoustique, afin d'introduire le concept d'ambisonie. Les principales caractéristiques de cette représentation (observations coïncidentes, directivités théoriques constantes suivant la fréquence...) mais également les limites de cette représentation dans le cadre d'une captation réelle ont été présentées et seront par la suite rappelées lorsqu'il s'agira d'appliquer l'analyse de scène à des contenus ambisoniques. Dans le chapitre suivant, des méthodes de séparation de sources pouvant s'appliquer au format ambisonique sont décrites, dans le cas d'un encodage respectant les propriétés théoriques énoncées dans cette partie.

2. Etat de l'art sur l'analyse de scène multi-capteurs

Dans le cadre de cette thèse, l'analyse de scène sonore a pour but d'apporter une réponse aux questions suivantes :

- Combien de sources sonores sont présentes dans un contenu ambisonique (comptage) ?
- Où sont-elles situées (localisation) ?
- Quelles sont-elles (séparation) ?
- Comment évoluent-elles au cours du temps (suivi) ?

On voit que deux problématiques s'imbriquent alors : la cartographie de la scène sonore et l'extraction donc la séparation des sources identifiées. Pour établir une cartographie, il est nécessaire de connaître les caractéristiques spatiales et fréquentielles des capteurs à disposition (directivité, espacement inter-capteurs) et également de faire des hypothèses sur le milieu de propagation (milieu homogène ou hétérogène, champ libre ou milieu réverbérant, etc.). En fonction des signaux à disposition et de la problématique à traiter, un minimum d'hypothèses doit donc être posé. Les hypothèses posées systématiquement par la suite sont les suivantes :

- On dispose d'un contenu ambisonique encodé parfaitement sur toute la bande spectrale utilisée. Les observations sont donc coïncidentes et respectent les propriétés de directivité énoncées dans la partie précédente.
- Les différentes sources ont des directions d'arrivée différentes.
- Les sources sont considérées comme suffisamment éloignées du microphone pour être vues comme des ondes planes. De plus, étant donné qu'il n'y a pas de différence de phase entre les observations, l'encodage est invariant suivant la distance entre la source et le microphone et l'on ne cherchera pas à estimer celle-ci.
- Le nombre de sources est inférieur ou égal au nombre d'observations, en tout cas dans la bande de fréquences visée et sur l'intervalle de temps étudié. Cette hypothèse peut en pratique être mise en défaut, elle est cependant souvent vérifiée, en particulier lorsque l'on utilise des ordres ambisoniques supérieurs (9 canaux à l'ordre 2, 16 canaux à l'ordre 3).
- La position des sources évolue potentiellement au cours du temps, mais leur déplacement est petit au regard de la durée d'une trame d'analyse.

D'après les hypothèses que l'on vient d'énumérer, l'encodage microphonique de N sources en champ libre se formule sous la forme d'un mélange instantané de variables temporelles

discrètes, soit :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2.1}$$

où $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_M(t)]^T$ représente le vecteur des M observations ambisoniques, $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_N(t)]^T$ le vecteur des sources et $\mathbf{A}$ la matrice de mélange réelle de dimensions $M \times N$. Les observations $x_i(t)$ sont ainsi obtenues par la formule :

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \tag{2.2}$$

où a_{ij} est la contribution de la source j à l'observation i. Les coefficients de mélange relatifs à la source $s_j(t)$ d'incidence (θ_j, ϕ_j) s'expriment grâce aux fonctions harmoniques sphériques $Y_{mn}^{\sigma}(\theta_j, \phi_j)$, ce qui donne pour chaque observation :

$$x_i(t) = \sum_{j=1}^{N} Y_{mn}^{\sigma}(\theta_j, \phi_j) s_j(t) \tag{2.3}$$

les indices m, n et σ dépendant de i.
En présence de réverbération, le mélange devient convolutif, faisant intervenir une réponse impulsionnelle propre à chaque couple source-observation. L'équation du mélange devient donc :

$$x_i(t) = \sum_{j=1}^{N} a_{ij} * s_j(t) = \sum_{j=1}^{N} \sum_{\tau=0}^{T-1} a_{ij}(\tau) s_j(t-\tau) \tag{2.4}$$

où l'opérateur $*$ désigne le produit de convolution et T est la longueur de la réponse impulsionnelle. Un mélange convolutif peut également être ramené à un mélange instantané en opérant dans le domaine fréquentiel, avec pour chaque bande de fréquence des coefficients de mélange complexes.
La littérature foisonne de méthodes d'analyse de scène et de séparation de sources. Celles-ci peuvent être aveugles [25] [26] ou avec des connaissances préalables [27] [28], traitant des mélanges convolutifs [29], utilisant des factorisations en matrices non-négatives [30] ou encore des réseaux de neurones [31]. Une approche basée sur la séparation aveugle de sources (SAS) nous a semblé pertinente pour plusieurs raisons :

— Les sources à identifier ne sont pas connues *a priori*.

— L'encodage ambisonique de sources en champ libre peut être modélisé par une matrice de mélange instantanée, dont les termes correspondent aux coefficients ambisoniques associés à chaque source. Ce modèle correspond parfaitement au formalisme de la SAS.

— Dans le cas d'un mélange instantané déterminé, la SAS permet l'estimation à la fois de la matrice de séparation et de la localisation des sources grâce à la matrice de mélange, inverse de la matrice de séparation. Ainsi, les aspects cartographie et séparation de sources sont traités simultanément.

Initiée par Herault et Jutten [32] dans les années 80, la séparation aveugle de sources regroupe un ensemble de méthodes partageant un concept commun : identifier, à partir d'un jeu d'observations, les différentes sources qui les constituent et les coefficients de mélange associés, avec le moins possible de connaissances *a priori* sur la nature des sources. En ce qui nous concerne, il s'agit d'identifier, à partir d'un contenu ambisonique (appelé ensuite mélange), les source sonores enregistrées et les coefficients d'encodage associés, à partir desquels une estimation des directions d'arrivée (DOA pour *Direction Of Arrival*) est obtenue. La résolution du cas instantané fut historiquement la première à être étudiée, mais rapidement les algorithmes existants ont été étendus au cas convolutif (voir notamment les travaux de Shalvi et Weinstein [33], Loubaton et Regalia [34] ou Back [35]). Makino *et al.* [29] proposent une revue des méthodes de séparation de sources pour des mélanges convolutifs, appliquées au traitement de la parole. Le choix de la méthode d'analyse va varier suivant le type de mélange : instantané ou convolutif, sur- ou sous-déterminé. Lorsque le nombre de sources est supérieur au nombre d'équations, on a affaire à un système sous-déterminé. La matrice de mélange n'étant pas directement inversible, différents techniques permettent cependant de se rapprocher d'une solution exploitable. On peut citer des méthodes de régularisation sous contrainte [36] [37], dans lesquelles on détermine un critère à optimiser pour converger vers une solution (par exemple favoriser la solution qui explique les observations au moyen du plus petit nombre de sources), souvent associées à des méthodes parcimonieuses, basées sur l'hypothèse que les sources ont des supports temps/fréquence distincts [38] [8]. La parcimonie fréquentielle ou temporelle permet ainsi se rapprocher d'un cas déterminé ou sur-déterminé. Dans le cas d'un mélange bruité, la caractérisation du système se fait en identifiant deux sous-espaces : le sous-espace signal et le sous-espace bruit. De tels cas sont abordés notamment par J.-F. Cardoso et F. Souloumiac [39] ou Murata *et al.* [40].
Les deux premières parties de ce chapitre proposent un panorama des méthodes de SAS avec une description des principaux algorithmes existants. La troisième partie aborde les différentes approches d'analyse de scène existantes dédiées à des contenus ambisoniques. Enfin, la dernière section se penche sur la formalisation du *beamforming* pour la séparation de sources et les problématiques associées.

2.1. Séparation aveugle de sources - méthodes parcimonieuses

Méthode DUET L'approche DUET (*Degenerate Unmixing Estimation Technique*), formulée par A. Jourjine [41], permet de localiser et extraire N sources en conditions anéchoïques à partir de seulement deux observations non coincidentes, en faisant l'hypothèse que les sources ont des supports fréquentiels disjoints, soit

$$S_i(f)S_j(f) = 0 \tag{2.5}$$

pour tout f dès lors que $i \neq j$.
Après une décomposition des observations en sous-bandes fréquentielles, une amplitude

a_i et un retard t_i sont estimés pour chaque sous-bande en se basant sur l'équation de mélange théorique :

$$\begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-i\omega t_1} & \cdots & a_N e^{-i\omega t_N} \end{bmatrix} \cdot \begin{bmatrix} S_1(f) \\ \cdots \\ S_N(f) \end{bmatrix} \tag{2.6}$$

Dans chaque bande de fréquence f, un couple (a_i,t_i) correspondant à la source i active est estimée de la façon suivante :

$$\begin{cases} a_i &= \left|\left|\frac{X_2(f)}{X_1(f)}\right|\right| \\ t_i &= \frac{1}{2\pi f}\Im\{\log \frac{X_2(f)}{X_1(f)}\} \end{cases} \tag{2.7}$$

Une représentation de l'ensemble des couples (amplitude, retard) sous forme d'histogramme permet de repérer les zones contenant des sources actives, dont les contributions sont délimitées par une étape de *clustering*. L'extraction des sources se fait alors par inversion de la matrice de mélange estimée lorsque le nombre de sources est inférieur ou égal à deux, ou bien par masquage fréquentiel dans le cas sous-déterminé.

Méthode DEMIX DEMIX (Arberet *et al.* [42]) est une approche également basée sur une hypothèse de parcimonie temps-fréquence des sources dans un contenu stéréo, pour des observations coïncidentes ou non en l'absence de réverbération. Cette fois, une direction d'arrivée est estimée dans chaque bande de fréquence, à partir d'une analyse en composantes principales (ACP) appliquée à un agrégat de points temps-fréquences voisins. Le ratio entre l'énergie de la première valeur propre identifiée et l'énergie des valeurs propres suivantes est utilisé comme indicateur de confiance, à la fois pour améliorer la précision de localisation en mettant en avant les indicateurs les plus robustes, et également pour l'étape de reconstruction des sources large-bande par *clustering* dit séquentiel (voir [43]). Dans le cas anéchoïque, DEMiX a l'avantage de ne pas souffrir de l'ambiguité introduite par le repliement de la phase lors de l'estimation du délai et de fournir une localisation beaucoup plus précise.

2.2. Séparation aveugle de sources - méthodes statistiques

Les méthodes de SAS décrites dans cette section sont des approches probabilistes, basées sur les propriétés statistiques réelles ou supposées des signaux étudiés. La plupart du temps, celles-ci sont basées sur l'hypothèse que les sources à extraire sont indépendantes, ou tout du moins que leurs cumulants croisés sont nuls jusqu'à un certain ordre. On distinguent deux principales catégories :

— les méthodes tensorielles (diagonalisation de matrices de cumulants),

— celles basées sur la théorie de l'information et la minimisation de l'entropie des sources.

Lorsque l'hypothèse d'indépendance des sources est faite ou que les cumulants croisés sont supposés nuls au-délà du second ordre, on parle alors d'analyse en composantes indépendantes (ACI). L'analyse en vecteurs indépendants (AVI) [26] généralise l'ACI en séparant les sources conjointement sur différents jeux de données.
Les algorithmes exposés ici utilisent une formulation temporelle du problème, faisant donc intervenir des variables réelles, cependant la plupart peuvent être déclinés pour traiter des variables complexes pour s'adapter à une approche fréquentielle de la séparation de sources.

2.2.1. Notions de statistiques

Nous définissons ici les principales notions statistiques utilisées par la suite. Les développements mathématiques associés à la théorie de l'information sont accessibles dans [44] et [45], pour les estimateurs statistiques d'ordres supérieurs on peut se référer à l'ouvrage de J.-L. Lacoume *et al.* [46]. Dans tous les développements qui vont suivre, les variables sont supposées centrées, ce qui est le cas des signaux acoustiques traités. Par ailleurs les estimateurs sont exprimés dans le domaine temporel mais ceux-ci peuvent être généralisés à des variables complexes.

2.2.1.1. Densité de probabilité

La densité de probabilité $p_x(a)$ d'une variable aléatoire x est une fonction dont l'intégrale vaut 1 et qui décrit la répartition des valeurs prises par cette variable.
Est définie comme gaussienne une variable x dont la densité de probabilité $p_x(a)$ suit la loi suivante :

$$p_x(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{a^2}{2\sigma^2}} \tag{2.8}$$

avec σ l'écart-type de la variable. En séparation de sources, une hypothèse de gaussianité ou non-gaussianité des sources est souvent faite, car cela a une influence sur le choix des estimateurs statistiques utilisés, comme il est montré par la suite.

2.2.1.2. Entropie et information mutuelle

L'entropie est interprétée comme l'incertitude sur la valeur d'une variable aléatoire. L'entropie de la variable x, appelée aussi entropie marginale, notée $H(x)$, est définie par :

$$H(x) = -\int_{-\infty}^{\infty} p_x(a) \log p_x(a) da \tag{2.9}$$

où $p_x(a)$ désigne la densité de probabilité de la variable discrète x, c'est-à-dire ici la probabilité que la variable x prenne une valeur a donnée. Par mesure de lisibilité, on notera par la suite $p(x)$ la densité de probabilité de la variable x. On note qu'une variable gaussienne aura l'entropie maximale possible, parmi toutes les variables de même variance.

On exprime l'entropie conditionnelle de deux variables x_i et x_j, soit l'entropie de x_i connaissant x_j :

$$H(x_i|x_j) = H(x_i, x_j) - H(x_j) \tag{2.10}$$

où $H(x_i, x_j)$ est l'entropie conjointe :

$$H(x_i, x_j) = -\sum_{x_i \in \mathcal{X}_\rangle} \sum_{x_j \in \mathcal{X}_|} p(x_i, x_j) \log p(x_i, x_j) \tag{2.11}$$

On en déduit l'information mutuelle :

$$I(x_i; x_j) = H(x_i) - H(x_i|x_j) \tag{2.12}$$

Deux variables x_i et x_j sont indépendantes si $p(x_i, x_j) = p(x_i)p(x_j)$. Autrement dit, deux variables sont indépendantes si la valeur de l'une n'influe pas sur la valeur de l'autre. On définit alors l'information mutuelle comme la divergence de Kullback-Leibler entre $p(x_i, x_j)$ et $p(x_i)p(x_j)$, soit :

$$I(x_i; x_j) = \sum_{x_i \in \mathcal{X}_\rangle} \sum_{x_j \in \mathcal{X}_|} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{2.13}$$

L'indépendance des variables se traduit donc par $I(x_i, x_j) = 0$ et en termes d'entropie par :

$$H(x_i, x_j) = H(x_i) + H(x_j) \tag{2.14}$$

Ces notions sont généralisables à un vecteur $\mathbf{x}$ composé de N variables aléatoires.

2.2.1.3. Vraisemblance

On définit la vraisemblance L_x d'un processus aléatoire discret $x(t)$ de dimensions $1 \times T$ comme étant le produit des densités de probabilité associées à chaque valeur de x [47], soit :

$$L_x = \prod_{t=1}^{T} p(x(t)) \tag{2.15}$$

Lorsque la densité de probabilité de x est inconnue et que l'on cherche à approcher celle-ci, la vraisemblance sera maximale pour la fonction approchant le mieux $p(x)$. Celle-ci pouvant prendre des valeurs très faibles, on utilise en général le logarithme de la vraisemblance.

2.2.1.4. Covariance

La covariance de deux variables x_i et x_j est énoncée par la formule suivante :

$$r_{ij} = \mathrm{E}\{x_i x_j\} \tag{2.16}$$

où l'opérateur E{.} désigne l'espérance.
Pour un vecteur-colonne $\mathbf{x}$ de N variables x_j, on définit la matrice de covariance $\mathbf{R_x}$ carrée et symétrique de dimensions $N \times N$:

$$\mathbf{R_x} = \mathrm{E}\{\mathbf{x}\mathbf{x}^T\} \tag{2.17}$$

dont les composants sont les r_{ij} avec $1 \leqslant i, j \leqslant N$. Les termes diagonaux r_{ii} correspondent à la variance des variables x_i. Les variables x_i et x_j sont dites décorrélées si $r_{ij} = 0$.
La notion de matrice de covariance peut être généralisée pour estimer la corrélation entre un jeu de signaux et une version retardée de ceux-ci :

$$\mathbf{R_x}(\tau) = \mathrm{E}\{\mathbf{x}(t)\mathbf{x}(t-\tau)^T\} \tag{2.18}$$

où τ désigne le retard temporel. Dans la suite du document, pour des raisons de lisibilité, on écrira $\mathbf{R_x}$ en lieu et place de $\mathbf{R_x}(0)$.

2.2.1.5. Cumulants

Pour décrire les statistiques d'ordres supérieurs à 2, on introduit la notion de cumulant. Les cumulants sont des estimateurs statistiques permettant une description plus complète des données et de la relation liant un ensemble de variables entre elles [46]. Dans le cas d'une variable gaussienne, tous les cumulants d'ordre supérieur ou égal à 3 sont nuls, c'est pourquoi l'hypothèse de non-gaussianité des sources recherchées est posée lorsque l'on cherche à tirer partie des statistiques d'ordres supérieurs. Propriété souvent utilisée pour l'ACI, les cumulants croisés de variables indépendantes sont théoriquement nuls quel que soit l'ordre. L'expression du cumulant d'ordre n d'une variable x noté κ^n est obtenue en dérivant n fois en 0 la seconde fonction caractéristique de cette variable, soit :

$$\Psi_x(u) = \log \mathrm{E}\{e^{iux}\} \tag{2.19}$$

et

$$\kappa^n = (-i)^n \frac{\mathrm{d}^n \Psi_x(u)}{\mathrm{d}u^n} \Big|_{u=0} \tag{2.20}$$

Souvent utilisé en séparation de sources, on nomme kurtosis de la variable x_i son cumulant d'ordre 4, tel que :

$$\kappa_i^4 = \mathrm{Cum}^{(4)}(x_i, x_i, x_i, x_i) = \frac{\mathrm{E}\{x_i^4\}}{\mathrm{E}\{x_i^2\}^2} - 3 \tag{2.21}$$

Des variables possédant un kurtosis positif ou négatif sont appelées respectivement sur-gaussienne ou sous-gaussienne, alors que celui-ci est nul pour les variables gaussiennes.

Les cumulants croisés s'obtiennent par généralisation des équations 2.19 et 2.20 à des variables multidimensionnelles. Pour un jeu $\mathbf{x}$ constitué de N variables, le tenseur $\mathbf{Q}$, de dimensions $N \times N \times N \times N$, est appelé tenseur de quadricovariance et contient l'ensemble des cumulants d'ordre 4. Celui-ci se compose des termes Q_{ijkl} avec $1 \leqslant i, j, k, l \leqslant N$, tels que :

$$\begin{aligned} Q_{ijkl} = \mathrm{Cum}^{(4)}(x_i, x_j, x_k, x_l) = \mathrm{E}\{x_i x_j x_k x_l\} \\ - \mathrm{E}\{x_i x_j\}\,\mathrm{E}\{x_k x_l\} \\ - \mathrm{E}\{x_i x_k\}\,\mathrm{E}\{x_j x_l\} \\ - \mathrm{E}\{x_i x_l\}\,\mathrm{E}\{x_j x_k\} \end{aligned} \tag{2.22}$$

2.2.1.6. Analyse en composantes principales

L'analyse en composantes principales (ACP) appliquée à un jeu d'observations $\mathbf{x}$ consiste à trouver un ensemble de combinaisons linéaires, noté sous la forme matricielle $\mathbf{V}^T$, tel que les variables $\mathbf{z} = \mathbf{V}^T\mathbf{x}$ sont décorrélées, soit $\mathrm{E}\{z_i z_j\} = 0$ pour tout $i \neq j$. L'ACP opère une diagonalisation de la matrice de covariance $\mathbf{R_x}$, par une décomposition de celle-ci en valeurs propres et en vecteurs propres orthogonaux. $\mathbf{V}$ contient les vecteurs propres, tandis que les valeurs propres correspondent aux termes diagonaux de la matrice de covariance $\mathbf{D}$ des variables ainsi décorrélés. On obtient donc la relation :

$$\mathbf{R}_x\mathbf{V} = \mathbf{D}\mathbf{V} \tag{2.23}$$

ou encore

$$\mathbf{V}^T\mathbf{R}_x\mathbf{V} = \mathbf{D} \tag{2.24}$$

2.2.2. Méthodes basées sur les statistiques d'ordre 2

Nous nous intéressons ici au cas déterminé, c'est-à-dire un système comportant autant d'observations que de sources.

Séparation de signaux gaussiens non-stationnaires De façon générale, une simple étape de décorrélation ne suffit pas à séparer correctement des signaux indépendants. On peut cependant arriver à une extraction des sources lorsque des informations supplémentaires sont disponibles, et notamment lorsque la distribution statistique et la variance des signaux évoluent au cours du temps.
Des méthodes de séparation ont été mises au point uniquement à partir de statistiques du second ordre pour des signaux gaussiens non-stationnaires. Elles sont basées sur l'analyse conjointe des observations à différents instants, et nécessitent que la distribution statistique des sources varie dans le temps. Une des premières techniques de séparation de deux signaux non-stationnaires a été mise au point par Weinstein en se basant sur la minimisation des densités spectrales croisées à différents instants [48]. Pham et Cardoso décrivent

dans [49] une méthode s'appuyant sur la diagonalisation conjointe non-orthogonale de matrices de covariance sur K trames temporelles successives, sous l'hypothèse de non-stationnarité des sources d'une trame à l'autre. La matrice de covariance des signaux extraits à la trame k, notée $\hat{\mathbf{R}}_{\mathbf{s},\mathbf{k}}$, s'exprime en fonction de $\hat{\mathbf{R}}_{\mathbf{x},\mathbf{k}}$ et de la matrice de séparation $\mathbf{B}$:

$$\hat{\mathbf{R}}_{\mathbf{s},\mathbf{k}} = \mathbf{B}\hat{\mathbf{R}}_{\mathbf{x},k}\mathbf{B}^T \tag{2.25}$$

La matrice $\mathbf{B}$ rendant $\hat{\mathbf{R}}_{\mathbf{s},\mathbf{k}}$ la plus diagonale possible pour chaque trame est alors recherchée. Le critère à minimiser est le suivant :

$$C = \sum_{k=1}^{K} \frac{T_k}{T} \cdot \text{off}(\hat{\mathbf{R}}_{\mathbf{s},k}) = \sum_{k=1}^{K} \frac{T_k}{T} \cdot \text{off}(\mathbf{B}\hat{\mathbf{R}}_{\mathbf{x},k}\mathbf{B}^T) \tag{2.26}$$

avec $\frac{T_k}{T}$ le rapport entre la longueur de la trame k et la longueur totale de toutes les trames, et $\text{off}(\hat{\mathbf{R}}_{\mathbf{s},k}) = \text{D}\{\hat{\mathbf{R}}_{\mathbf{s},k} | \, \text{diag}(\hat{\mathbf{R}}_{\mathbf{s},k})\}$ la divergence de Kullback-Leibler exprimant ici la "non-diagonalité" de la matrice $\hat{\mathbf{R}}_{\mathbf{s},k}$. Cette méthode a ensuite été étendue par H. Boumaraf pour un mélange convolutif de sources [50]. Cette approche est similaire à d'autres méthodes de diagonalisation conjointe comme JADE pour des statistiques du 4ème ordre, que nous aborderons plus loin.

Algorithme WASOBI - séparation de signaux gaussiens d'après un modèle AR WASOBI pour *Weights-Adjusted Second-Order Blind Identification* effectue la diagonalisation conjointe de matrices d'intercorrelation [51]. Basé sur les hypothèses que les sources sont indépendantes, gaussiennes, avec une durée d'autocorrélation de K échantillons, l'algorithme effectue la diagonalisation conjointe de K matrices $\mathbf{R}_\mathbf{x}(\tau)$, pour $0 \leq \tau \leq K-1$.

2.2.3. Méthodes basées sur les statistiques d'ordres supérieurs

Lorsque les sources sont supposées non-gaussiennes et indépendantes, les statistiques d'ordres supérieurs à 2 peuvent aider à identifier et séparer celles-ci, on parle alors d'analyse en composantes indépendantes (ACI). En pratique, l'étude des statistiques d'ordres supérieurs se limite souvent à l'ordre 4. On peut identifier deux groupes d'algorithmes d'ACI : les méthodes tensorielles qui diagonalisent des tenseurs de cumulants et constituent une extension des méthodes de décorrélation aux ordres supérieurs à 2, et les méthodes basées sur la théorie de l'information, cherchant les sources ayant l'entropie marginale la plus faible. Bien que ces deux approches soient différentes sur la forme, elles consistent toutes deux à minimiser l'information mutuelle entre les sources, dans un cas en minimisant les cumulants croisés, dans l'autre en minimisant la somme des entropies marginales.

Blanchiment des données Blanchir un jeu de variables revient à les décorréler et les normaliser en énergie. C'est une étape préliminaire à la plupart des algorithmes d'ACI, qui se traduit par une ACP suivie d'une normalisation des composantes principales par

leurs valeurs propres. Á l'issue du blanchiment, on obtient un ensemble de variables $\mathbf{z}$ telles que :

$$\mathrm{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{Id} \tag{2.27}$$

avec $\mathbf{Id}$ une matrice identité.
Une matrice de mélange $\mathbf{A}$, de dimensions $M \times N$, peut être décomposée en valeurs singulières, soit :

$$\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{W}^T \tag{2.28}$$

avec $\mathbf{V}$ et $\mathbf{D}$ s'apparentant aux vecteurs propres et valeurs propres obtenus par ACP et $\mathbf{W}$ une matrice unitaire $N \times N$. En estimant les matrices $\mathbf{V}$ et $\mathbf{D}$, le blanchiment permet :

— de ramener le problème à l'identification d'une matrice $\mathbf{W}$ unitaire, possédant moins de degrés de liberté que la matrice de mélange,

— de réduire potentiellement la dimensionnalité du problème en présence de valeurs propres nulles (soit $N < M$), en conservant uniquement le sous-espace signal.

Le problème de séparation, initialement posé comme $\mathbf{x} = \mathbf{A}\mathbf{s}$, devient :

$$\mathbf{z}(t) = \mathbf{W}^T\mathbf{s}(t) \tag{2.29}$$

ou encore

$$\mathbf{s}(t) = \mathbf{W}\mathbf{z}(t) \tag{2.30}$$

$\mathbf{W}$ représente donc la matrice de séparation des observations blanchies. L'ACI fait souvent appel à des algorithmes de gradient dérivant une fonction de coût par rapport à $\mathbf{W}$ afin de la minimiser. La recherche de la matrice de séparation se fait alors le plus souvent ligne par ligne, c'est-à-dire source par source. Ce sont les méthodes que l'on retrouve sous le nom de *one-unit algorithms* dans la littérature. Par la suite, la notation w_i représente la i^{e} ligne de la matrice de séparation, agencée en vecteur colonne. Ainsi, l'estimation de la source s_i se fait suivant l'opération :

$$\mathbf{s_i}(t) = w_i^T\mathbf{z}(t) \tag{2.31}$$

2.2.3.1. ACI par méthodes tensorielles

Les méthodes tensorielles d'ACI ont été initiées par Hérault et Jutten au début des années 80 [52] et développées par Cardoso [53] et Comon [54]. Elles consistent à trouver la matrice qui diagonalise au mieux un ou plusieurs tenseurs de cumulants, c'est-à-dire qui annule au mieux les cumulants croisés qui représentent la dépendance entre les variables. Les cumulants utilisés sont en général d'ordre 4. Différentes méthodes de diagonalisation ont été imaginées, nous abordons ici la première méthode appelée FOBI et l'algorithme JADE qui est le plus répandu et qui est une généralisation de FOBI montrant une plus grande robustesse.

Méthode FOBI FOBI pour *Fourth Order Blind Identification* est une méthode élaborée par Cardoso [53]. Elle consiste à diagonaliser la matrice $\mathbf{\Omega}$ composée du sous-ensemble de cumulants d'ordre 4 :

$$\mathbf{\Omega} = E\{\mathbf{z}\mathbf{z}^T||\mathbf{z}||^2\} \tag{2.32}$$

D'après les propriétés du blanchiment énoncées précédemment, on cherche alors à identifier la matrice $\mathbf{W}$ telle que :

$$\mathbf{\Omega} = E\{\mathbf{z}\mathbf{z}^T||\mathbf{z}||^2\} = \mathbf{W}^T E\{\mathbf{s}\mathbf{s}^T||\mathbf{s}||^2\}\mathbf{W} \tag{2.33}$$

Cette méthode est simple et possède un coût de calcul faible, équivalent à celui d'une ACP. En contre-partie, une condition nécessaire à la séparation effective des sources réside dans le fait que les valeurs propres de la matrice $\mathbf{\Omega}$ doivent être uniques, ce qui signifie que les kurtosis des sources doivent être distincts. Aussi, en pratique, FOBI échoue à séparer des sources ayant des distributions identiques [25].

Algorithme JADE L'algorithme JADE (*Joint Approximate Diagonalization of Eigenmatrices*) est une extension de la méthode FOBI par Cardoso et Souloumiac [39]. Au lieu de diagonaliser une matrice d'un sous-ensemble des cumulants d'ordre 4, JADE diagonalise conjointement plusieurs matrices de cumulants d'ordre 4, ce qui permet de prendre en compte un nombre de cumulants d'ordre 4 supérieur à FOBI et ainsi d'augmenter la robustesse de la séparation.
Les approches tensorielles ont l'avantage d'une implémentation simple et d'une vitesse de calcul très rapide pour des systèmes de faibles dimensions, cependant le coût de calcul augmente rapidement lorsqu'il s'agit de diagonaliser des matrices de grande taille, aussi elles ne sont donc pas adaptées lorsque le nombre d'observations devient trop important. Par ailleurs, les cumulants d'ordre 4 sont très sensibles aux valeurs extrêmes (*outliers*) et une quantité importante de données est souvent nécessaire pour limiter la variance des estimateurs.

2.2.3.2. ACI par minimisation de l'entropie

Une autre approche consiste à réduire l'information mutuelle entre les variables, non pas en annulant les cumulants croisés, mais en exprimant celle-ci en fonction de l'entropie des sources recherchées. La principale différence entre les méthodes exploitant cette approche se trouve alors dans la façon dont l'entropie va être estimée.

Negentropie et non-gaussianité Le théorème central limite établit qu'une somme infinie de variables aléatoires indépendantes tend vers une variable aléatoire gaussienne. Ainsi, rechercher les composantes indépendantes au sein d'un mélange peut être vu comme la recherche des signaux les moins gaussiens possibles, d'où le concept de non-gaussianité [25]. Sachant qu'une variable gaussienne est celle possédant l'entropie maximale parmi toutes les variables de même énergie, s'éloigner d'une distribution gaussienne est équivalent à minimiser l'entropie marginale des sources.

La non-gaussianité d'une variable s est souvent exprimée en faisant appel à l'entropie négative, ou negentropie, définie comme la différence entre l'entropie d'une variable gaussienne s_g de même variance que s et l'entropie de la variable s, soit :

$$N(s) = H(s_g) - H(s) \tag{2.34}$$

La negentropie des sources extraites est souvent utilisée pour l'ACI comme un critère à maximiser, car elle éloigne les sources extraites d'une distribution gaussienne et doit à terme les rendre plus indépendantes.

Algorithme FastICA FastICA est un algorithme basé sur la maximisation de la negentropie des signaux estimés et utilisant la méthode du point fixe, décrit dans [55]. Dans cet article, les auteurs proposent d'utiliser au choix deux fonctions de mesure non-linéaire $G(s_i)$ pour approcher l'entropie :

$$G(s_i) = \begin{cases} G_1(s_i) & = & \frac{1}{a_1} log(cosh(a_1 s_i)) \\ G_2(s_i) & = & -e^{\frac{-s_i^2}{2}} \end{cases} \tag{2.35}$$

avec $1 \leqslant a_1 \leqslant 2$. La negentropie est alors approchée par l'équation :

$$N(s_i) = [E\{G(s_g)\} - E\{G(s_i)\}]^2 \tag{2.36}$$

s_g étant un signal gaussien de même énergie que s_i. La fonction G peut également être choisie différemment, en fonction d'*a priori* sur la distribution des sources.
Les lignes w_i de la matrice de séparation sont estimées les unes après les autres par un algorithme de point-fixe, en maximisant la negentropie des composantes extraites. En posant g et g' comme étant les dérivées première et seconde des fonctions de mesure G, le déroulement de l'algorithme est le suivant :

Début
Initialisation aléatoire
Tant que $\sum_i N(s_i)$ **n'a pas convergé**
Pour chaque composante
1. Poser $w_i+ = E\{\mathbf{z}g(w_i^T\mathbf{z})\} - E\{g'(w_i^T\mathbf{z})\}\mathbf{w}_i$,
2. Mettre à jour w_i : $w_i = \frac{w_i+}{||w_i||+}$,

Fin
Fin
Fin

Une étape d'orthogonalisation des vecteurs w_i est également opérée à chaque itération de l'algorithme, de façon à ne pas converger plusieurs fois vers la même solution.

Algorithme EFICA EFICA est une optimisation de FastICA mise au point par koldovsky *et al.* [56]. Cette version de FICA ajuste la fonction de coût associée à chaque composante recherchée, de façon à converger plus rapidement et diminuer la variance introduite des estimateurs en présence d'un nombre réduit d'échantillons.

Algorithme Infomax La version de l'algorithme de Amari [57] cherche à minimiser l'information mutuelle $I(\mathbf{W})$ telle que :

$$I(\mathbf{W}) = -H(\mathbf{s}) + \sum_{a=1}^{N} H(s_i) \tag{2.37}$$

Ce critère peut être exprimé en utilisant la divergence de Kullback-Leibler, soit :

$$I(\mathbf{W}) = \int p(\mathbf{s}) \cdot log \frac{p(\mathbf{s})}{\prod_{i=1}^{n} p(s_i)} d\mathbf{s} \tag{2.38}$$

où $p(s_i)$ désigne la densité de probabilité marginale de s_i. Cela correspond à l'expression de l'information mutuelle de l'équation 2.13.
La densité de probabilité de chaque source est approchée par une série de Gram-Charlier, faisant intervenir les moments d'ordres supérieurs.

La structure temporelle des signaux n'est pas prise en compte dans les méthodes basées sur l'entropie seule. Celle-ci ne dépend en effet pas de l'ordre dans lequel les échantillons sont disposés, ce qui peut paraître surprenant étant donné que l'agencement temporel des données est un vecteur important de l'information. A défaut d'exploiter pleinement la structure temporelle des signaux, différents algorithmes d'ACI ont intégré une modélisation des signaux avec un modèle source-filtre auto-régressif (AR) où le signal excitateur est supposé indépendant (temporellement) et identiquement distribué (hypothèse i.i.d.). Cette modélisation a pour but de traiter non plus directement le mélange de sources mais un mélange de signaux i.i.d. pour lesquels les estimateurs statistiques possèdent une variabilité moins forte. La longueur du modèle AR est généralement fixée au départ, puis les coefficients sont estimés itérativement.

Algorithme AR-MoG AR-MoG [58] modélise les sources comme des variables i.i.d $p_i(t)$ passées au travers d'un processus AR $\mathbf{a}$ de longueur P, soit :

$$s_i(t) = u_i(t) - \sum_{p=1}^{P-1} a_i(p) s_i(t-p) \tag{2.39}$$

La densité de probabilité de $u_i(t)$ est supposée être un mélange de gaussiennes (MoG pour *Mixture of Gaussians*). Celle-ci s'exprime de la manière suivante :

$$p(u_i(t)) = \sum_{q=1}^{Q} \mathcal{N}(u_i(t)|\mu_{iq}, \nu_{iq}) \cdot \pi_{iq} \tag{2.40}$$

où $\mathcal{N}$ symbolise une distribution gaussienne, dépendant des paramètres μ sa moyenne et ν son écart-type et π_{iq} représente une pondération.
Une estimation de la log-vraisemblance est obtenue à partir de ces paramètres, soit :

$$L = N \det W^T + \sum_{i=1}^{M} \sum_{t=1}^{N} \sum_{q=1}^{Q} \gamma_{itq} \log(\frac{\mathcal{N}(u_i(t)|\mu_{iq}, \nu_{iq}) \cdot \pi_{iq}}{\gamma_{itq}}) \tag{2.41}$$

où γ_{itq} est un facteur de pondération.
Les paramètres sont estimés de façon à maximiser la vraisemblance, suivant un procédé estimation-maximisation (EM) :

Début

Initialisation

Tant que L **n'a pas convergé**

1. Etape E :

   Optimisation du paramètre γ_{itq},

   Calcul de la vraisemblance et test de convergence,

2. Etape M :

   Estimation de $\mathbf{W}^T$ par un algorithme de gradient,

   Estimations successives de $\mathbf{a}$, μ_{iq}, ν_{iq} et π_{iq}.

Fin

Fin

Algorithme ERM-ARG ERM-ARG, pour *Entropy Rate Minimisation - Autoregressive Generalized gaussian distribution*, minimise l'entropie de signaux supposés issus d'un filtrage AR, dont le processus excitateur $u_i(t)$ possède une distribution gaussienne généralisée [59]. La modélisation AR est ici vue comme un cas particulier de sources markoviennes. La fonction de coût minimisée reflète l'information mutuelle, soit :

$$J(s_1, ..., s_n) = \sum_{i=1}^{N} H(u_i) - \log |\det \mathbf{W}^T| \tag{2.42}$$

Les lignes w_i^T de la matrice sont alors mises à jour l'une après l'autre par un algorithme de gradient minimisant J.

Les algorithmes précédents procèdent à une unique estimation de l'entropie d'après un modèle pré-défini, en mettant si besoin à jour les paramètres de ce modèle. Des méthodes basées sur un jeu de plusieurs estimateurs de l'entropie ont été mises au point notamment par G. Fu, X. Li et T. Adali.

Algorithme EBM EBM [60] pour *Entropy Bound Minimization* calcule K estimations de l'entropie pour chaque composante, à chaque itération de l'algorithme. Chacune correspond à l'entropie maximale possible, connaissant un estimateur statistique non linéaire associé à la composante s_i. On nomme fonction de mesure cet estimateur $G_k(s_i)$, avec $1 \leq k \leq K$. La densité de probabilité permettant l'entropie maximale connaissant $G_k(s_i)$ est $p_k(s_i)$, prenant la forme :

$$p_k(s_i) = Ae^{-as^2 - bs - cG(s_i)} \tag{2.43}$$

où les constantes A, a, b, c sont obtenues par des méthodes numériques en prenant en compte les contraintes de moyenne nulle, de variance unité. Ce choix est justifié par le principe d'entropie maximale [61], stipulant que l'estimation la moins biaisée de l'entropie pour un ensemble réduit d'observations est celle la moins contraignante, donc la plus élevée. L'entropie maximale $H_k(s_i)$ s'exprime alors comme :

$$H(s_i) \leq H_k(s_i) = -\log A + a + c\mu_{k,g} \tag{2.44}$$

avec $\mu_{k,g} = E[G_k(s_i)]$.
EBM évalue l'entropie maximale pour un ensemble de K fonctions de mesure et conserve finalement la valeur la plus faible, selon le principe du maximum de vraisemblance, soit :

$$H(s_i) = \underset{1 \leq k \leq K}{\arg\min}\, H_k(s_i) \tag{2.45}$$

Une descente de gradient classique, décrite dans [60], est appliquée pour minimiser la fonction de coût J_i associée à la i^{e} composante :

$$J(s_i, w_i) = H(z_i) - \log |h_i^T w_i| \tag{2.46}$$

où le vecteur h_i est de norme unité et orthogonal à tous les autres vecteurs w_j, $i \neq j$. Le terme $\log |h_i^T w_i|$ permet de préserver l'orthogonalité des vecteurs de séparation.

Algorithme ERBM ERBM (*Entropy Rate Bound Minimization*), est une extension de l'algorithme EBM, auquel est couplée une modélisation AR des sources estimées [62]. Comme pour AR-MoG, chaque signal extrait $s_i(t)$ est considéré comme la résultante du filtrage d'un signal i.i.d $u_i(t)$ par un filtre inversible a_i de longueur P tel que :

$$u_i(t) = s_i * a_i = \sum_{p=0}^{P-1} a_i(p) s_i(t-p) \tag{2.47}$$

Cette fois, les coefficients $a_i(k)$ sont estimés par minimisation du ratio d'entropie (*entropie rate*) associée à $u_i(t)$. Le ratio d'entropie H_r correspond à la quantité d'information apportée par chaque échantillon temporel :

$$H_r(x[n]) = H(x[n]|x[n-1], x[n+1], x[n-2], \dots) \tag{2.48}$$

Pour un signal i.i.d, le ratio d'entropie correspond à l'entropie, les échantillons étant indépendants.
La fonction de coût globale minimisée par l'algorithme est donc :

$$J(u_i, w_i, a_i) = H_r(u_i) - \log |h_i^T w_i| \tag{2.49}$$

Pour chaque source, on optimise ainsi alternativement la matrice de séparation et les coefficients AR, jusqu'à convergence. L'estimation des coefficients AR se fait par une approche identique à EBM. Le pseudo-code d'ERBM est le suivant :

Début

Initialisation de $\mathbf{W}$ par EBM

Initialisation aléatoire de a_i pour $1 \leq i \leq N$

Estimation de u_i et $H(u_i)$ pour $1 \leq i \leq N$

Tant que $\sum_{i}^{N} J(u_i)$ **n'a pas convergé**

Pour chaque composante i

1. Mise à jour de a_i par EBM
2. $u_i = (w_i^T \cdot \mathbf{z}) * a_i$
3. Estimation de $H_k(u_i)$ pour les K fonctions de mesure
4. $H(u_i) = \arg\min H_k(u_i)$
5. Calcul de $\frac{\partial J(u_i)}{\partial w_i}$
6. Mise à jour de w_i par descente de gradient

Fin

Fin

Fin

2.2.3.3. AVI - analyse en vecteurs indépendants

L'analyse en vecteurs indépendants (AVI) est une extension de l'ACI pour laquelle la séparation s'effectue conjointement sur différents jeux de données, en supposant qu'il existe une corrélation entre les composantes d'un jeu à l'autre. Le concept est évoqué pour la première fois par T. Kim dans [63] et l'on retrouve un panorama des méthodes existantes actuellement dans [26]. Pour K jeux d'observations, contenant N observations chacun et dans le cas de de systèmes déterminés, le problème de séparation aveugle de sources, initialement posé par l'équation 2.1 devient alors :

$$\mathbf{x}^{[k]} = \mathbf{A}^{[k]} \cdot \mathbf{s}^{[k]} \tag{2.50}$$

avec :

- k le jeu de données considéré ($1 \leq k \leq K$),
- $\mathbf{x}^{[k]}$ les N observations du jeu k,
- $\mathbf{A}^{[k]}$ une matrice de mélange $N \times N$,
- $\mathbf{s}^{[k]}$ les N composantes indépendantes du jeu k .

On suppose donc ici que les composantes s_m^k et s_m^{k+1} sont corrélées alors que les composantes s_m^k s_n^k sont indépendantes pour $m \neq n$. L'AVI cherche alors à minimiser l'entropie marginale de chaque composante, et à maximiser l'information mutuelle entre les composantes liées entre elles. Une fonction de coût générale peut ainsi être posée [26] :

$$I(\mathscr{W}) = \sum_{n=1}^{N} \Big(\sum_{k=1}^{K} H(s_n^{[k]}) - I(\mathbf{s}_n) \Big) - \sum_{k=1}^{K} \log |det(\mathbf{W^{[k]}}|) - C \tag{2.51}$$

avec

— $\mathscr{W}$ un tenseur de dimensions $N \times N \times K$ concaténant les matrices de séparation $\mathbf{W}^{[\mathbf{k}]}$ pour tous les k,

— $s_n^{[k]}$ la n^{e} composante estimée dans le jeu k,

— $\mathbf{s}_n$ contenant tous les $s_n^{[k]}$ liés entre elles.

— $I_r(\mathbf{s}_n)$ l'information mutuelle de $\mathbf{s}_n$.

Deux cas de figures peuvent être identifiés [64] :

— Lorsqu'il existe une dépendance linéaire entre les jeux de données : algorithme MCCA *Multiset Canonical Correlation Analysis* [62] ou JBSS *Joint Blind Source Separation* [65] .

— Lorsque cette dépendance est non-linéaire : extension de la MCCA à des fonctions non linéaires des observations par Todros [66] ou prise en compte de la dépendance aux ordres supérieurs [67].

2.3. Analyse de scène basée sur l'ambisonie

Les propriétés de l'encodage ambisonique (signaux coïncidents, matrice de mixage instantanée dépendant uniquement de l'angle d'incidence) ont débouché sur plusieurs techniques d'analyse de scène. Les méthodes dirAC et HARPEX sont présentées ici, elles ont en commun d'effectuer une décomposition temps-fréquence-espace de la scène sonore en vue de faire de l'*upmix*, c'est à dire de reconstruire à partir d'une captation à l'ordre m, un champ sonore ambisonique d'ordre $m' \geqslant m$, en calculant les harmoniques sphériques d'ordres supérieurs associées aux sources identifiées. Ces méthodes, similaires à l'algorithme DUET [41], posent l'hypothèse d'une parcimonie plus ou moins forte dans le domaine temps/fréquence.

Méthode dirAC V. Pulkki propose une méthode d'analyse et d'upmix basée sur une captation ambisonique au format B (ambisonique ordre 1), nommée dirAC pour *Directional Audio Coding*. Celle-ci est fondée sur une hypothèse forte de parcimonie puisqu'elle suppose qu'à chaque instant et dans chaque sous-bande de fréquences, une seule source est active.
dirAC est basée sur la technique SIRR (*Spatial Impulse Response Rendering*) mise au point avec J.Merimaa [68], qui consiste, après une décomposition temps-fréquence du champ sonore, à l'exprimer comme la résultante d'une onde plane incidente et d'un champ diffus [9] [69]. La principale différence avec la méthode SIRR réside dans la décomposition temps-fréquence utilisée : alors que SIRR se base sur une analyse par transformée de Fourier à court terme (STFT, *Short-Time Fourier Transform*), dirAC décompose le format B par un filtrage ERB (*Equivalent Rectangular Bandwidth*), qui reproduit un modèle psychoacoustique.
L'analyse est basée sur l'intensité acoustique $\vec{I}$ qui représente le flux d'énergie acoustique à travers une surface élémentaire. Après application d'une STFT, on exprime le vecteur intensité $\vec{I}(k, \omega)$ pour un signal harmonique de pulsation ω pour la trame k de façon

analogue à l'équation 1.28, soit :

$$\vec{I}(k,\omega) = \frac{1}{2}\Re\{p(k,\omega)\vec{v}^*(k,\omega)\} = \frac{1}{2}\Re\Big\{p^*(k,\omega)\left|\begin{array}{l} v_x(k,\omega)\cdot\vec{e_x} \\ v_y(k,\omega)\cdot\vec{e_y} \\ v_z(k,\omega)\cdot\vec{e_z}\end{array}\right.\Big\} \tag{2.52}$$

Grâce à l'équation d'Euler linéarisée et aux propriétés du format B, on trouve la relation suivante :

$$\begin{aligned}\vec{I}(k,\omega) &= \tfrac{1}{2\rho_0 c_0}\Re\Big\{p^*(k,\omega)\left|\begin{array}{l} X(k,\omega)\cdot\vec{e_x} \\ Y(k,\omega)\cdot\vec{e_y} \\ Z(k,\omega)\cdot\vec{e_z}\end{array}\right.\Big\} \\ &= \tfrac{1}{2\rho_0 c_0}\Re\Big\{W^*(k,\omega)\left|\begin{array}{l} X(k,\omega)\cdot\vec{e_x} \\ Y(k,\omega)\cdot\vec{e_y} \\ Z(k,\omega)\cdot\vec{e_z}\end{array}\right.\Big\}\end{aligned} \tag{2.53}$$

avec $W(k,\omega)$, $X(k,\omega)$, $Y(k,\omega)$ et $Z(k,\omega)$ les composantes fréquentielles du format B. La direction d'incidence $(\theta(k,\omega),\phi(k,\omega))$ va alors correspondre à la direction de propagation du flux d'énergie, soit :

$$\theta(k,\omega) = \arctan 2\frac{I_y(k,\omega)}{I_x(k,\omega)} \tag{2.54}$$

$$\phi(k,\omega) = \arctan 2\frac{I_z(k,\omega)}{\sqrt{I_x^2(k,\omega)+I_y^2(k,\omega)}} \tag{2.55}$$

où l'opérateur arctan 2 est une variante à deux arguments de la fonction arctan, permettant de lever l'indétermination de 180° sur l'angle calculé.
Le caractère diffus du champ acoustique s'exprime par un coefficient de diffusivité [69] :

$$\Phi(k,\omega) = 1 - \frac{\sqrt{3}\|\Re\{W^*(k,\omega)\vec{V}(k,\omega)\}\|}{|W^*(k,\omega)|^2 + \|\vec{V}(k,\omega)\|^2} \tag{2.56}$$

avec $\vec{V}$ le vecteur des composantes $X(k,\omega)$, $Y(k,\omega)$, $Z(k,\omega)$.
On peut aisément interpréter le terme $\Phi(k,\omega)$. Dans le cas d'un champ acoustique constitué à l'instant k d'une seule onde plane incidente de pulsation ω, le champ ambisonique respecte la condition $W^2(k,\omega) = 3(X^2(k,\omega)+Y^2(k,\omega)+Z^2(k,\omega))$. C'est le cas pour lequel la diffusivité du champ sera nulle et la directivité maximale. On obtient alors :

$$\Phi = 1 - \frac{\sqrt{3}\|\Re\{W^*\vec{V}\}\|}{|W^*|^2+\|\vec{V}\|^2} = 1 - \frac{\sqrt{3}\sqrt{W^2\cdot(X^2+Y^2+Z^2)}}{|W|^2+\sqrt{X^2+Y^2+Z^2}^2} = 1 - \frac{\sqrt{3}\sqrt{3\cdot W^4}}{3\cdot W^2} = 0 \tag{2.57}$$

où l'indice (k,ω) est sous-entendu. À l'inverse, un champ complètement diffus engendrera des composantes directionnelles X,Y et Z nulles et donc $\Phi = 1$.

Cette approche permet une décomposition mathématique du champ sonore mais comporte des limites. D'une part, la représentation temps-fréquence ne suffit pas pour identifier une source large-bande, il faut dans ce cas passer par une étape de regroupement des bins appelée *clustering*. De plus, un effet de salle réel n'est jamais parfaitement isotrope et l'hypothèse d'une unique source par bande de fréquences peut souvent être mise en défaut.
Un algorithme analogue est proposé par J. Palacino [70] [71], appliqué à une captation non pas ambisonique mais issue de trois microphones cardioïdes coïncidents.

Méthode HARPEX Egalement basée sur l'ambisonie d'ordre 1, HARPEX (*High Angular Resolution Planewave Expansion*) est une méthode mise au point par S. Berge *et al.* [8] [72] permettant de décomposer, après STFT, chaque bande de fréquences en deux ondes planes d'incidences distinctes.
Pour chaque trame k et chaque fréquence angulaire ω, les composantes $W_k(\omega)$, $X_k(\omega)$, $Y_k(\omega)$ et $Z_k(\omega)$ s'expriment sous forme matricielle :

$$\mathbf{X}_{k,\omega} = \begin{bmatrix} W_r & W_i \\ X_r & X_i \\ Y_r & Y_i \\ Z_r & Z_i \end{bmatrix} \tag{2.58}$$

laissant apparaître explicitement leurs parties réelles et imaginaires. HARPEX décompose chaque bande de fréquences en une somme de deux ondes planes d'amplitudes complexes, soit :

$$\mathbf{X}_{k,\omega} = \underbrace{\begin{bmatrix} 1 & 1 \\ x_1 & x_2 \\ y_1 & y_2 \\ z_1 & z_2 \end{bmatrix}}_{\mathbf{A}_{k,\omega}} \cdot \underbrace{\begin{bmatrix} S_{1r} & S_{1i} \\ S_{2r} & S_{2i} \end{bmatrix}}_{\mathbf{S}_{k,\omega}} \tag{2.59}$$

avec $\mathbf{A}_{k,\omega}$ contenant les coefficients de mélange ambisoniques des ondes 1 et 2 et $\mathbf{S}_{k,\omega}$ leurs amplitudes complexes.
On sous-entend par la suite les indices (k, ω). $\mathbf{X}$ est d'abord décomposé à l'aide d'un algorithme de décomposition QR tel que :

$$\mathbf{X} = \mathbf{QR} \tag{2.60}$$

où $\mathbf{Q}$ est une matrice de dimensions 4x2 dont les colonnes sont orthonormées et $\mathbf{R}$ est une matrice de dimensions 2x2 triangulaire supérieure. Des transformations sont ensuite opérées aux matrices $\mathbf{Q}$ et $\mathbf{R}$ de façon à ce que chaque colonne de $\mathbf{Q}$ soit conforme aux propriétés d'encodage du format B pour une onde plane, soit :

— $3w_i^2 = x_i^2 + y_i^2 + z_i^2$ pour chaque colonne i,
— le coefficient de mélange du canal omnidirectionnel égal à 1.

On a donc, pour chaque point temps/fréquence, une décomposition de champ sonore en deux ondes planes d'amplitudes complexes. De ces coefficients peuvent alors être déduites les directions d'incidence des ondes planes. Lorsque la décomposition en deux ondes planes n'est mathématiquement pas possible (réverbération ou nombre de sources trop important), une solution alternative est alors utilisée, décrite dans [73], incluant une troisième composante omnidirectionnelle, contribuant uniquement au canal W. Cette approche comporte les mêmes limitations que diRAC, à ceci près que deux sources peuvent être identifiées dans chaque bande de fréquences en champ libre. En présence d'un nombre de sources plus important ou dans le cas d'un mélange réverbérant, la décomposition HARPEX ne permet pas de retrouver les directions d'arrivée réelles des sources.
Ces deux méthodes ont donc plusieurs limitations qui les rendent peu utilisables en pratique :

— une hypothèse de parcimonie forte (une ou deux sources par bande de fréquences),

— un manque de robustesse en milieu réverbérant,

— une décomposition par bande de fréquences mais sans lien avec la réalité physique de la scène, à savoir potentiellement des sources à larges spectres qui nécessiterait une étape de *clustering* afin d'agréger les contributions d'une même source.

Régularisation parcimonieuse Une méthode d'analyse parcimonieuse a été utilisée par N. Epain, A. Wabnitz *et al.* pour analyser une captation ambisonique dans le domaine temporel [37] [74] ou fréquentiel [10] [11] [75], dans une formulation sous-déterminée du problème d'analyse de scène. Considérant un nombre N de sources sonores potentielles réparties sur un maillage sphérique de l'espace, N étant grand devant le nombre de canaux ambisoniques. Ces sources, notées $\mathbf{s}$, contribuent au signal ambisonique $\mathbf{x}$ d'ordre m au travers d'une matrice de mélange ambisonique $\mathbf{Y}$ dépendant de leurs coordonnées sphériques, soit le système d'équations :

$$\mathbf{x}(t) = \mathbf{Y}\mathbf{s}(t) \tag{2.61}$$

où $\mathbf{Y}$ a pour dimensions $(m+1)^2 \times N$ sachant que $(m+1)^2 \ll N$.
Le système d'équations ainsi posé est donc largement sous-déterminé, le nombre de sources potentielles étant très supérieur au nombre de signaux ambisoniques. L'approche parcimonieuse choisie consiste à rechercher la solution minimisant l'énergie totale de l'ensemble, c'est-à-dire à minimiser le critère $\|s\|_{12}$ tel que :

$$\|\mathbf{s}\|_{12} = \sum_{i=1}^{N} \sqrt{\sum_{t=1}^{T} \hat{s}_i^2(t)} \tag{2.62}$$

sous la contrainte $\mathbf{x}(t) = \mathbf{Y}\mathbf{s}(t)$. Pour réduire la complexité du problème et optimiser le temps de calcul, une décomposition en valeurs singulières (SVD, *Singular Value Decomposition*) est préalablement appliquée aux signaux temporels ambisoniques.

Application de l'ACI à une captation ambisonique L'analyse en composantes indépendantes appliquée à l'ambisonie a été étudiée par N.Epain *et al.* [12] [76] sur des signaux voix d'une durée de quelques secondes. FastICA est appliqué à un contenu HOA d'ordre 4 contenant un mélange anéchoïque de signaux de parole, simulé à partir d'un modélisation numérique de microphone sphérique. Un filtrage passe-bande est préalablement appliqué aux signaux HOA avant l'analyse afin de ne conserver que la zone de validité de l'encodage, de façon à pouvoir formuler le problème de séparation de sources avec l'équation 2.1.
Une fois la matrice de mélange $\mathbf{A}$ estimée par ACI, la direction d'incidence (θ_i, ϕ_i) de chaque source s_i est déterminée en projetant chaque colonne $\mathbf{A}_i$ de la matrice de mélange sur un maillage $\mathbf{Y}$ de coefficients harmoniques sphériques et en choisissant les valeurs angulaires maximisant la corrélation entre les vecteurs A_i et $\mathbf{Y}_j$, soit :

$$(\theta_i, \phi_i) = \arg\max_j \frac{\mathbf{A}_i^t \cdot \mathbf{Y}_j}{\|\mathbf{A}_i\|\|\mathbf{Y}_j\|} \tag{2.63}$$

A partir de ces directions d'arrivée, la matrice de mélange finale est construite en utilisant les équations de l'encodage ambisonique, puis les sources sont extraites par inversion matricielle de la matrice de mélange, soit :

$$\hat{\mathbf{s}}(t) = \mathbf{A}^{-1}\mathbf{x}(t) \tag{2.64}$$

Dans le cas où les signaux des capsules (format A) sont accessibles, une extraction des sources directement depuis les signaux microphoniques est décrite dans [76], via une matrice de séparation globale.

2.4. Formation de voies pour la séparation de sources

La formation de voies consiste en une combinaison d'observations ayant des directivités distinctes afin de synthétiser de nouvelles observations possédant des directivités propres. Dans le cas d'un contenu multicanal de type ambisonique, les observations sont coïncidentes et les propriétés de directivité constantes en fonction de la fréquence, la formation de voies revient donc à une combinaison linéaire des observations pondérées par un facteur de gain réel. Dans ce cadre, la séparation spatiale de sources revient à une opération de formation de voies sous contraintes, recherchant pour chaque voie formée un gain unitaire dans la direction d'arrivée de la source à extraire et un gain nul dans la direction d'arrivée des sources interférentes.
Un contenu ambisonique constitué de N canaux offre N degrés de liberté pour la formation de voies permettant donc de séparer N sources. Lorsque le nombre de sources à séparer est inférieur au nombre de canaux, les degrés de liberté restants peuvent être mis à profit pour optimiser les propriétés de directivité des composantes extraites en fonction de l'application visée.
Cette section formalise tout d'abord la formation de voies pour une simple focalisation dans une direction de l'espace puis aborde la formation de voies pour la séparation de sources et les problématiques associées.

2.4.1. Focalisation ambisonique

La formation d'une voie s à partir d'un contenu ambisonique $\mathbf{x}$ d'ordre m constitué de M canaux est formalisée par l'expression :

$$s = \mathbf{bx} \tag{2.65}$$

où $\mathbf{b}$ est un vecteur-ligne de coefficients de dimensions $1 \times M$.
Le vecteur $\mathbf{D}_{\mathbf{x}}^{m}$ des directivités ambisoniques jusqu'à l'ordre m s'exprime suivant la formule :

$$\mathbf{D}_{\mathbf{x}}^{m}(\theta,\phi) = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta\cos\phi \\ \sqrt{3}\sin\theta\cos\phi \\ \sqrt{3}\sin\phi \\ \cdots \\ \tilde{Y}_{mn}^{\sigma}(\theta,\phi) \end{bmatrix} \tag{2.66}$$

La directivité formée $\mathcal{D}_{s}^{m}(\theta,\phi)$ est donc une combinaison linéaire des termes de $\mathbf{D}_{\mathbf{x}}^{m}$, c'est-à-dire :

$$\mathcal{D}_{s}^{m}(\theta,\phi) = b_1 + b_2\sqrt{3}\cos\theta\cos\phi + b_3\sqrt{3}\sin\theta\cos\phi + b_4\sqrt{3}\sin\phi + \cdots + b_m Y_{mn}^{\sigma}(\theta,\phi) \tag{2.67}$$

Une approche naturelle pour focaliser dans la direction (θ_i,ϕ_i) matérialisée par le vecteur unitaire $\vec{u_i}$ est d'exprimer un vecteur $\mathbf{b}_i$ comme un vecteur de coefficients harmoniques sphériques fonction de $\vec{u_i}$:

$$\mathbf{b}_i = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ \cdots \\ b_m \end{bmatrix} = \begin{bmatrix} g_0 \\ g_1\sqrt{3}\cos\theta_i\cos\phi_i \\ g_1\sqrt{3}\sin\theta_i\cos\phi_i \\ g_1\sqrt{3}\sin\phi_i \\ \cdots \\ g_m\tilde{Y}_{mn}^{\sigma}(\theta_i,\phi_i) \end{bmatrix} \tag{2.68}$$

qui sont pondérés à l'aide d'un vecteur $\mathbf{g} = [g_1, \cdots, g_m]$ dont les termes dépendent de l'ordre ambisonique. En posant $\mathbf{b}_i$ de cette façon, on obtient une fonction de directivité dont le gain maximal se trouve dans la direction $\vec{u} = \vec{u_i}$ et dont la valeur varie selon la distance angulaire γ par rapport la position $\vec{u_i}$.
On peut alors exprimer la fonction de directivité $\mathcal{D}_{s}^{1}(\theta,\phi)$ uniquement en fonction des variables g_m et γ. A l'ordre 1, l'équation 2.67 se ramène à l'expression :

$$\mathcal{D}_{s}^{1}(\theta,\phi) = b_1 + b_2\sqrt{3}\cos\theta\cos\phi + b_3\sqrt{3}\sin\theta\cos\phi + b_4\sqrt{3}\sin\phi \tag{2.69}$$

Par combinaison des équations 2.68 et 2.69, on arrive à la formulation suivante :

$$\begin{aligned} \mathcal{D}_{s}^{1}(\theta,\phi) &= b_1 + \sqrt{b_1^2+b_2^2+b_3^2}(\sin\theta\sin\theta_i + \cos\phi\cos\phi_i(\cos\theta - \cos\theta_i)) \\ &= b_1 + \sqrt{b_1^2+b_2^2+b_3^2}\cos\gamma \\ &= g_0 + 3g_1\cos\gamma \end{aligned} \tag{2.70}$$

La généralisation à l'ordre M est obtenue grâce aux propriétés liant les fonctions de Legendre et les fonctions harmoniques sphériques décrites dans [13]. En utilisant les équations 2.67 et 2.68, on exprime la directivité formée à l'ordre m :

$$\begin{aligned}
\mathcal{D}_s^m(\theta,\phi) &= \sum_{j=1}^{(m+1)^2} b_j \tilde{Y}_{mn}^{\sigma}(\theta,\phi) \\
&= \sum_{m'=0}^{m} g_{m'} \sum_{n,\sigma} \tilde{Y}_{m'n}^{\sigma}(\theta_i,\phi_i)\tilde{Y}_{m'n}^{\sigma}(\theta,\phi) \\
&= \sum_{m'=0}^{m} (2m'+1) g_{m'} \sum_{n,\sigma} Y_{m'n}^{\sigma}(\theta_i,\phi_i) Y_{m'n}^{\sigma}(\theta,\phi) \\
&= \sum_{m'=0}^{m} (2m'+1) g_{m'} P_{m'}(\cos\gamma)
\end{aligned} \tag{2.71}$$

où P_m est le polynôme de Legendre d'ordre m. On peut donc exprimer la directivité ainsi formée comme une fonction de $\cos\gamma$ faisant intervenir les fonctions de Legendre et des facteurs de pondération dépendant de l'ordre.
Une formulation similaire se retrouve dans [6] où la formation de voies est utilisée pour le décodage d'un contenu ambisonique sur haut-parleurs.

2.4.2. Optimisation de la focalisation

La formulation générale exprimée dans la section précédente suggère qu'il est possible d'optimiser le *beamforming* en adaptant les coefficients $g_{m'}$.

Gain unité La contrainte la plus courante est de vouloir conserver un gain unitaire dans la direction pointée. Pour cela, sachant que $P_m(0) = 1$ pour tout m, on procède en normalisant les coefficients $g_{m'}$, soit :

$$\tilde{g}_{m'} = \frac{g_{m'}}{\sum_{m'=0}^{m} (2m'+1) g_{m'}} \tag{2.72}$$

Formation de voie *in-phase* Si l'on cherche à manipuler localement un contenu ambisonique, par exemple pour déplacer un élément sonore dans la scène, on peut chercher la formation de voie ayant le support le plus compact possible afin d'éviter de perturber toute la scène sonore. Dans ce but, on peut utiliser le vecteur $\mathbf{g}$ associé à la solution nommée *in-phase*, qui est initialement proposé par D. Malham [77] pour le décodage ambisonique sur haut-parleurs (tableau 2.1). La solution *in-phase* est caractérisée par un gain positif décroissant de façon monotone depuis le point de focalisation, allant jusqu'à s'annuler dans la direction opposée, soit à $\gamma = \pi$. Les diagrammes de directivité associés sont donc dépourvus de lobes secondaires. La figure 2.1 illustre les directivités formées pour une focalisation dans la direction (0°,0°) en fonction de l'ordre ambisonique. On remarque que la directivité formée à l'ordre 1 correspond à une fonction cardioïde, puis que l'utilisation des ordres supérieurs tend à diminuer progressivement la largeur du lobe principal.

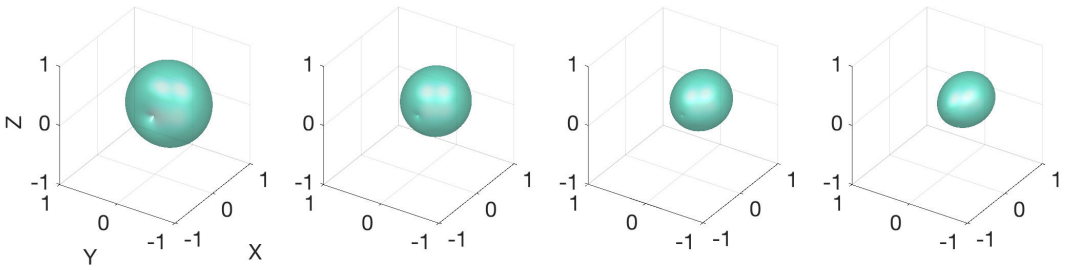

FIGURE 2.1. – Formation de voie *in-phase* en fonction de l'ordre ambisonique dans la direction (0°,0°). De gauche à droite : ordres 1 à 4. Gain unité à (0°,0°), gain nul à (180° ,0°)

	in-phase				max-SNR			
m	1	2	3	4	1	2	3	4
g_0	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{9}$	$\frac{1}{16}$	$\frac{1}{25}$
g_1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{3}{20}$	$\frac{2}{15}$	$\frac{1}{4}$	$\frac{1}{9}$	$\frac{1}{16}$	$\frac{1}{25}$
g_2	x	$\frac{1}{30}$	$\frac{1}{20}$	$\frac{2}{7}$	x	$\frac{1}{9}$	$\frac{1}{16}$	$\frac{1}{25}$
g_3	x	x	$\frac{1}{140}$	$\frac{1}{70}$	x	x	$\frac{1}{16}$	$\frac{1}{25}$
g_4	x	x	x	$\frac{1}{630}$	x	x	x	$\frac{1}{25}$

TABLEAU 2.1. – Coefficients de pondération $g_{m'}$ en fonction de l'ordre ambisonique m' et de l'ordre maximal m. A gauche : solution *in-phase*. A droite : solution max-SNR.

Maximisation du SNR Pour la séparation de sources et le débruitage, il est intéressant de chercher à effectuer une formation de voie minimisant le niveau de réverbération ou de bruit ambiant. Le facteur de directivité $\mathcal{F}_s^m$, évoqué notamment par B. Rafaely [78], permet d'estimer le niveau de réduction de bruit ou de réverbération obtenu par formation de voie à l'ordre m, en supposant que celui-ci puisse être modélisé comme un champ diffus isotrope. le facteur de directivité est formulé ainsi :

$$\mathcal{F}_s^m = \frac{|\mathcal{D}_s^m(\theta_i,\phi_i)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_{-\pi/2}^{\pi/2}|\mathcal{D}_s^m(\theta,\phi)|^2\cos\phi d\phi d\theta} \tag{2.73}$$

Le problème de maximisation de $\mathcal{F}_s^m$ sous la contrainte $\max|D_s^m(\theta,\phi)| = |D_s^m(\theta_i,\phi_i)|$ se résout en incorporant l'équation 2.71 dans l'équation 2.73, donnant alors :

$$\begin{aligned}\mathcal{F}_s^m &= \frac{\left(\sum_{m'=0}^{m'} g_{m'}P_{m'}(1)\right)^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_{-\pi/2}^{\pi/2}\left(\sum_{m'=0}^{m} g_{m'}P_{m'}(\cos\gamma)\right)^2\cos\phi d\phi d\theta} \\ &= \frac{\left(\sum_{m'=0}^{m} g_{m'}(2m'+1)\right)^2}{\sum_{m'=0}^{m} g_{m'}^2(2m'+1)}\end{aligned} \tag{2.74}$$

Solution	max-SNR	*in-phase*
F^1	4	3
F^2	9	5
F^3	16	7
F^4	25	9
F^M	$(M+1)^2$	$2M+1$

TABLEAU 2.2. – Facteurs de directivités (valeurs linéaires) obtenus avec les méthodes de formation de voies max-SNR et *in-phase* en fonction de l'ordre ambisonique.

La dérivation de cette expression permet d'établir que les maxima du facteur de directivité sont obtenus pour $g_\alpha = g_\beta \quad \forall \alpha$ et β (tableau 2.1). Les diagrammes de directivités associées à la solution max-SNR entre les ordres 1 et 4 sont représentés sur la figure 2.2. Le tableau 2.2 expose les facteurs de directivité correspondant aux méthodes max-SNR

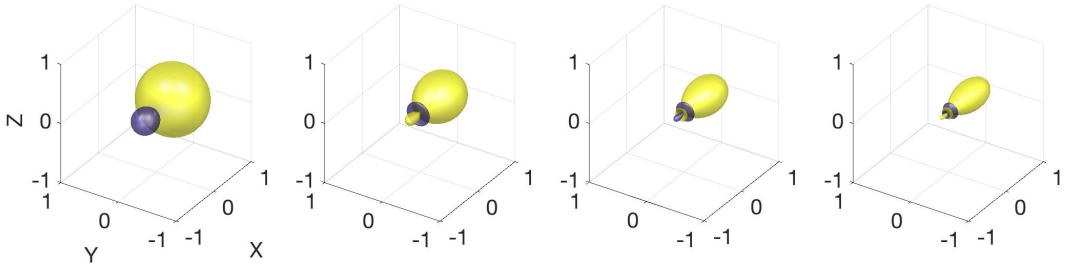

FIGURE 2.2. – Diagrammes de directivité : formation de voie dans la direction (0°,0°) maximisant le facteur de directivité, en fonction de l'ordre ambisonique. De gauche à droite : ordres 1 à 4. En jaune : valeur positive. En violet : valeurs négatives (inversion de phase).

et *in-phase* en fonction de l'ordre. Il est intéressant de constater qu'à l'ordre m, le facteur de directivité est de $M = (m+1)^2$ avec la solution max-SNR et de $2m+1$ avec la solution *in-phase*. Les valeurs obtenues en termes de gain maximum SNR permettent de connaître les performances maximales atteignables en termes de débruitage/déréverbération pour les différents cas d'usage rencontrés.

2.4.3. Formation de voies pour la séparation de sources

La séparation de sources fait apparaître un nombre de contraintes pour la formation de voies égal au nombre de sources à séparer. Pour chaque voie formée, il s'agit alors d'obtenir un diagramme de directivité possédant un gain unitaire dans la direction de la source à extraire et un gain nul dans la direction de chaque source interférente.
La matrice de séparation $\mathbf{B}$, dont chaque ligne $\mathbf{b}_i$ contient les coefficients permettant d'extraire la i^{e} source, est formée par inversion de la matrice de mélange ambisonique estimée. La figure 2.3 illustre les directivités formées aux ordres 1, 2 et 3 pour séparer

trois sources situées dans le plan azimutal à respectivement 0°, 90° et 120°. A l'ordre 1, on remarque que les directivités formées n'ont pas nécessairement un gain maximal dans la direction de chaque source extraite, avec un facteur d'environ 1.1 entre le gain dans la direction de la source et le gain maximal. L'ordre 2 permet ici d'obtenir une directivité plus sélective et de repositionner la source extraite au centre du lobe de directivité principal.

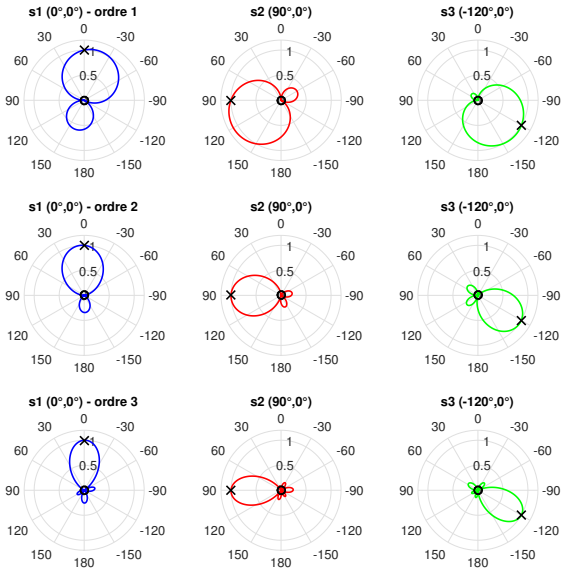

FIGURE 2.3. – Formation de voie pour la séparation de trois sources situées dans le plan horizontal, à des azimuths respectifs de 0°, 90° et −120°. Croix : source extraite. Zéros : sources interférentes. Haut : ordre 1. Centre : ordre 2. Bas : ordre 3.

Plus les sources à extraire sont proches spatialement, plus les contraintes liées à la séparation vont avoir une influence visible sur les figures de directivité obtenues. Mathématiquement, on relie le conditionnement de la matrice de mélange avec la capacité à conserver un facteur de directivité élevé lors du *beamforming*. Le cas extrême de plusieurs sources possédant la même incidence est un cas insoluble, où la matrice de mélange n'est plus de rang plein et ne peut pas être inversée. Pour illustrer ce phénomène, la figure 2.4 représente les directivités formées lorsque deux des sources sont éloignées de seulement 30°. A l'ordre 1, les contraintes imposent une directivité éloignée d'un *beamforming* idéal tel que décrit dans la section précédente, en particulier pour la seconde composante qui présente un facteur 3 entre le gain maximal et le gain dans la direction de la source extraite. Cela se traduit par une baisse du facteur de directivité et mécaniquement par un réhaussement du niveau de réverbération et d'interférences résiduelles. L'utilisation de l'ordre 2, et *a fortiori* des ordres supérieurs, permet dans ce cas de conserver une focalisation dans la direction de la source extraite malgré la proximité de sources interférentes. Par ailleurs, les degrés de liberté conservés aux ordres supérieurs permettent d'envisager une optimisation sous contrainte de la formation de voies. Cette piste n'a pas

été approfondie au cours des travaux de thèse.

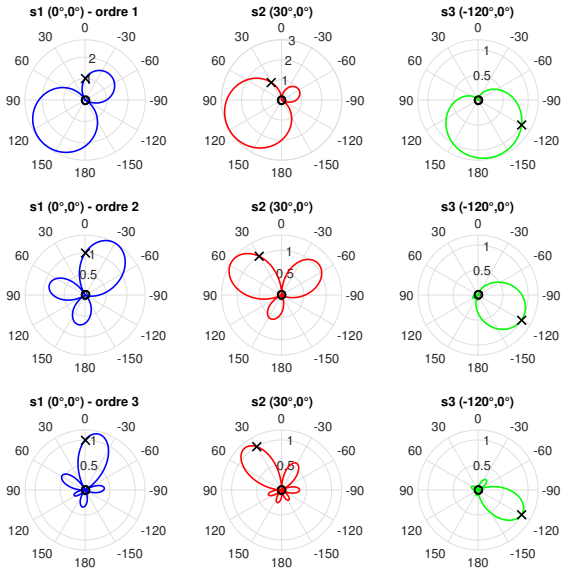

FIGURE 2.4. – Formation de voie pour la séparation de trois sources situées dans le plan horizontal, à des azimuths respectifs de 0°, 30° et −120°. Croix : source extraite. Zéros : sources interférentes. Haut : ordre 1. Centre : ordre 2. Bas : ordre 3.

2.4.4. Synthèse

Nous avons abordé dans cette partie la formalisation de la formation de voies appliquée à un contenu ambisonique, dans le cas d'une focalisation simple ou pour la séparation spatiale de sources. Pour la formation de voie simple, les degrés de libertés disponibles permettent une optimisation de la directivité obtenue, en fonction de critères comme la minimisation des lobes secondaires ou la maximisation du rapport signal sur bruit. Pour la séparation de sources, les contraintes inhérentes réduisent les possibilités d'optimisation du *beamforming* et peuvent potentiellement faire apparaître des phénomènes indésirables, comme une augmentation du niveau de réverbération, en particulier lorsque les sources à séparer sont proches. Ces désagréments se trouvent minimisés par l'utilisation d'ordres ambisoniques plus élevés.

2.5. Conclusion

Ce chapitre a permis de résumer d'un côté les principales approches généralistes de séparation aveugle de sources applicables à un contenu ambisonique : les méthodes parcimonieuses, tensorielles et entropiques, basées sur les statistiques des signaux (ACI) ou sur un modèle acoustique (DUET, DEMIX, DIRAC). Parmi les méthodes d'ACI, des études comparatives disponibles dans la littérature permettent de se contenter d'un nombre restreint d'algorithmes à implémenter pour sélectionner la méthode la plus robuste pour

l'analyse de scène ambisonique.
Des études préliminaires nous ont amenés à écarter les méthodes parcimonieuses diRAC et HARPEX de par leur manque de robustesse et des hypothèses de parcimonie trop restrictives. Le choix des algorithmes de SAS évalués s'est basé d'une part sur les performances comparatives trouvées dans la littérature [26] et d'autre part sur leur représentativité des différentes approches existantes. WASOBI et JADE illustrent les méthodes tensorielles, en diagonalisant des matrices de cumulants respectivement d'ordres 2 et 4. Infomax et EFICA sont des approches basées sur la minimisation de l'entropie où celle-ci est approchée par une fonction non-linéaire. EBM fait intervenir un jeu de plusieurs fonctions pour l'estimation de l'entropie, tandis que ERBM y ajoute une modélisation source-filtre des signaux. Ces six algorithmes, JADE, EFICA, Infomax, WASOBI, EBM et ERBM, sont évalués dans le chapitre suivant pour la séparation de sources appliquée à des mélanges ambisoniques instantanés ou réverbérants.

3. Analyse de scène - expérimentations sur des contenus synthétiques

Les méthodes de SAS décrites dans le chapitre 2 permettent toutes *a priori* d'identifier la matrice de séparation menant aux sources initiales, sous réserve de respecter certaines hypothèses comme l'indépendance des sources et un nombre de sources inférieur ou égal au nombre d'observations. La matrice de mélange estimée, obtenue par pseudo-inversion de la matrice de séparation, permet la localisation des sources dans le cas d'une captation ambisonique parfaite, grâce aux propriétés de l'ambisonie décrites dans le premier chapitre.
Cependant, dans un environnement acoustique réverbérant, l'effet de salle introduit une grande quantité de sources secondaires corrélées aux signaux directs ainsi qu'un champ diffus décorrélé, mettant ainsi à mal l'hypothèse d'indépendance des sources, de même que celle d'un mélange sur-déterminé. L'objectif de ce chapitre est d'évaluer les capacités de localisation et de séparation de sources des principales méthodes de SAS décrites précédemment, en présence ou non de réverbération, à l'aide de contenus ambisoniques synthétiques.
La première partie de ce chapitre est une description détaillée de l'algorithme d'analyse de scène, articulée autour de l'étape de séparation aveugle de sources. Dans une seconde partie sont énoncés les critères d'évaluation utilisés, basés sur la localisation et la séparation obtenues par SAS. Enfin, dans un troisième temps, les expérimentations sont détaillées et les résultats obtenus sont exposés en fonction de plusieurs paramètres comme l'ordre ambisonique, le nombre de sources, la taille de trame d'analyse. Les expérimentations couvrent deux cas de figure : un mélange de sources instantané (encodage microphonique parfait, sans effet de salle) et un cas plus complexe couplant un encodage parfait avec un effet de salle simulé par la méthode des sources images.

3.1. Algorithme d'analyse de scène

Le schéma de la figure 3.1 décrit les principales étapes de l'algorithme d'analyse de scène destiné à évaluer les méthodes de SAS sur des contenus synthéthiques. Afin de se placer dans un contexte temps-réel, le contenu ambisonique est découpé en trames (étape 1), qui sont traitées séparément (étapes 2 à 5).
Ici sont évaluées uniquement les étapes de séparation et de localisation. Les méthodes de comptage de sources mises au point durant la thèse ne sont pas décrites ici, on considérera simplement que l'on connaît le nombre de sources et que l'ambiguïté de permutation est parfaitement levée en se basant sur la direction d'arrivée de référence des sources

encodées.

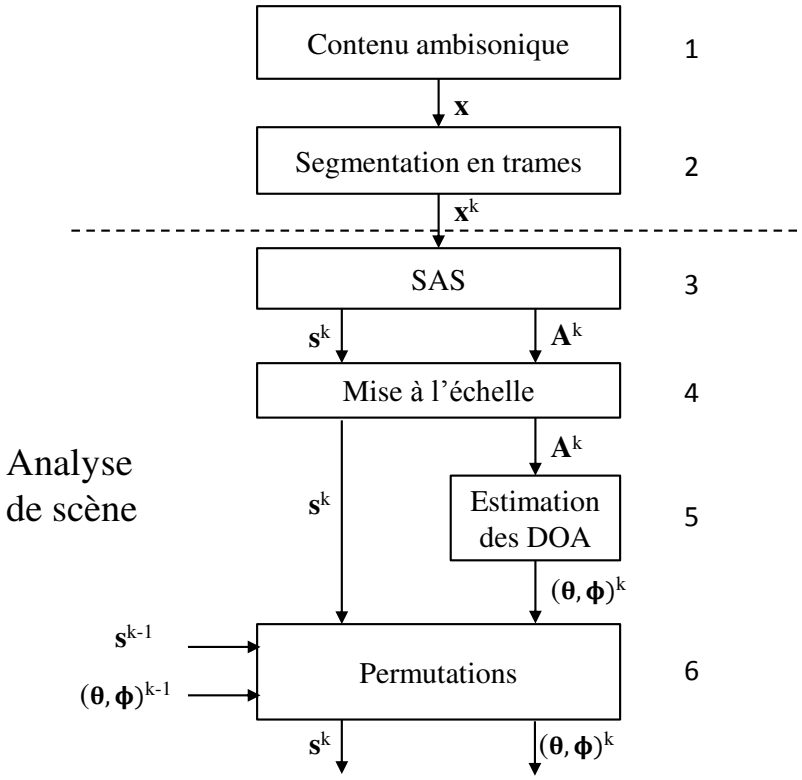

FIGURE 3.1. – Schéma-bloc de l'algorithme d'analyse de scène pour l'évaluation des méthodes de séparation de sources avec un encodage ambisonique parfait.

3.1.1. Génération des contenus ambisoniques

La description détaillée de la génération des contenus ambisoniques et de l'effet de salle synthétique est l'objet de l'annexe A.
Les contenus ambisoniques sont générés à partir de sources de voix monophoniques (en français ou en anglais), encodées avec des coefficients de mélange dépendant de leur direction d'arrivée dans le cas instantané et à l'aide de réponses impulsionnelles synthétiques dans le cas réverbérant. Les sources de voix sont issues d'un corpus d'Orange Labs de phrases phonétiquement équilibrées, l'autre partie provient de la base de données [79] du projet collaboratif SiSEC 2008 portant sur l'évaluation de méthodes de séparation de sources sonores [80].
Des coordonnées sphériques (θ_j,ϕ_j,r_j) sont attribués à chaque source j. Dans le cas instantané, des coefficients d'encodage ambisoniques relatifs aux coordonnées (θ_j,ϕ_j) sont calculés afin de simuler la contribution de la source à chaque observation, selon l'équa-

tion 2.3. Dans le cas réverbérant, l'outil de sources images RoomSim [81] basé sur les travaux d'Allen et Berkley [82] est utilisé pour simuler la réponse d'une salle rectangulaire de dimensions 7 m × 6 m × 4.5 m (figure A.1), possédant un TR_{60} d'environ 350 ms. Chaque réflexions est ici encodée comme une onde plane au format ambisonique en fonction de sa direction d'arrivée. Les réponses des quatre premières composantes ambisoniques sont représentées sur la figure 3.2.

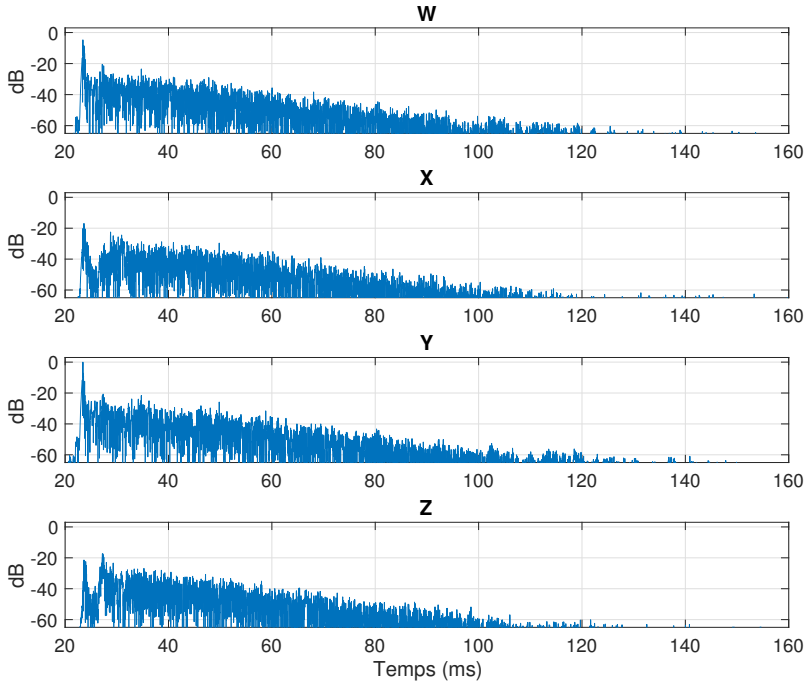

FIGURE 3.2. – SRIR ambisonique d'ordre 1 simulée à partir de l'outil RoomSim pour une source positionnée $(90°, 0°, 2\text{ m})$ par rapport au microphone (amplitude normalisée, en dB)

.

La contribution de chaque source s_j au contenu multicanal généré, appelée source-image et notée $\mathbf{x}_j$, est calculée en convoluant, pour chaque canal m, la source mono par la réponse impulsionnelle de salle $\text{SRIR}_{j,m}$:

$$x_{j,m}(t) = s_j * \text{SRIR}_{j,m} = \sum_{\tau=0}^{T-1} \text{SRIR}_{j,m}(\tau)\dot{s}_j(t-\tau) \tag{3.1}$$

La source-image est donc obtenue par la formule :

$$\mathbf{x}_j(t) = s_j * \mathbf{SRIR}_j = \begin{bmatrix} s_j * \mathrm{SRIR}_{j,1} \\ s_j * \mathrm{SRIR}_{j,2} \\ \dots \\ s_j * \mathrm{SRIR}_{j,M} \end{bmatrix} \tag{3.2}$$

Dans le cas d'un mélange instantané, $\mathrm{SRIR}_{j,m}$ est un dirac dont l'amplitude correspond à la valeur de la $m^{\text{ième}}$ harmonique sphérique dans la direction (θ_j, ϕ_j).
Pour construire une scène $\mathbf{x}$ constituée de N sources sans bruit additionnel, les sources-images sont ensuite sommées :

$$\mathbf{x} = \sum_{j=1}^{N} \mathbf{x}_j \tag{3.3}$$

3.1.2. Analyse-synthèse

La figure 3.3 schématise le procédé d'analyse-synthèse (étape 2) utilisé pour l'analyse trame-à-trame. Le contenu ambisonique $\mathbf{x}$ est découpé de manière à obtenir des trames de taille L se recouvrant sur une durée L_{overlap}. Le recouvrement permet de lisser les estimateurs au cours du temps et d'éviter les discontinuités entre les trames successives. Celui-ci est fixé par la suite à la valeur $L_{\text{overlap}} = \frac{L}{2}$ pour faciliter la reconstruction. La trame k est obtenue de la façon suivante :

$$\mathbf{x}^k[n] = w_{\text{in}}[n]\mathbf{x}[n + (k-1)\frac{L}{2})] \tag{3.4}$$

où w_{in} est une fenêtre de Hann de longueur L.
Après séparation et permutations, les signaux $\mathbf{s}^k$ extraits pour chaque trame k sont sommés par *Overlap-and-Add* pour reconstruire les signaux complets. L'utilisation conjointe d'une fenêtre de Hann et d'un recouvrement de 50% permet une reconstruction parfaite des signaux en sortie.

3.1.3. Séparation aveugle de sources

Un algorithme de séparation aveugle de sources parmi ceux présentés précédemment est appliqué à chaque trame $\mathbf{x}^k$, permettant d'obtenir les signaux estimés $\mathbf{s}^k$ et une matrice de séparation $\mathbf{B}^k$ (étape 3). Six algorithmes ont été évalués, chacun représentant une approche différente pour la séparation de sources : WASOBI, JADE, EFICA, Infomax, EBM et ERBM.
Par la suite, l'indice des trames k sera sous-entendu afin d'alléger les notations, sauf mention contraire.

3.1.4. Résolution des ambiguïtés de signe et d'amplitude

Comme évoqué dans le chapitre précédent, la plupart des méthodes de séparation aveugle de sources ne permettent pas de résoudre les ambiguïtés relatives au signe et à l'amplitude

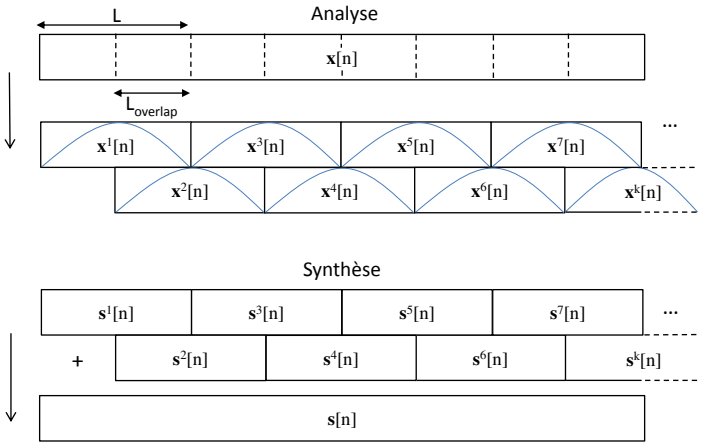

FIGURE 3.3. – Schéma du procédé d'analyse-synthèse. En haut : analyse avec un recouvrement de 50% (fenêtrage en bleu). En bas : reconstruction par *Overlap-and-Add* des signaux extraits.

des sources estimées. En se basant sur le formalisme ambisonique, ces indéterminations peuvent être facilement levées (étape 4). La matrice de mélange estimée $\mathbf{A}$, inverse de la matrice de séparation, est d'abord calculée. Pour chaque colonne $\mathbf{A}_j = [a_{1j}, a_{2j}, \cdots, a_{Mj}]^T$ de la matrice de mélange, le premier terme correspond au coefficient de mélange associé au canal omnidirectionnel (tableau 1.1). Celui-ci étant par définition égal à 1, il suffit de normaliser chaque colonne de la matrice de mélange par son premier terme a_{1j} pour pouvoir ensuite extraire les signaux avec la bonne amplitude et le bon signe en inversant la matrice de mélange normalisée :

$$\begin{cases} \mathbf{A}_j &= \frac{\mathbf{A}_j}{a_{1j}} \quad \forall j \\ \mathbf{B} &= \mathbf{A}^{-1} \end{cases} \tag{3.5}$$

3.1.5. Estimation des directions d'arrivée

L'équation 1.17 permet de relier directement les coefficients de mélange ambisoniques estimés à une direction d'arrivée (étape 5). De manière analogue à la méthode dirAC, les coordonnées sphériques (θ_j, ϕ_j) de la composante s_j sont obtenues à partir de la colonne $\mathbf{A}_j$ de la matrice de mélange estimée et du jeu d'équations suivantes :

$$\begin{cases} \theta_j &= \arctan 2 \frac{a_{3j}}{a_{2j}} \\ \phi_j &= \arctan 2 \frac{a_{4j}}{\sqrt{a_{2j}^2 + a_{3j}^2}} \end{cases} \tag{3.6}$$

Les coefficients relatifs aux ordres supérieurs contiennent également une information directionnelle, permettant de localiser les sources. Néanmoins, l'utilisation de l'ordre 1 est suffisante pour déduire les coordonnées et plus simple d'un point de vue calculatoire.

3.1.6. Résolution de l'ambiguïté de permutation

La troisième ambiguïté relative à la séparation aveugle de sources est la permutation des composantes. Afin d'effectuer la reconstruction, il est nécessaire de placer les sources extraites dans le même ordre d'une trame à l'autre, si tant est qu'elles soient toujours actives (étape 6). L'approche développée durant la thèse est basée sur la distance angulaire entre les composantes successives s_j^{k-1} et s_j^k.

Minimisation de la distance angulaire Une façon de lever l'ambiguïté de permutation est de chercher l'ordre de permutation qui minimise la distance entre les sources s_j^{k-1} et s_j^k. On définit la distance angulaire γ entre les sources s_i et s_j ramenées sur une sphère unité avec comme coordonnées angulaires (θ_i,ϕ_i) et (θ_j,ϕ_j) par la formule suivante :

$$\gamma(s_i, s_j) = \text{acos}(\sin\theta_j \sin\theta_i + \cos\phi_j \cos\phi_i(\cos\theta_j - \cos\theta_i)) \tag{3.7}$$

On définit également le vecteur de permutation $\mathbf{p}$ de dimensions $1 \times N$ contenant les indices de 1 à N permutés et la fonction de coût $c_{\text{angle}}(\mathbf{p})$ telle que :

$$c_{\text{angle}}(\mathbf{p}) = \sum_{j=1}^{N} \gamma(s_j^{k-1}, s_{\mathbf{p}(j)}^k) \tag{3.8}$$

La permutation est réalisée en sélectionnant le vecteur de permutation $\mathbf{p}_{\min}$ qui minimise la fonction de coût c_{angle} :

$$\mathbf{p}_{\min} = \arg\min_{\mathbf{p}} c_{\text{angle}}(\mathbf{p}) \tag{3.9}$$

Recherche exhaustive ou itérative Pour N composantes, le nombre de combinaisons de permutations possibles est de $N!$. Lorsque N est relativement petit (en pratique inférieur ou égal à 7), on choisit de tester de façon exhaustive toutes les combinaisons et de choisir celle minimisant la fonction de coût. Lorsque le nombre de combinaisons devient trop important, une variante itérative est utilisée. Celle-ci consiste à permuter les composantes une à une en recherchant le couple de composantes d'indices j et $\mathbf{p}(j)$ $(s_j^{k-1}, s_{\mathbf{p}(j)}^k)$ les plus proches, en fixant alors $\mathbf{p}(j)$, et en procédant de la même manière par déflation sur les composantes restantes. Pour l'évaluation de l'analyse de scène présentée dans ce chapitre, le nombre de sources et leurs positions respectives sont supposés connus, les permutations se font donc à partir des directions d'arrivées réelles en recherchant les sources estimées les plus proches des sources réelles. On va considérer par la suite que l'ambiguïté de permutation est parfaitement résolue.

3.2. Critères d'évaluation objectifs

La séparation de sources permet d'estimer à la fois les sources et la matrice de mélange associée. On peut donc évaluer l'analyse de scène en se basant soit sur les directions d'arrivées estimées grâce à la matrice de mélange, soit sur la séparation des signaux extraits.

3.2.1. Critères basés sur les directions d'arrivée

La matrice de mélange permet l'estimation de la direction d'arrivée (θ_j,ϕ_j) de chaque composante extraite j (équation 3.6). La différence entre la DOA estimée et la DOA réelle ($\theta_{j,\text{ref}}$,$\phi_{j,\text{ref}}$) de chaque source est ici quantifiée en terme de distance angulaire sur la sphère unité, notée $\gamma(s_j, s_{j,\text{ref}})$, suivant l'équation :

$$\gamma(s_j, s_{j,\text{ref}}) = \text{acos}(\sin\theta_{j,\text{ref}} \sin\theta_j + \cos\phi_{j,\text{ref}} \cos\phi_j (\cos\theta_{j,\text{ref}} - \cos\theta_j)) \tag{3.10}$$

Ce critère prend ainsi en compte l'écart à la position réelle à la fois en azimuth et en élévation. La valeur médiane de la distance angulaire, pour toutes les sources estimées et sur l'ensemble des K trames traitées, est utilisée comme critère d'évaluation :

$$\gamma^{\text{median}} = \text{median}_{k,j}\{\gamma(s_j^k, s_{j,\text{ref}}^k)\} \tag{3.11}$$

Pour quantifier les écarts de localisation en azimuth et en élévation, la valeur absolue de l'erreur commise est utilisée, dont on retient la valeur médiane pour l'ensemble des sources :

$$\text{e}_\theta^{\text{median}} = \text{median}_{k,j}\{|\theta_j^k - \theta_{j,\text{ref}}^k|\} \tag{3.12}$$

$$\text{e}_\phi^{\text{median}} = \text{median}_{k,j}\{|\phi_j^k - \phi_{j,\text{ref}}^k|\} \tag{3.13}$$

3.2.2. Critères basés sur les signaux

3.2.2.1. Toolbox BSSEval

Parmi les outils d'évaluation objective basés sur les signaux extraits, il existe notamment ceux mis au point par E. Vincent *et al.* [83] regroupés dans la *toolbox* BSSEval [84], qui décomposent chaque signal extrait s_j en un signal cible s_{target}, un signal interférent e_{interf}, un bruit résiduel e_{noise} et un ensemble d'artefacts e_{artif}, soit :

$$s_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \tag{3.14}$$

Différents scores sont déduits de ces composantes :

— un rapport signal-sur-interférence (SIR pour *Signal-to-Interference Ratio*) :

$$\text{SIR} = 10\log_{10}\frac{\text{var}\{s_{\text{target}}\}}{\text{var}\{e_{\text{interf}}\}} \tag{3.15}$$

— un rapport signal-sur-artefacts (SAR) :

$$\text{SAR} = 10\log_{10}\frac{\text{var}\{s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}\}}{\text{var}\{e_{\text{artif}}\}} \tag{3.16}$$

— un rapport signal-sur-bruit (SNR) :

$$\text{SNR} = 10\log_{10}\frac{\text{var}\{s_{\text{target}} + e_{\text{interf}}\}}{\text{var}\{e_{\text{noise}}\}} \tag{3.17}$$

— un rapport signal-sur-somme des dégradations (SDR pour *Signal-to-Distorsion Ratio*) :

$$\mathrm{SDR} = 10\log_{10}\frac{\mathrm{var}\{s_{\mathrm{target}}\}}{\mathrm{var}\{e_{\mathrm{interf}} + e_{\mathrm{artif}} + e_{\mathrm{noise}}\}} \tag{3.18}$$

Cette approche possède l'avantage d'être utilisable quel que soit le type de mélange (instantané, convolutif, variant dans le temps). Néanmoins, les scores obtenus restent des valeurs approchées, basées sur la projection des sources extraites sur les signaux de référence. De plus, dans notre cas, l'absence de traitements fréquentiels et de bruit additionnel rend moins pertinente l'utilisation des composantes e_{artif} et e_{noise}.

3.2.2.2. Outils d'évaluation mis en place

Le cadre expérimental mis en place ici permet d'obtenir aisément les valeurs de SIR exactes. Le mélange est instantané ou convolutif, mais les paramètres de mélange restent constants au cours du temps (sources fixes). De plus, les sources estimées sont extraites par combinaison linéaire des observations, via une matrice de séparation estimée. Ainsi, connaissant les sources mono et les réponses impulsionnelles utilisées pour générer la scène, le SIR exact peut être obtenu en appliquant la matrice de séparation à chaque source-image décrite par l'équation 3.2. La séparation de source permet d'obtenir la matrice de séparation $\mathbf{B}$ dont la $i$ème ligne $\mathbf{b}_i$ est un vecteur permettant d'extraire la source-cible estimée $\hat{s}_i(t)$. Ainsi, on peut écrire $\hat{s}_i(t)$ telle que :

$$\hat{s}_i(t) = \mathbf{b}_i\mathbf{x} = \mathbf{b}_i\sum_{j=1}^{N}\mathbf{x}_j = \sum_{j=1}^{N}\mathbf{b}_i\mathbf{x}_j \tag{3.19}$$

Le SIR est alors donné par le rapport entre l'énergie de la contribution de la source-image cible et l'énergie des contributions interférentes, soit pour la source s_i :

$$\mathrm{SIR}_i^{tot} = 10\log_{10}\frac{\mathrm{var}\{\mathbf{b}_i\mathbf{x}_i(t)\}}{\mathrm{var}\{\sum_{j=1,j\neq i}^{N}\mathbf{b}_i\mathbf{x}_j(t)\}} \tag{3.20}$$

Ce score prend en compte la totalité des contributions de chaque source pour chaque composante i extraite, soit le champ direct et le champ réverbéré. En présence de réverbération, on peut affiner cette analyse en décomposant les SRIR en champ direct / champ reverbéré, par un simple fenêtrage temporel (voir figure 1.13) :

$$\mathbf{SRIR} = \mathbf{SRIR}^{\mathrm{direct}} + \mathbf{SRIR}^{\mathrm{reverb}} \tag{3.21}$$

On forme alors les sources-images directes :

$$\mathbf{x}_j^{\mathrm{direct}}(t) = s_j * \mathbf{SRIR}_j^{\mathrm{direct}} \tag{3.22}$$

Cette approche permet de calculer un SIR uniquement basé sur l'annulation du champ direct, soit :

$$\mathrm{SIR}_i^{\mathrm{direct}} = 10\log_{10}\frac{\mathrm{var}\{\mathbf{b}_i\mathbf{x}_i^{\mathrm{direct}}(t)\}}{\mathrm{var}\{\sum_{j=1,j\neq i}^{N}\mathbf{b}_i\mathbf{x}_j^{\mathrm{direct}}(t)\}} \tag{3.23}$$

Les signaux étant non stationnaires, les sources-images utilisées pour le calcul des SIR sont normalisées en énergie pour chaque trame : l'énergie de la source-image cible est normalisée, ainsi que la somme des sources-images interférentes, de façon à avoir un SIR initial de 0 dB.

3.3. Analyse d'un mélange instantané

Pour évaluer les performances des méthodes d'ACI, des contenus comprenant entre $N = 2$ et $N = 8$ sources simultanées sont générés, puis analysés avec des trames d'analyse de tailles variables, entre 512 points et 16384 échantillons. Cela correspond pour des signaux à 16 kHz à des durées comprises entre 32 ms et 1 s.
Pour les scènes constituées de 2 ou 3 sources, celles-ci sont disposées dans le plan horizontal avec un espacement compris entre 10° et 150°. Au-delà, les sources sont disposées de façon quasi-régulière sur la sphère, de façon à conserver une matrice de mélange de rang plein. Pour 8 sources réparties régulièrement, cela revient à avoir une distance angulaire caractéristique de 60° entre les sources.
Un encodage ambisonique à l'ordre 1 est appliqué lorsque $N \leqslant 4$, l'ordre 2 est utilisé dès lors que $N \geqslant 4$, sachant que dans le cas instantané le blanchiment des données réduit la dimension du problème au nombre de sources effectives.

3.3.1. Localisation des sources

La figure 3.4 présente les résultats obtenus en termes d'erreur angulaire (équation 3.11) commise sur l'estimation des DOA des sources extraites. Deux principales observations peuvent être faites. Tout d'abord, l'ensemble des algorithmes convergent vers la bonne direction d'arrivée lorsque les trames d'analyse sont suffisamment longues, et ce quel que soit le nombre de sources dans le mélange.
Ensuite, les scores obtenus montrent clairement une précision accrue de l'algorithme ERBM par rapport aux autres méthodes, avec une erreur médiane inférieure à 4° quels que soient le nombre de sources et la longueur des trames. *A contrario*, l'erreur de localisation avec JADE et Infomax est significativement supérieure aux autres méthodes. Par ailleurs, on peut considérer qu'en présence de 6 ou 8 sources, une erreur angulaire médiane supérieure à 20° est équivalente à une localisation aléatoire, étant donné la faible distance entre les sources sur la sphère.
On note enfin que parmi les autres méthodes, WASOBI est la plus performante sur des trames courtes (moins de 4096 échantillons), progressivement rattrapée par EBM et EFICA pour des trames plus longues.

3.3.2. Séparation des sources

Les SIR normalisés (tels que définis par l'équation 3.20) sont représentés sur la figure 3.5. Comme on peut s'y attendre, les scores de SIR sont étroitement corrélés avec la capacité à localiser précisément les sources. Ainsi, l'augmentation de la taille de trame permet d'augmenter le SIR pour l'ensemble des algorithmes. ERBM présente des performances

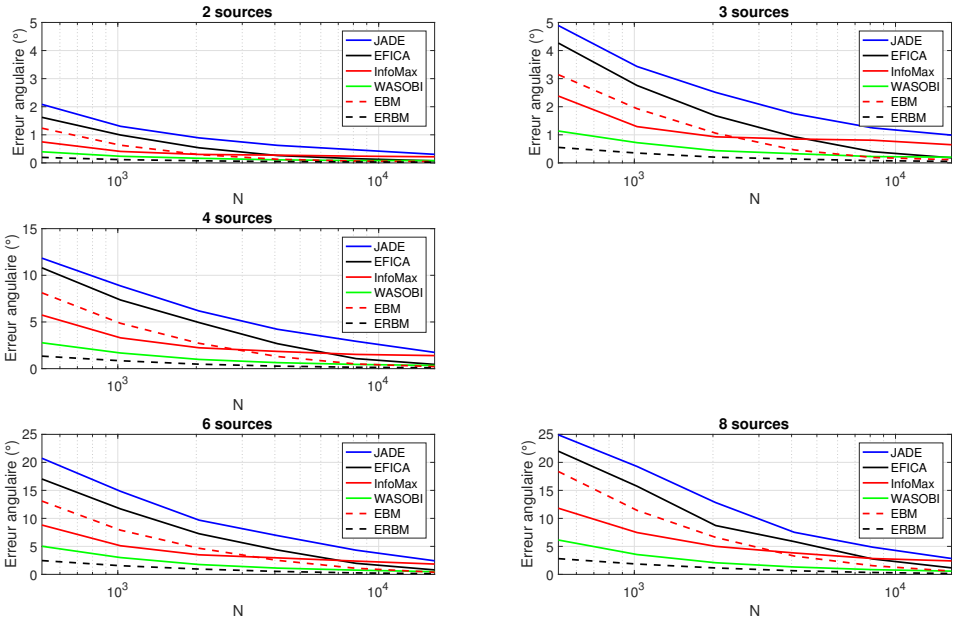

FIGURE 3.4. – Mélange instantané de signaux de parole. Erreurs de DOA (valeur mediane, en degrés) en fonction de la taille de trame. Algorithmes testés : JADE, WASOBI, Infomax, EFICA, EBM et ERBM.

de séparation robustes, avec un SIR médian supérieur à 20 dB même dans la configuration la moins favorable (8 sources, trames de 512 points). ERBM présente ainsi une amélioration d'au moins 6 dB par rapport à WASOBI ou EBM, pour l'ensemble des configurations testées.
L'augmentation de la taille de trame permet pour l'ensemble des algorithmes une augmentation du SIR, avec des scores supérieurs à 20 dB pour des trames d'une seconde ($N = 16384$ points), même avec l'algorithme le moins performant (JADE).
Inversement, l'augmentation du nombre de sources entraîne une dégradation du SIR, expliquée par la complexité accrue du problème et la nécessité d'avoir une quantité de données plus importante pour aboutir à une séparation de qualité équivalente. ERBM nécessite ainsi environ 2048 points pour obtenir un SIR de 40 dB dans le cas d'un mélange de 8 sources, au lieu de 512 points en présence de 2 sources.
Au vu de ces résultats, il apparaît clairement que ERBM est l'algorithme le plus performant pour localiser et extraire des signaux de voix dans un mélange instantané. WASOBI semble être une bonne alternative à ERBM pour travailler sur des trames courtes, mais présente un infléchissement des performances au-delà de 4096 échantillons. EBM et EFICA ont des performances analogues, avec des performances en retrait sur des trames courtes mais une forte progression au-delà de 2048 points d'analyse. InfoMax est classé 3^{e} pour des trames inférieurs à 1024 points, mais est distancé lorsque la taille de trame augmente. Quant à JADE, ses performances sont systématiquement en-deçà des autres

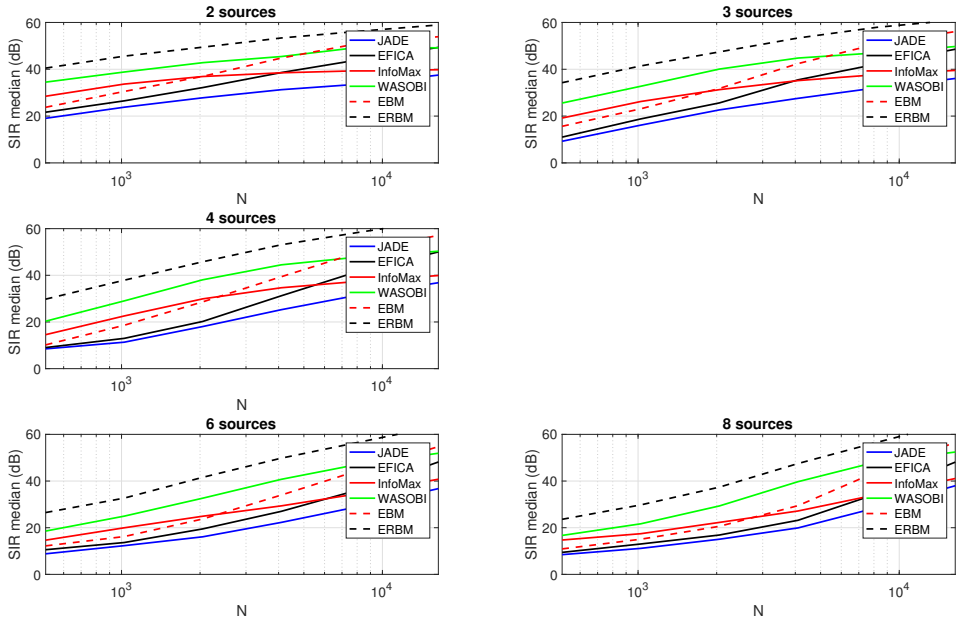

FIGURE 3.5. – Mélange instantané de signaux de parole. SIR médian (dB) en fonction de la taille de trame. Algorithmes testés : JADE, WASOBI, Infomax, EFICA,EBM et ERBM.

méthodes évaluées.

3.4. Mélange réverbérant

Ici sont présentés les résultats de localisation et de séparation en présence d'un effet de salle synthétique (salle virtuelle décrite en Annexe A). L'étude est restreinte au cas de deux sources, positionnées à 2.3 m d'un microphone ambisonique virtuel, avec comme coordonnées sphériques respectives $(0°, 15°)$ et $(120°, 0°)$ par rapport au micro.

3.4.1. Localisation des sources

Les algorithmes ERBM, EBM, WASOBI et EFICA, appliqués pour des trames de taille variable, donnent les performances de localisation présentées en figure 3.6. On remarque tout d'abord que pour un contenu d'ordre 1, la hiérarchie entres algorithmes est la même que pour le cas instantané, avec des erreurs de localisation pour ERBM et WASOBI encore une fois nettement inférieures à EFICA et EBM, même si des trames longues tendent à réduire l'écart entre les différents algorithmes.
L'apport des composantes d'ordre 2 pour la localisation apparaît sur la colonne de droite, en particulier pour ERBM dont l'erreur angulaire médiane est divisée par deux, passant de 10° à 5°. Pour WASOBI en revanche, la localisation en élévation est moins précise qu'à

l'ordre 1 tandis que la localisation azimutale reste aussi précise. La présence de premières réflexions en provenance du sol et du plafond semble ainsi perturber la localisation malgré la présence d'un nombre plus élevé de composantes ambisoniques.

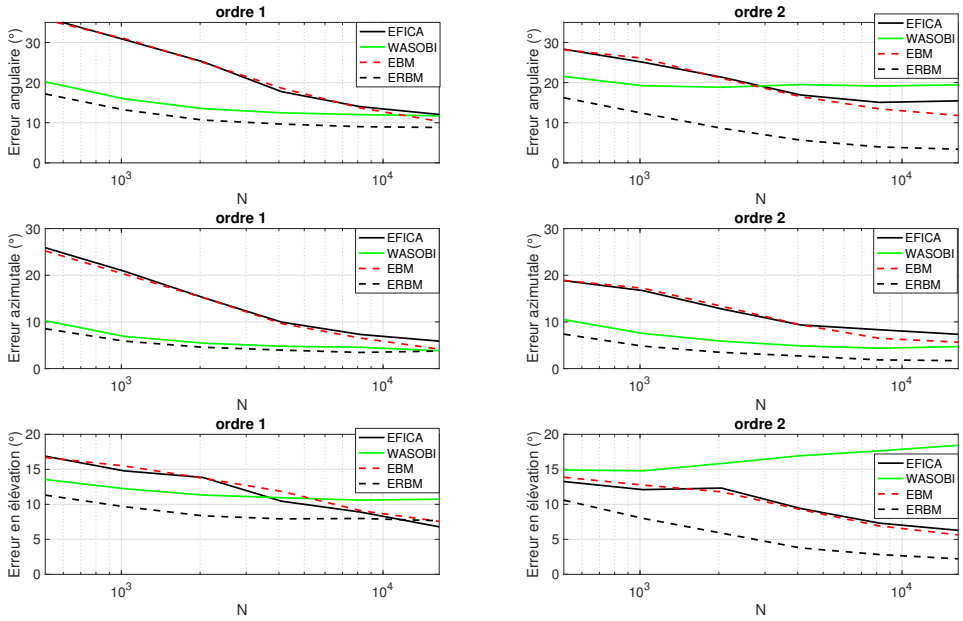

FIGURE 3.6. – Mélange réverbérant de deux signaux de parole. Erreurs de DOA (valeur mediane, en degrés) en fonction de la taille de trame et de l'ordre ambisonique. Algorithmes testés : WASOBI, EFICA, EBM et ERBM.

3.4.2. Séparation des sources

Les performances en termes de séparation de sources sont étudiées suivant deux critères : le SIR direct (équation 3.23) et le SIR total (équation 3.20) prenant en compte la contribution totale de chaque source.
Concernant l'annulation des chemins directs (figure 3.7), les performances à l'ordre 1 sont en adéquation avec les résultats de localisation, ERBM et WASOBI affichant des scores entre 20 dB et 30 dB suivant la longueur de trame, tandis que EFICA et EBM sont en retrait d'environ 10 dB.
A l'ordre 2, les scores obtenus sont sensiblement identiques, mais on note des résultats étonnamment élevés pour WASOBI, si l'on se réfère à l'erreur d'estimation de l'élévation. La configuration géométrique de la scène explique ces résultats, car la séparation de sources positionnées quasiment dans le même plan azimutal est plus sensible aux erreurs de localisation azimutale. On le visualise sur la figure 3.8, avec la comparaison des directivités formées pour extraire la source 1, lorsque la source 2 est exactement localisée (en haut) et lorsqu'une erreur de 20° est commise, soit en élévation (au centre) ou bien en azimuth (en bas). On visualise clairement qu'une erreur en élévation n'aura que

très peu d'effet sur la qualité de la séparation du fait de la polarisation azimutale du *beamforming*, alors qu'en cas d'erreur azimutale, la source 2 se retrouve sur un lobe de directivité secondaire, dégradant le SIR direct.

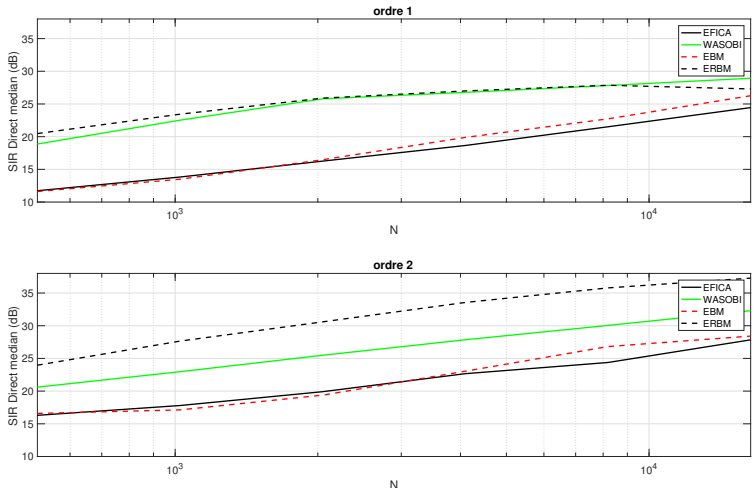

FIGURE 3.7. – Mélange réverbérant synthétique de deux signaux de parole. SIR direct médian (dB) en fonction de la taille de trame et de l'ordre ambisonique. Algorithmes testés : WASOBI, EFICA, EBM et ERBM

Le SIR total se retrouve sur la figure 3.9 et fait apparaître des performances à l'ordre 1 similaires entre ERBM et WASOBI d'un côté et EFICA et EBM de l'autre. Les scores, inférieurs alors à 7 dB s'expliquent par le faible facteur de directivité de la formation de voies à l'ordre 1 ambisonique (section 2.4), entraînant la présence d'une réverbération résiduelle non négligeable, même en cas de bonne localisation des sources.
A l'ordre 2 toutefois, la sélectivité accrue permet d'améliorer le SIR entre 2 et 3 dB pour ERBM et WASOBI. De plus, la meilleure localisation associée à ERBM par rapport à WASOBI entraîne une différence d'environ 1 dB pour des trames courtes, qui tend à s'estomper pour des trames longues lorsque la localisation azimutale de WASOBI se précise.
Au vu de ces résultats, il apparaît que ERBM est plus robuste à la présence d'effet de salle que WASOBI, en termes de localisation et de séparation des champs directs. Les performances d'EFICA et EBM sont nettement en deçà, que ce soit en termes de localisation ou de séparation.

3.5. Blanchiment temporel des données par ERBM

La seule différence entre ERBM et EBM résidant en une décorrélation temporelle des sources, il est intéressant de se pencher sur l'apport de cette étape sur la localisation en milieu réverbérant.
ERBM (*Entropy Rate Bound Minimization*) est un algorithme d'ACI basé sur la mi-

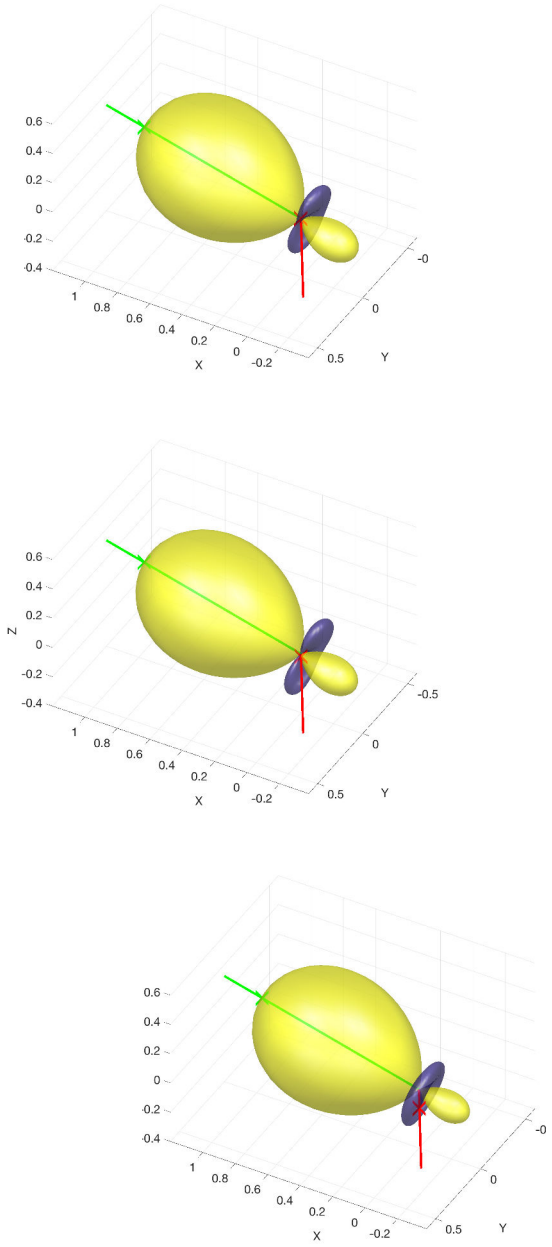

FIGURE 3.8. – Beamforming ordre 2 pour séparer deux sources, cible $(0°, 15°)$ et interférente $(120°, 0°)$. Haut : localisation exacte. Centre : erreur de 20° en élévation sur la source interférente. Bas : erreur de 20° en azimuth sur la source interférente. Croix : positions réelles.

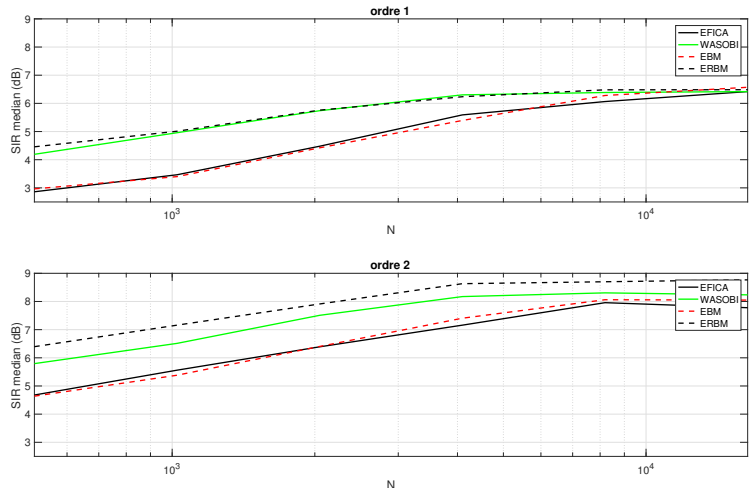

FIGURE 3.9. – Mélange réverbérant synthétique de deux signaux de parole. SIR médian (dB) en fonction de la taille de trame. Algorithmes testés : WASOBI, EFICA, EBM et ERBM.

nimisation du ratio d'entropie (*entropy rate*) des sources estimées (voir section 2.2.3). Au cours de la séparation, ERBM procède à un filtrage de type MA (*Moving Average*) des composantes estimées afin de les rendre indépendantes et identiquement distribuées (i.i.d). On parlera par la suite de signaux "blanchis" temporellement.
Dans un signal i.i.d., chaque échantillon est indépendant des autres avec une même loi de probabilité. L'entropie d'un tel signal est équivalente à son ratio d'entropie, qui correspond au taux d'entropie par échantillon, ou qui peut être vu comme l'information apportée par un échantillon connaissant tous les échantillons précédents. L'algorithme procède alors à la minimisation du ratio d'entropie des signaux i.i.d. estimés, en optimisant alternativement la matrice de séparation et le filtre associé à chaque source par une descente de gradient.
Le blanchiment des données était initialement utilisé pour réduire la variance des estimateurs statistiques et donc augmenter la robustesse de l'analyse. Nous avons observé que le blanchiment des signaux permettait également, dans le cas d'un mélange réverbérant, de discriminer le champ direct des premières réflexions et donc améliorer la localisation des sources. Nous allons mettre ici cette propriété en avant, par l'étude du cas simple d'un mélange instantané d'une source et d'une réflexion, version retardée et atténuée de la source principale.

3.5.1. Blanchiment des données

Le blanchiment de la composante s_i s'apparente à une étape de déconvolution permettant l'extraction d'un signal i.i.d. u_i à partir d'un filtre a_i de longueur p par l'opération :

$$u_i = a_i * s_i \tag{3.24}$$

Concrètement, une matrice de convolution $\mathbf{C}$ est construite à partir de s_i, telle que :

$$\mathbf{C}[n] = \begin{bmatrix} s_i[n] \\ s_i[n-1] \\ ... \\ s_i[n-p+1]) \end{bmatrix} \tag{3.25}$$

pour calculer $u_i[n]$:

$$u_i[n] = a_i^T \cdot \mathbf{C}[n] \tag{3.26}$$

L'estimation de a_i se fait par minimisation de l'entropie de $u_i[n]$, avec une approche analogue à celle d'EBM décrite en 2.2.3. Lorsque le signal traité est effectivement issu d'un signal i.i.d coloré, l'étape de déconvolution permet de retrouver quasiment à l'identique le signal excitateur, sous réserve que le filtre initial soit inversible à l'aide d'un modèle MA de longueur p.
Le résultat du blanchiment est illustré par la figure 3.10. Celle-ci représente la fonction d'autocorrélation d'un signal i.i.d. centré et de loi uniforme (en rouge), celle du signal i.i.d coloré par un filtre MA (20 coefficients) en bleu et celle du signal blanchi par le module de déconvolution d'ERBM (en pointillés noirs) en prenant $p = 10$. Le blanchiment est ici effectué correctement au second ordre, avec une fonction d'autocorrélation du signal estimé se rapprochant d'un dirac, même si des oscillations apparaissent à $\tau = \pm 20$ échantillons, artefacts causés par le filtre de blanchiment. Le graphique de droite illustre un cumulant croisé d'ordre 4 (kurtosis normalisé) dont la formule générale pour une variable $s[n]$ et sa version retardée $s[n-\tau]$est :

$$\mathrm{Cum}^4(\tau) = \mathrm{E}\{s^2[n]s^2[n-\tau]\} - E\{s^2[n]\}E\{s^2[n-\tau]\} - 2\,\mathrm{E}\{s[n]s[n-\tau]\}^2 \tag{3.27}$$

De façon analogue à la fonction d'autocorrélation, le kurtosis tend à s'annuler pour $\tau \neq 0$ après l'étape de blanchiment, ce qui caractérise l'indépendance entre les variables $u[n]$ et $u[n-\tau]$ pour τ non nul, même si la taille limitée du filtre de blanchiment ne permet pas de supprimer complètement la coloration du signal comme en témoignent les oscillations du kurtosis. On remarque par ailleurs un kurtosis négatif en $\tau = 0$, caractéristique d'une variable possédant une distribution uniforme.
En utilisant cette fois un filtre blanchisseur de $p = 50$ échantillons, on obtient le résultat de la figure 3.11. L'autocorrélation et le kurtosis s'annulent ici quasiment parfaitement dès lors que $\tau \neq 0$.
Lorsque le signal coloré est cette fois un signal de parole de 4 s (voix de femme, figure 3.12), un blanchiment efficace peut être obtenu en prenant $p = 5$. Le kurtosis conserve néanmoins une valeur non nulle pour l'ensemble des τ car il est impossible de blanchir un signal de parole aussi parfaitement qu'un signal synthétique. La qualité du blanchiment temporel est donc fonction à la fois de la longueur du filtre estimé et de la nature des signaux, notamment de la validité du modèle i.i.d + filtre et du caractère inversible du filtre.

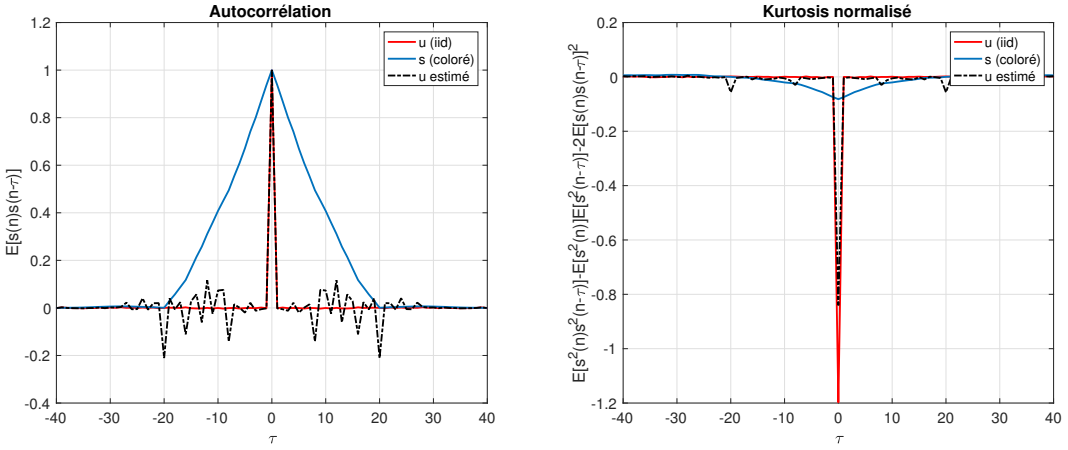

FIGURE 3.10. – Blanchiment d'un signal coloré synthétique. Cumulants d'ordre 2 (gauche) et 4 (droite). En rouge : signal i.i.d.. En bleu : signal coloré. En pointillés noirs : signal blanchi. Filtre de blanchiment de $p = 10$ échantillons.

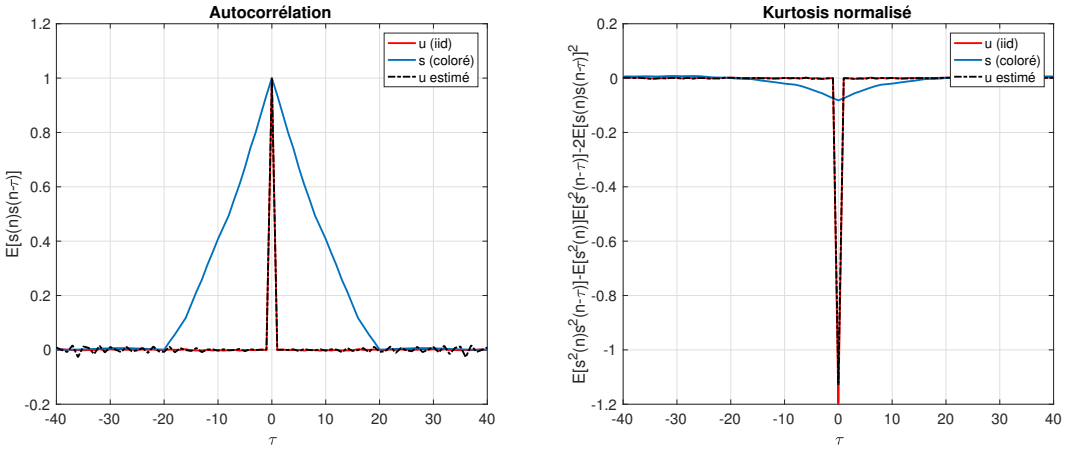

FIGURE 3.11. – Blanchiment d'un signal coloré synthétique. Cumulants d'ordre 2 (gauche) et 4 (droite). En rouge : signal i.i.d.. En bleu : signal coloré. En pointillés noirs : signal blanchi. Filtre de blanchiment de $p = 50$ échantillons.

3.5.2. Identification signal direct/signal réfléchi

La plupart du temps, la présence d'effet de salle dégrade la localisation et séparation des sources, en grande partie à cause de l'incapacité des méthodes d'ACI à dissocier les signaux directs des premières réflexions. Cela s'explique par la forte corrélation liant généralement le champ direct et les premières réflexions.
En revanche, cet inconvénient disparaît si les sources sont i.i.d et que l'on considère les réflexions comme des versions retardées et atténuées du signal principal. En effet, une propriété fondamentale d'un signal i.i.d. $s_1[n]$ est que toute version retardée

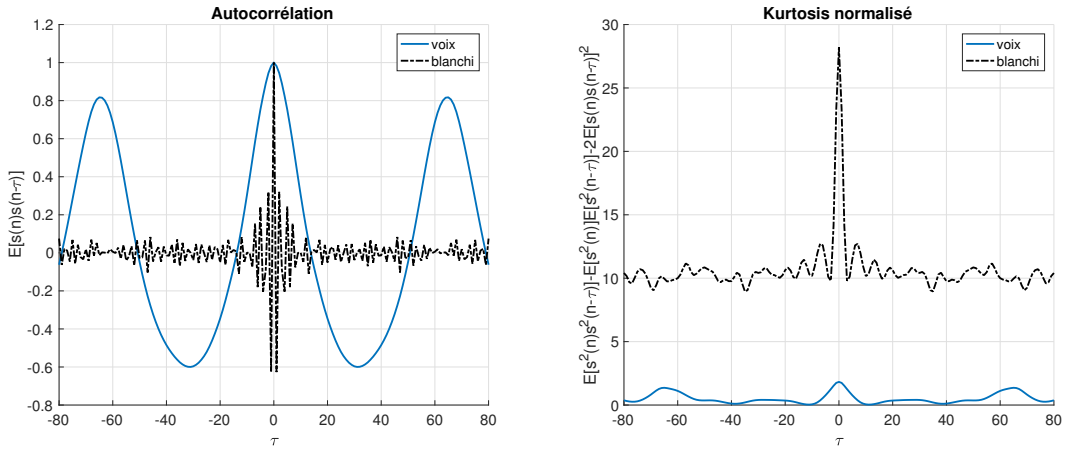

FIGURE 3.12. – Blanchiment d'un signal de parole. Cumulants d'ordre 2 (gauche) et 4 (droite). En bleu : signal vocal. En pointillés noirs : signal blanchi. Filtre de blanchiment de $p = 5$ échantillons.

$s_2[n] = a_2 \cdot s_1[n - \tau_2]$ lui est statistiquement indépendant dès lors que $\tau_2 \neq 0$. Ainsi, la plupart des méthodes d'ACI sont capables de procéder à la séparation des signaux s_1 et s_2, même si physiquement les deux signaux sont issus de la même source d'excitation, à condition que celle-ci soit i.i.d..

On perçoit alors l'intérêt d'ERBM pour l'identification du champ direct et des premières réflexions en présence de sources sonores structurées temporellement, car la modélisation i.i.d. décrite précédemment permet théoriquement de rendre le signal direct et les signaux réfléchis statistiquement indépendants, et donc séparables.

En conservant la même source colorée que précédemment, la figure 3.13 illustre sur le graphique de gauche la fonction d'autocorrélation (en bleu) d'un mélange source-réflexion $x[n] = s_1[n] + s_2[n] = s_1[n] + a_2 \cdot s_1[n - \tau_2]$ ($a_2 = 0.4$, $\tau_2 = 15$ échantillons), qui ne permet pas de discriminer la contribution du signal direct et celle du signal réverbéré. La fonction d'autocorrélation du mélange blanchi par un filtre de longueur $p = 10$ permet alors de distinguer nettement les deux contributions avec des pics de corrélation pour $\tau = \pm\tau_2$ et il en est de même pour le cumulant d'ordre 4.

Le résultat du blanchiment pour un mélange de signaux de voix, un retard $\tau_2 = 27$ échantillons et un blanchiment de 4 échantillons est visible sur la figure 3.14. Comme pour le signal synthétique précédent, le blanchiment permet de visualiser clairement les contributions du champ direct et du champ réfléchi, sans pour autant obtenir des signaux parfaitement i.i.d..

Il est tentant de vouloir augmenter la longueur du filtre de blanchiment afin de rendre les signaux estimés les plus indépendants possibles. Cependant, un filtre trop long, en particulier supérieur à τ_2, va blanchir à la fois le signal direct et le signal réfléchi. La réflexion sera alors vue comme une coloration supplémentaire que l'algorithme de blanchiment tentera d'annuler. Le problème est que le signal direct et la réflexion ne sont

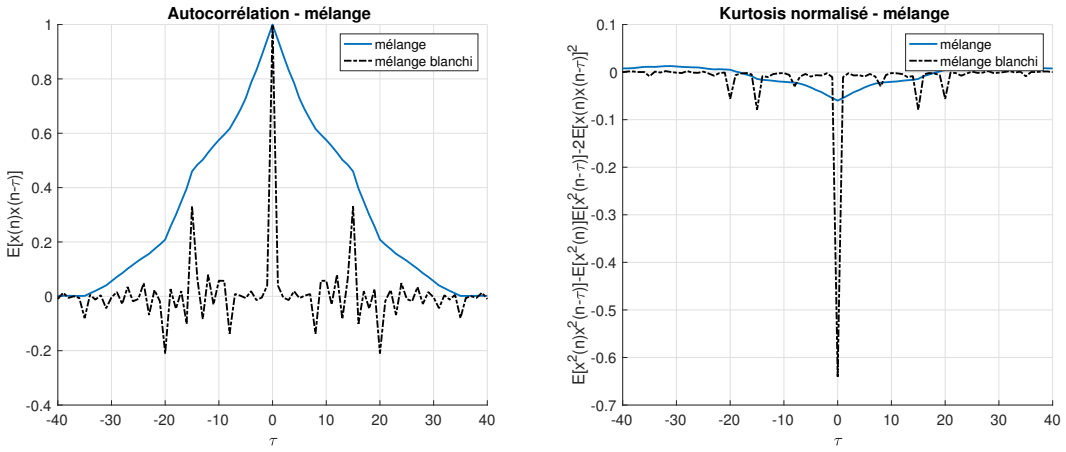

FIGURE 3.13. – Blanchiment d'un signal coloré synthétique (mélange source-réflexion, $\tau_2 = 15$ échantillons). Cumulants d'ordre 2 (gauche) et 4 (droite). En bleu : mélange coloré. En pointillés noirs : mélange blanchi. Filtre de blanchiment de $p = 10$ échantillons.

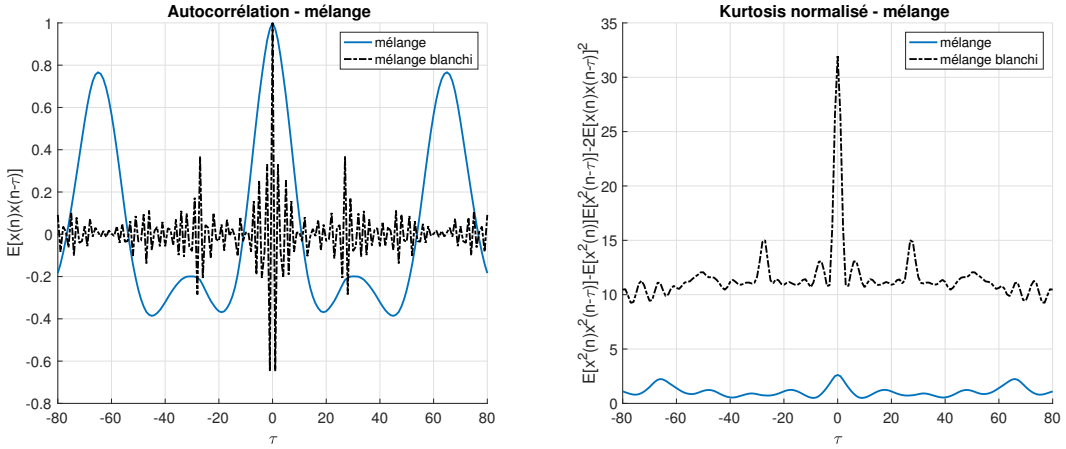

FIGURE 3.14. – Blanchiment d'un signal de parole (mélange source-réflexion, $\tau_2 = 27$ échantillons). Cumulants d'ordre 2 (gauche) et 4 (droite). En bleu : mélange. En pointillés noirs : mélange blanchi. Filtre de blanchiment de $p = 5$ échantillons.

pas colinéaires, il devient donc impossible pour l'algorithme d'estimer correctement les coefficients de mélange du signal direct, et donc les informations de localisation. De plus, la direction pointée lors de la séparation va varier à chaque itération du gradient de la matrice de séparation $\mathbf{W}$, ce qui fait varier le poids relatif des signaux directs et réverbérés dans les composantes extraites et empêche une convergence correcte des estimateurs d'entropie. On retrouve ainsi sur la figure 3.15 les estimateurs statistiques d'ordre 2 et 4 du mélange blanchi en prenant $p = 50$. La résultante est cette fois quasiment i.i.d et la

contribution du signal réfléchi n'est plus visible, avec un seul pic de corrélation en $\tau = 0$.

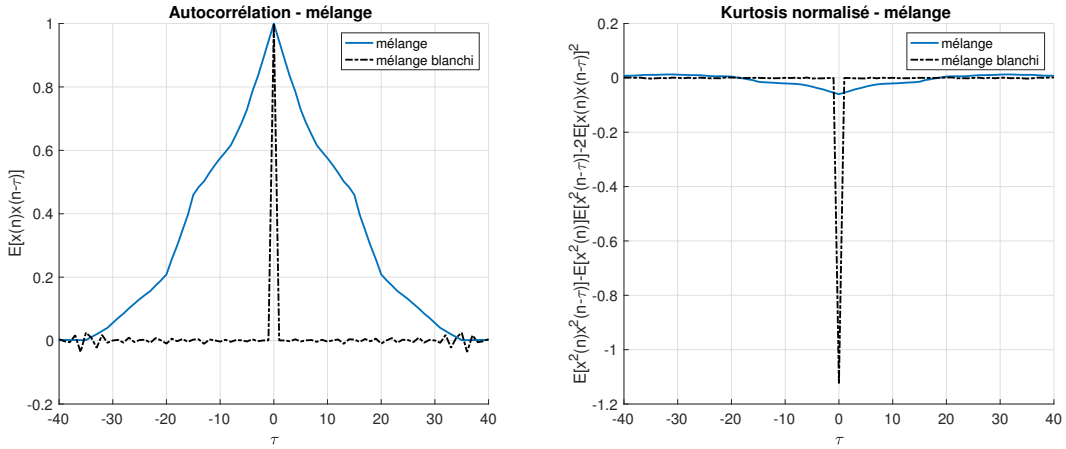

FIGURE 3.15. – Blanchiment d'un signal coloré synthétique (mélange source-réflexion, $\tau_2 = 15$ échantillons). Cumulants d'ordre 2 (gauche) et 4 (droite). En bleu : mélange coloré. En pointillés noirs : mélange blanchi. Filtre de blanchiment de $p = 50$ échantillons.

En pratique, il a été observé que pour des signaux de parole, un filtre dont la longueur est comprise entre $p = 5$ et $p = 10$ échantillons était généralement suffisant pour tirer pleinement avantage du module de blanchiment. Cela correspond à une différence de marche de 10-20 cm pour une fréquence d'échantillonnage de 16 kHz, largement inférieure à la différence de marche source-réflexion. En pratique, la distance signal direct-signal réfléchi est donc rarement un facteur limitant.

3.5.3. Limitations pour des signaux périodiques

Un signal périodique peut difficilement être approché par un modèle signal i.i.d. + filtre. De plus, le ratio d'entropie d'un signal périodique est nul. En effet, la connaissance d'un nombre d'échantillons égal ou supérieur à une période d'oscillation suffit à prédire parfaitement les échantillons suivants, ceux-ci ne portent donc pas d'information. Des études préliminaires réalisées sur des mélanges de sources instantanés ont confirmé la difficultés pour ERBM d'opérer en présence de signaux musicaux ou périodiques. En pratique, les signaux traités ne sont jamais parfaitement périodiques, mais les limitations de la modélisation i.i.d.-filtre expliquent les mauvaises performances observées, l'utilisation du ratio d'entropie n'étant pas appropriée à ce type de signaux.

3.6. Conclusion

Nous avons dans ce chapitre détaillé la structure de l'algorithme d'analyse de scène et les premiers résultats obtenus pour des mélanges ambisoniques instantanés et réverbérants.

Ceux-ci ont mis en évidence les performances de l'algorithme d'analyse en composantes indépendantes ERBM par rapport aux autres méthodes de la littérature que sont JADE, EFICA, Infomax et WASOBI. Les résultats obtenus ont par ailleurs fait l'objet d'un acte de congrès [85] et d'une présentation à la 60ème conférence de l'AES à Louvain. La modélisation source-filtre opérée par ERBM pour chaque composante estimée permet de limiter l'influence des premières réflexions sur la localisation des sources en milieu réverbérant. Les propriétés associées à cette modélisation permettent à ce stade d'envisager l'utilisation de l'algorithme à des contenus ambisoniques réels, ce qui est l'objet du chapitre suivant.

4. Analyse de contenus ambisoniques réels

Le chapitre 3 a permis de mettre en évidence la supériorité de l'algorithme ERBM par rapport à d'autres méthodes d'ACI pour localiser et séparer des sources, dans le cas de mélanges instantanés et de contenus réverbérants synthétiques.
La question qui se pose à ce stade est la suivante : qu'en est-il des contenus réels ?
L'analyse de contenus réels nécessite de prendre en compte certaines modifications par rapport aux simulations numériques, qui concernent les propriétés physiques des sources (signal émis, géométrie, directivité), l'effet de salle et le système de captation.
La directivité des sources et le modèle de propagation acoustique évoluent. Les sources ponctuelles omnidirectionnelles deviennent des sources réelles, possédant un certain étalement et un rayonnement anisotrope. Par ailleurs, l'hypothèse onde plane n'est valide que lorsque la distance source-microphone est grande devant la longueur d'onde.
L'effet de salle est également plus complexe que le modèle paramétrique utilisé précédemment et inclut des phénomènes tels que la diffusion et la diffraction du champ acoustique dues aux obstacles rencontrés par l'onde sonore.
L'encodage microphonique, qui jusqu'à présent respectait parfaitement les équations théoriques du formalisme ambisonique, est maintenant sujet aux dégradations induites par les limitations physiques du système de captation. Il devient alors nécessaire d'adapter le traitement des contenus en fonction de la validité ou non de l'encodage ambisonique pour chaque bande de fréquences.
Dans ce quatrième chapitre, on propose donc de poursuivre l'étude des performances de l'algorithme d'analyse de scène basé sur ERBM en l'appliquant cette fois à des contenus ambisoniques réels. On décrit tout d'abord les conditions expérimentales ainsi que les modifications algorithmiques introduites pour s'adapter aux conditions réelles, avant de détailler les résultats obtenus à la fois en termes d'analyse de scène et de séparation de sources.

4.1. Génération des contenus sonores

Les scènes sonores sont créées à partir de réponses impulsionnelles microphoniques mesurées en conditions réelles, converties en réponses ambisoniques par une matrice de filtres. Comme pour les contenus synthétiques, des sources de voix sont convoluées par les réponses impulsionnelles puis leurs contributions respectives sont additionnées. L'annexe B détaille la procédure d'acquisition des réponses impulsionnelles de salle et les différents environnements acoustiques.

Une salle, de dimensions 5.15 m × 4.08 m × 2.30 m, dont les murs sont équipés de panneaux absorbants amovibles, est exploitée pour générer les réponses impulsionnelles de salles. Deux configurations acoustiques ont été retenues pour la génération des contenus :

— Une configuration absorbante (figure B.1) dans laquelle les parois sont intégralement recouvertes de blocs de mousse, dont le TR_{60} est d'environ 120 ms.

— Une configuration réverbérante (illustré par la figure 4.1), pour laquelle les panneaux absorbants sont enlevés dans leur intégralité. Le TR_{60} associé est alors de 350 ms.

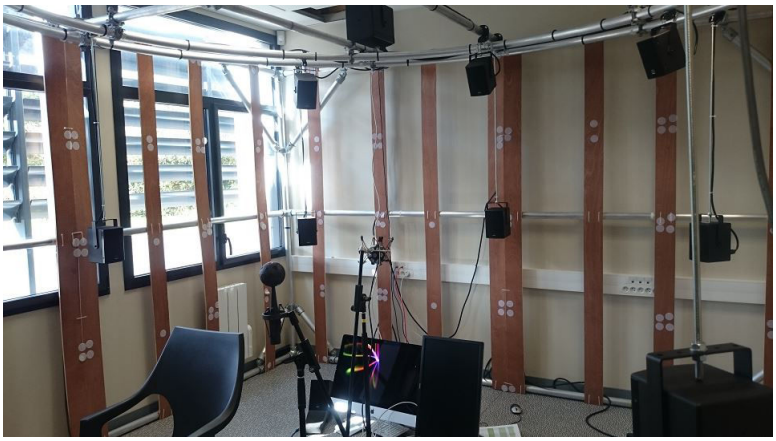

FIGURE 4.1. – Salle aux parois peu absorbantes utilisée pour les mesures de réponses impulsionnelles.

Sur la figure 4.2, on visualise une réponse impulsionnelle omnidirectionnelle dans les deux configurations de salle, avec dans les deux cas un premier front d'onde aux alentours de 23 ms, suivi de la première réflexion à 27 ms causée par le sol de la pièce (différence de marche d'environ 1.1 m), suivie d'une décroissance de l'énergie à peu près logarithmique jusqu'au niveau de bruit ambiant.

4.2. Adaptation de l'algorithme à une captation ambisonique réelle

Comme nous l'avons vu dans la partie 1.2.4.3, la validité de l'encodage ambisonique est restreinte à une plage de fréquences dépendant des caractéristiques du microphone, et évoluant en fonction de l'ordre ambisonique synthétisé (figure 1.11). Il est donc nécessaire de prendre en compte les limitations de l'encodage ambisonique pour optimiser à la fois la localisation et la séparation des sources.

4.2.1. Approche choisie pour la localisation de sources

Afin de tirer parti du formalisme ambisonique pour l'analyse de scène, il est nécessaire de restreindre le spectre du contenu analysé à une bande de fréquences dans laquelle la

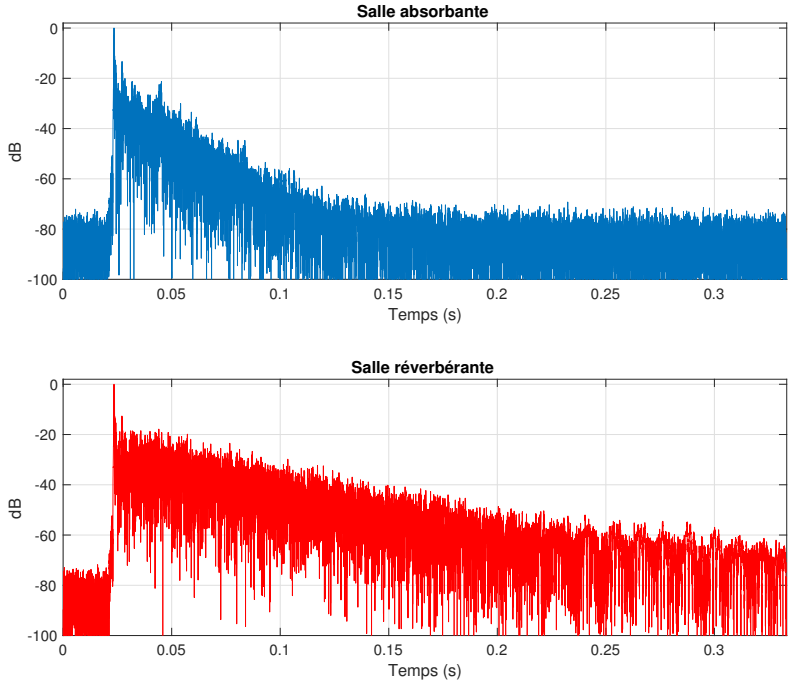

FIGURE 4.2. – Réponses impulsionnelles omnidirectionnelles (en dB normalisés). En haut : salle absorbante. En bas : salle réverbérante.

directivité théorique est respectée par l'ensemble des composantes HOA. Plus l'on monte dans les ordres ambisoniques, plus cette plage est réduite, l'utilisation de contenus HOA réels implique donc d'appliquer un filtrage passe-bande à l'ensemble des canaux, pour ne conserver que la plage de fréquences de l'ordre le plus élevé qui est le plus restrictif [12]. En pratique, si l'on reprend les valeurs issues du tableau 1.2, un contenu ambisonique d'ordre 1 pourra être analysé entre 200 et 8000 Hz. Si l'on souhaite tirer parti des composantes d'ordre 2, il faut alors se restreindre à la bande de fréquences 800-8000 Hz, et ainsi de suite. On suppose bien sûr que les sources sonores à identifier ont un contenu spectral assez riche pour être détectées dans la bande de fréquences restreinte utilisée.
Si l'on analyse un contenu ordre 2 suivant ce procédé, l'ordre 1 valide entre 200 et 800 Hz n'est plus exploité. Au vu des résultats de localisation obtenus sur des contenus synthétiques dans le chapitre 3, il est raisonnable de penser que cette perte d'information va être compensée par l'augmentation du nombre de canaux, qui permet une résolution spatiale plus fine et la localisation d'un nombre plus important de sources simultanées.

4.2.2. Approche choisie pour la séparation de sources

Si l'on ne tient pas compte des dégradations de l'encodage et que l'on choisit une matrice de séparation constante pour toutes les fréquences, la directivité obtenue est altérée dès

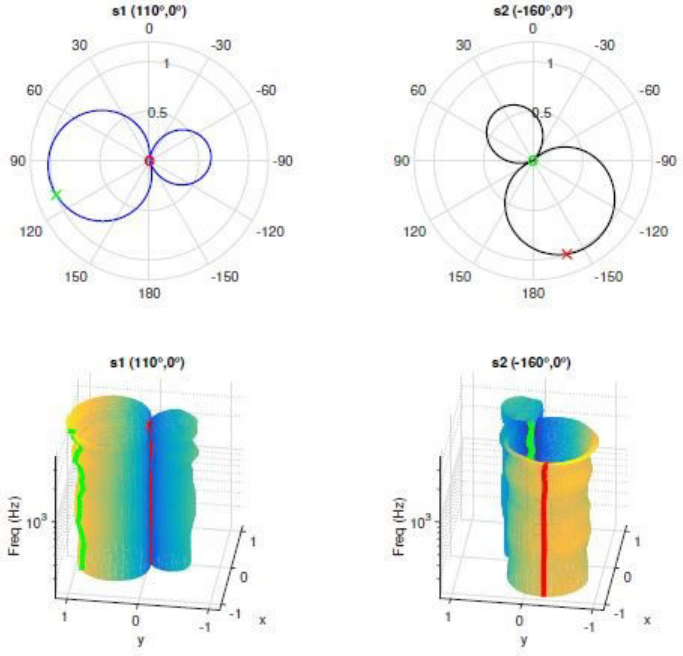

FIGURE 4.3. – Directivités ordre 1 dans le plan horizontal : séparation de deux sources positionnées à 110° et −160° d'azimuth. Matrice de séparation constante. En haut : directivités théoriques. En bas : directivités réelles fonction de la fréquence (200 Hz-4 kHz). Trait vert (resp. rouge) : gain pour la source 1 (resp. 2).

lors que les canaux utilisés ne respectent plus l'encodage théorique. A titre d'exemple, on retrouve sur la figure 4.3 une formation de voies pour séparer deux sources entre 200 Hz et 4 kHz à l'ordre 1 uniquement, et la même formation de voies faisant intervenir l'ordre 2 en gardant une matrice de séparation constante sur tout le spectre (figure 4.4). A l'ordre 1, la formation de voies permet de séparer correctement les deux sources car l'encodage microphonique est valide sur toute la plage de fréquences observée. Pour le contenu d'ordre 2 en revanche, on constate une déformation des figures de directivité au dessous de 800 Hz et donc une dégradation de la séparation obtenue.

Pour mettre à profit la sélectivité des ordres supérieurs tout en conservant une bande-passante la plus large possible, il est judicieux de traiter la séparation spatiale des sources en adaptant la formation de voies en fonction des plages de validité des différents ordres. La solution adoptée est d'utiliser pour le *beamforming* un sous-ensemble de canaux ambisoniques dans chaque bande de fréquences, conduisant dans le cas présent à une formation de voies utilisant exclusivement l'ordre 1 entre 200 Hz et 800 Hz, et incorporant l'ordre 2 pour les fréquences plus élevées. Cette approche permet, comme le montre la figure 4.5 de garantir une séparation de sources cohérente sur l'ensemble du spectre utile, en maintenant un gain unité pour la source cible et un gain nul pour la ou les source(s) interférente(s), cela en contrepartie d'une sélectivité plus faible en basses fréquences.

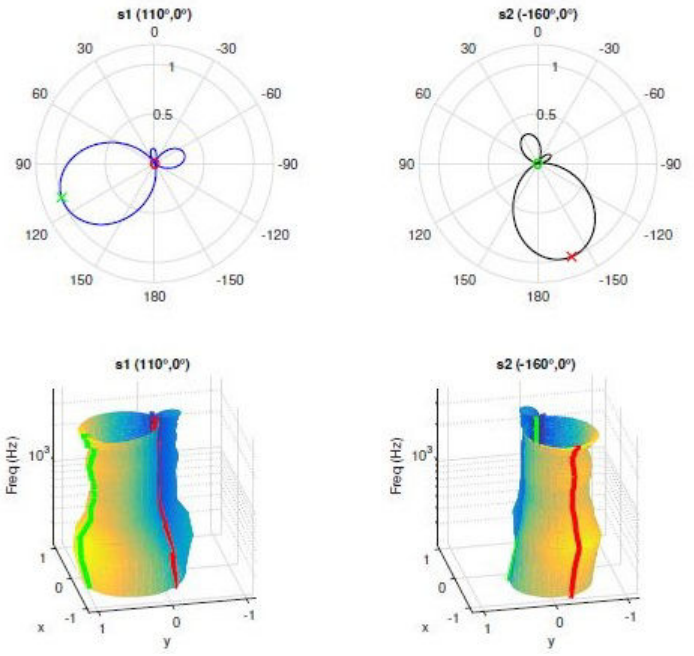

FIGURE 4.4. – Directivités ordre 2 dans le plan horizontal : séparation de deux sources positionnées à 110° et −160° d'azimuth. Matrice de séparation constante. En haut : directivités théoriques. En bas : directivités réelles fonction de la fréquence (200 Hz-4 kHz). Trait vert (resp. rouge) : gain pour la source 1 (resp. 2).

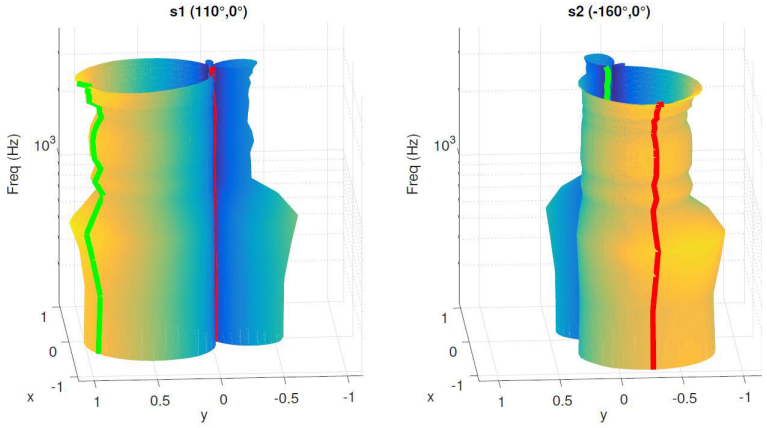

FIGURE 4.5. – Directivités en fonction de la fréquence dans le plan horizontal : séparation de deux sources positionnées à respectivement 110° et −160° d'azimuth. Matrice de séparation générée par sous-bande. Trait vert (resp. rouge) : gain associé à la source 1 (resp. 2).

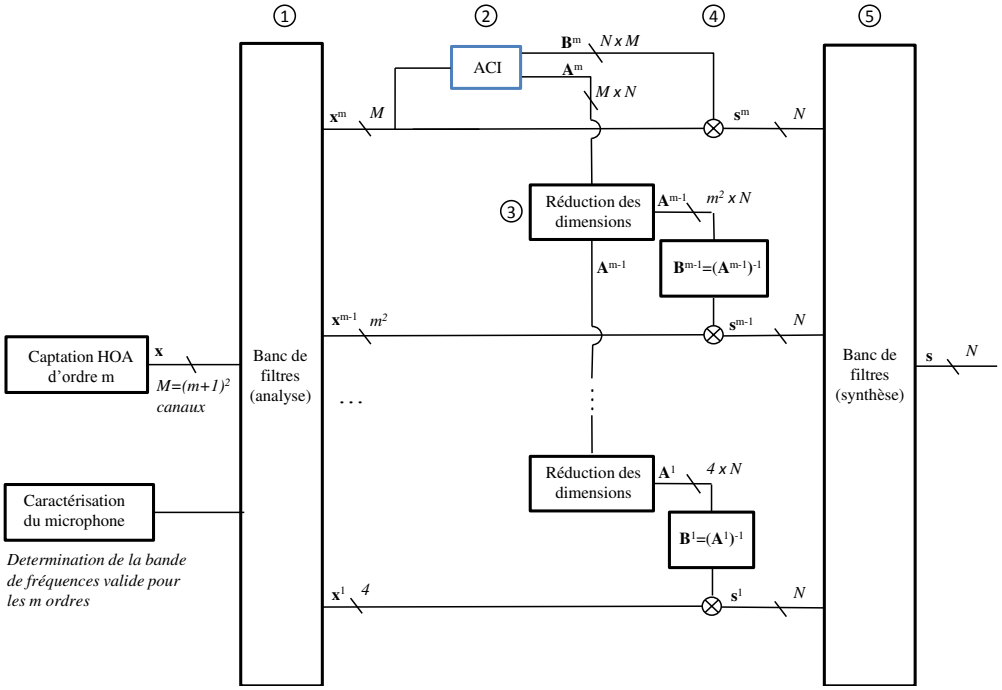

FIGURE 4.6. – Principe de la séparation de sources par sous-bandes dans un contenu ambisonique réel d'ordre m.

4.2.2.1. Algorithme d'analyse/séparation en sous-bandes

Le procédé pour l'analyse et le traitement en sous-bandes des contenus réels est synthétisé par l'algorithme de la figure 4.6.

1. Après une caractérisation de l'encodage, le contenu ambisonique $\mathbf{x}$ d'ordre m comprenant $M = (m+1)^2$ canaux est divisé en sous-bandes $\mathbf{x}^p$ $(1 < p < m)$ à l'aide d'un banc de filtre d'analyse. $\mathbf{x}^p$ est un contenu ambisonique d'ordre p possédant $(p+1)^2$ canaux (étape 1).
2. L'ACI est appliquée dans la sous-bande m possédant le maximum de canaux. Supposant connu le nombre de sources N, on obtient alors la matrice de séparation estimée $\mathbf{B}^m$ de dimensions $N \times M$ et la matrice de mélange estimée $\mathbf{A}^m$, pseudo-inverse de $\mathbf{B}^m$ (étape 2). Dans l'exemple présenté plus haut, on analyse donc un contenu d'ordre 2 dans la sous-bande 800 Hz-8 kHz.
3. Pour chaque sous-bande p, une sous-matrice de mélange $\mathbf{A}^p$ de dimensions $(p+1)^2 \times N$ est calculée à partir de $\mathbf{A}^m$ en tronquant les coefficients des ordres supérieurs à p (étape 3).
4. $\mathbf{B}^p$, pseudo-inverse de $\mathbf{A}^p$ est calculée afin d'extraire les sources $\mathbf{s}^p$, par l'opération $\mathbf{s}^p = \mathbf{B}^p\mathbf{x}^p$ (étape 4).
5. Finalement, le filtre de synthèse permet la reconstruction des sources pleine bande $\mathbf{s}$ à partir des contributions $\mathbf{s}^p$ des m sous-bandes (étape 5).

On note que le nombre de sources séparables sur l'ensemble du spectre est limité à $N = 4$ avec cette approche, soit le nombre de canaux disponibles dans la sous-bande de l'ordre 1.
L'ensemble de ces traitements a fait l'objet d'un brevet déposé durant la thèse. Par ailleurs, ce procédé peut s'appliquer également au décodage de contenus ambisoniques réels sur haut-parleurs, en appliquant de manière analogue un traitement différencié aux différentes sous-bandes afin d'optimiser la restitution.

4.2.3. Visualisation de signaux extraits

Le spectrogramme de la figure 4.7 permet de visualiser le procédé mis en oeuvre. Ici, on visualise la source cible, le mélange et la source extraite par sous-bandes sur l'intervalle 200 Hz-8 kHz à l'ordre 2 pour des signaux de parole.
Utilisant la méthode de reconstruction décrite dans la section 4.2, la source cible est extraite sur la plage de fréquences 200 Hz-8 kHz grâce à l'ordre 1 seul et en ajoutant l'ordre 2 au-delà de 800 Hz. La formation de lobes de directivité beaucoup plus larges en basses fréquences lorsque l'ordre 1 est utilisé seul entraîne la présence d'une plus grande quantité de réverbération et notamment de réverbération provenant de la source interférente, comme on peut le voir aux alentours de 2 s sur le spectrogramme du signal extrait, où la contribution de la source cible apparaît moins distinctement. Au-delà de 800 Hz, la formation de voies d'ordre 2 permet de réhausser efficacement le signal cible et de réduire le taux d'interférences.

4.3. Résultats

On évalue ici les performances de l'outil de localisation/séparation à travers des cas d'études se rapprochant des scénarios réels pouvant être rencontrés. Dans un cas réel, le nombre de sources actives peut être amené à varier, les sources peuvent se déplacer et l'environnement acoustique n'est pas maîtrisé.
Afin de prendre en compte ces trois facteurs, trois scénarios ont été retenus, permettant une évaluation de l'outil en fonction du nombre de sources, de leur localisation et de l'environnement acoustique. Les scènes sont ainsi constituées de :

1. deux sources éloignées,
2. deux sources proches,
3. trois sources.

Les deux premières configurations sont testées à la fois dans un environnement acoustique mat et en conditions plus réverbérantes (salles décrites en annexe B), tandis que la troisième configuration est évaluée uniquement en environnement réverbérant.
L'influence des paramètres d'analyse, à savoir la taille de trame d'analyse et l'ordre ambisonique, est également discutée ici.
Par souci de simplicité, les scores de SIR (équations 3.20 et 3.23) sont calculés uniquement sur la plage de fréquences utilisée pour l'ACI (tableau 1.2), c'est-à-dire avec un

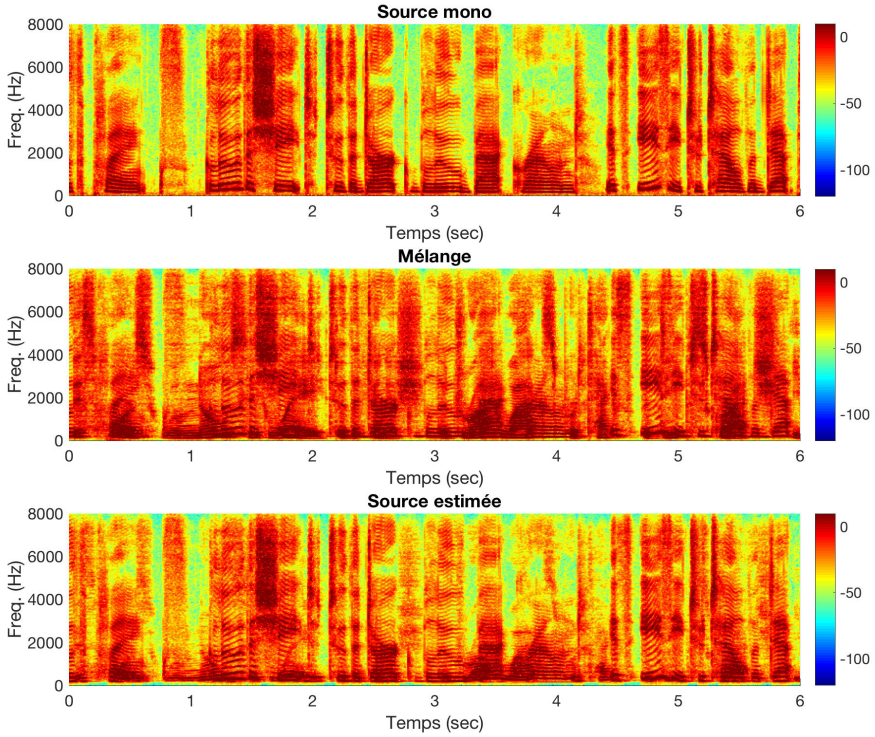

FIGURE 4.7. – Extraction de source en sous-bandes sur un contenu réel d'ordre 2 contenant deux sources (200 Hz-8 kHz). Haut : source cible. Milieu : mélange (canal omnidirectionnel). Bas : source localisée par ACI sur 500 ms et extraite en deux sous-bandes : 200 Hz-800 Hz et 800 Hz-8 kHz.

ordre ambisonique constant pour la séparation. Les SIR avec la reconstruction par sous-bandes ne sont donc pas exposés ici. Par ailleurs, les contributions des sources-images sont normalisées en énergie, ce qui donne un SIR initial de 0 dB en présence d'une seule source interférente, et de -3 dB en présence de deux sources interférentes.

4.3.1. Scénario 1 : deux sources éloignées

Salle mate Le premier cas étudié, le plus favorable, est une scène constituée de deux sources dans le plan azimutal éloignées d'environ 115°, captée dans la salle mate décrite précédemment. Les résultats en terme de localisation (moyenne et écart-type, figure 4.8) font apparaitre la même tendance que pour les contenus synthétiques du chapitre 3, à savoir une diminution de l'erreur de localisation lorsque la taille de trame augmente et lorsque l'on met à profit les ordres ambisoniques supérieurs. Dans ce cas de figure, le biais et la variance de localisation sont inférieurs à 5° au-delà de N=4096 échantillons, quel que soit l'ordre ambisonique utilisé.
Cependant, les scores de séparation (figure 4.9) montrent un intérêt des ordres supérieurs,

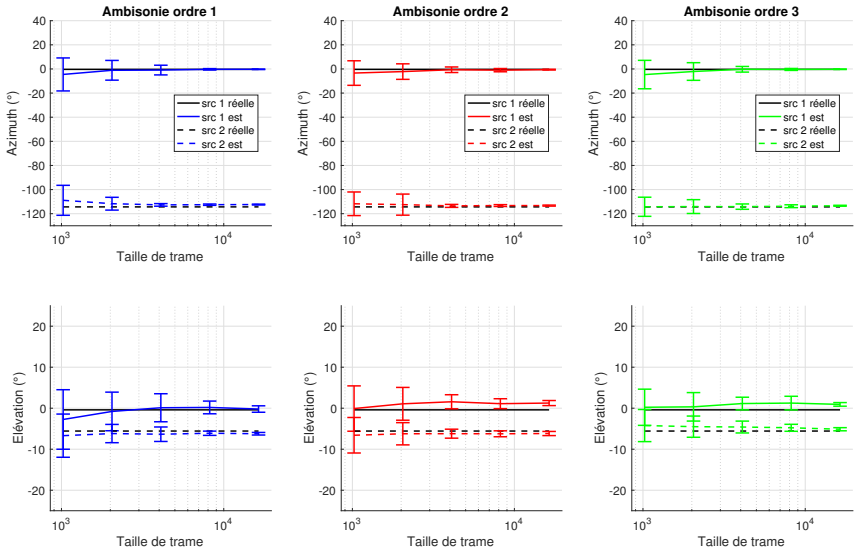

FIGURE 4.8. – Salle mate : localisation (moyenne/écart-type) de deux sources espacées de 120°, en fonction de la taille de trame et de l'ordre ambisonique.

même si l'ensemble des SIR sont supérieurs à 20 dB en ce qui concerne la séparation des champs directs (graphiques du haut), quel que soit l'ordre utilisé ou la taille de trame. Ainsi, alors qu'à l'ordre 1 la valeur médiane stagne aux alentours de 20 dB, on atteint à l'ordre 3 des scores supérieurs à 30 dB pour des trames longues. La différence de 10 dB entre les ordres 1 et 3, alors même que la localisation semble équivalente, est dûe à la formation de voies plus discriminante aux ordres supérieurs. En effet, un *beamforming* d'ordre élevé permet d'annuler le champ acoustique provenant d'une direction identifiée mais également d'atténuer fortement le champ sonore autour de cette direction, pour n'éclairer que la zone autour de la source ciblée. Les sources réelles ne sont pas tout à fait ponctuelles, l'annulation d'une source interférente se fait donc mieux lorsque la formation de voies est plus sélective, pour une localisation équivalente.
Le SIR total (figures du bas), prenant en compte la contribution complète des sources images, est en moyenne inférieur de 7 dB au SIR direct, mais reste au dessus de 13 dB à l'ordre 1 et dépasse les 20 dB à l'ordre 3. Ces valeurs sont à mettre en relation avec la faible contribution énergétique des champs réverbérés, permettant d'obtenir des valeurs de SIR total élevées grâce à l'annulation des champs directs interférents.

Salle réverbérante Le même scénario mais avec des sources captées cette fois en environnement plus réverbérant souligne l'intérêt de trames d'analyse plus longues pour diminuer à la fois le biais et la variance de localisation (figure 4.10). Alors qu'avec 1024 échantillons l'ordre 1 est incapable d'effectuer une localisation précise (environ 30° de biais et d'écart-type en azimuth pour la seconde source), la position estimée tend asymptotiquement vers la position réelle à mesure que le nombre d'échantillons augmente,

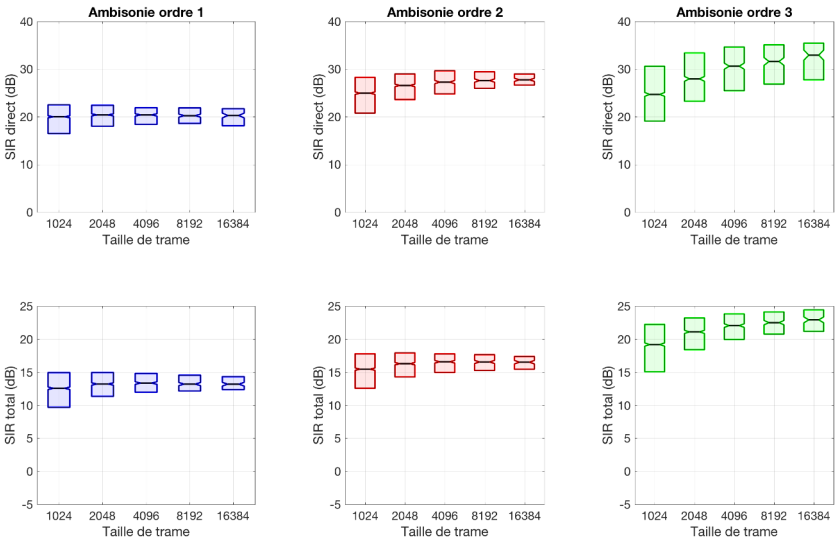

FIGURE 4.9. – Salle mate : SIR des sources extraites (quartiles, dB) pour deux sources espacées de 120°, en fonction de la taille de trame et de l'ordre ambisonique. En haut : SIR des signaux directs. En bas : SIR total.

jusqu'à une localisation précise à 5° près pour 16000 échantillons. On observe la même tendance pour les ordres supérieurs, avec une bien meilleure convergence des estimateurs de localisation malgré une plage de fréquences d'analyse plus restreinte (à l'ordre 3 : biais nul et 10° d'écart-type en azimuth dès 1024 échantillons).
Par ailleurs, cette configuration fait apparaître une différence significative de précision entre les sources 1 et 2 en ce qui concerne la localisation azimutale. La première source, positionnée à 0° d'azimuth est localisée avec un biais quasiment nul à l'ordre 1 même pour des trames de 1024 échantillons, et l'écart-type est deux fois moindre que pour la source 2. Cette différence est probablement due à la géométrie de la pièce (figure 4.1), où une paroi réfléchissante se trouve juste derrière la première source, renvoyant une première réflexion colinéaire à celle-ci.
Les dégradations de localisation par rapport à la salle traitée acoustiquement se ressentent en terme de séparation (figure 4.11). Les résultats soulignent la perte de précision de localisation à l'ordre 1, avec cette fois des SIR directs compris entre 15 dB et 18 dB suivant la taille de trame, et une plus forte variabilité. L'ordre 2, dont la localisation est moins dégradée, affiche des SIR directs compris entre 20 et 30 dB, tandis que l'ordre 3 présente des valeurs identiques à la salle traitée, dépassant les 30 dB à partir de 4096 échantillons. Le SIR total est également impacté, à la fois par la dégradation de la localisation à l'ordre 1, mais également par l'augmentation de la quantité de réverbération, qui fait baisser mécaniquement les scores obtenus même lorsque les sources sont correctement localisées. Le SIR est ainsi limité à respectivement 6 dB, 9 dB et 13 dB pour les ordres 1, 2 et 3 même lorsque la localisation est précise. La sélectivité du *beamforming* est alors

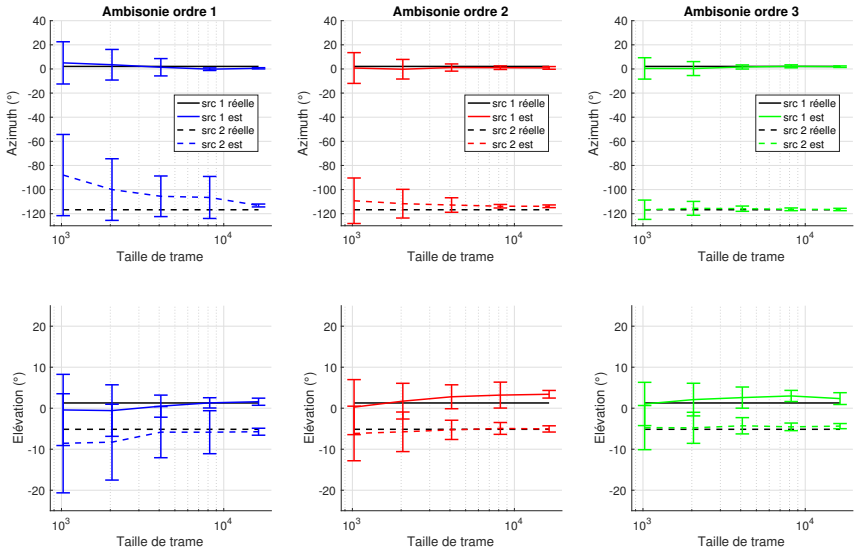

FIGURE 4.10. – Salle réverbérante : localisation (moyenne/écart-type) de deux sources espacées de 120°, en fonction de la taille de trame et de l'ordre ambisonique.

le facteur limitant pour la séparation.

4.3.2. Scénario 2 : deux sources proches

Si les deux sources sont à présent plus rapprochées, à environ 20° de distance angulaire, l'enjeu n'est plus seulement de localiser chaque source avec précision mais également de pouvoir les discriminer l'une de l'autre.

Salle mate Sur la figure 4.12, on aperçoit les limites de l'ordre 1 même en environnement mat. En dessous de 2048 points, les intervalles de confiance des DOA estimées se recouvrent, en azimuth et en élévation, ce qui indique que les contributions des deux sources ne sont pas correctement identifiées sur un grand nombre de trames. Au delà de 4096 points, l'écart-type se réduit et les contributions des deux sources sont nettement identifiées, avec cependant un biais de $-10°$ constant sur la source 2. L'apport des ordres supérieurs apparaît nettement ici, avec les deux sources clairement identifiées à l'ordre 2 dès 1024 points, avec un biais quasiment nul et un écart-type inférieur à 10°.
La séparation des sources (figure 4.13) donne de moins bons résultats qu'avec des sources éloignées dans un environnement identique. A l'ordre 1, Le SIR direct ne dépasse pas 10 dB pour des trames longues (celui-ci était de 20 dB pour des sources éloignées), ce qui est cohérent avec la précision de la localisation, tandis que le SIR total reste inférieur à 5 dB (12 dB pour des sources éloignées). Aux ordres supérieurs, on atteint des valeurs de SIR direct supérieures à 20 dB dès lors que la taille de trame dépasse 4096 points à l'ordre 2 ou 2048 points à l'ordre 3.

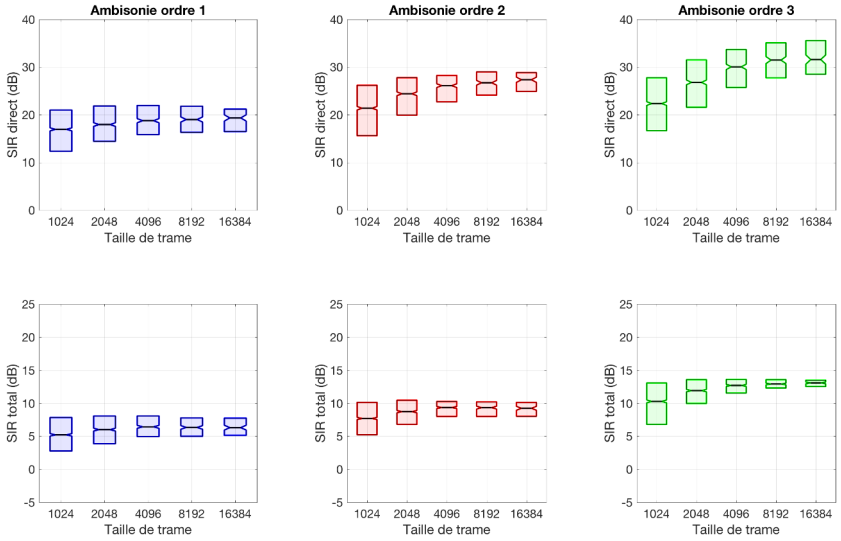

FIGURE 4.11. – Salle réverbérante : SIR des sources extraites (quartiles, dB) pour deux sources espacées de 120°, en fonction de la taille de trame et de l'ordre ambisonique. En haut : SIR des signaux directs. En bas : SIR total.

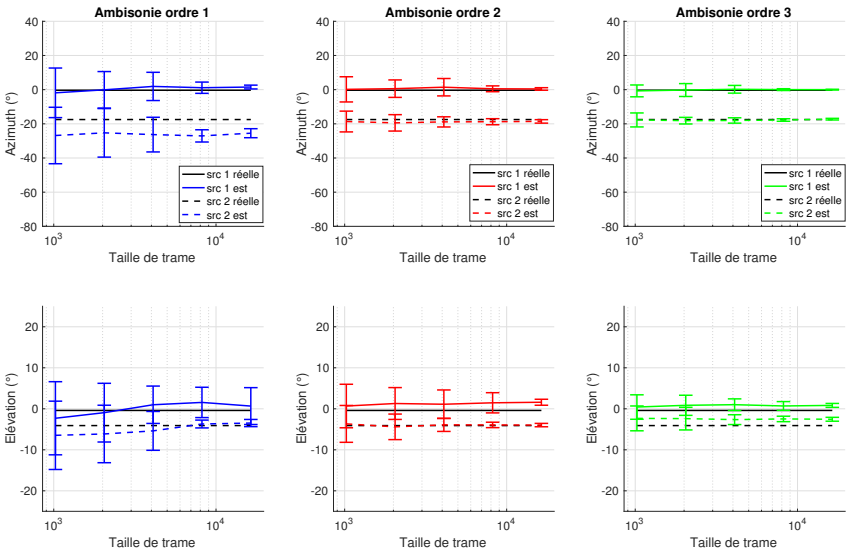

FIGURE 4.12. – Salle mate : localisation (moyenne/écart-type) de deux sources espacées de 20°, en fonction de la taille de trame et de l'ordre ambisonique.

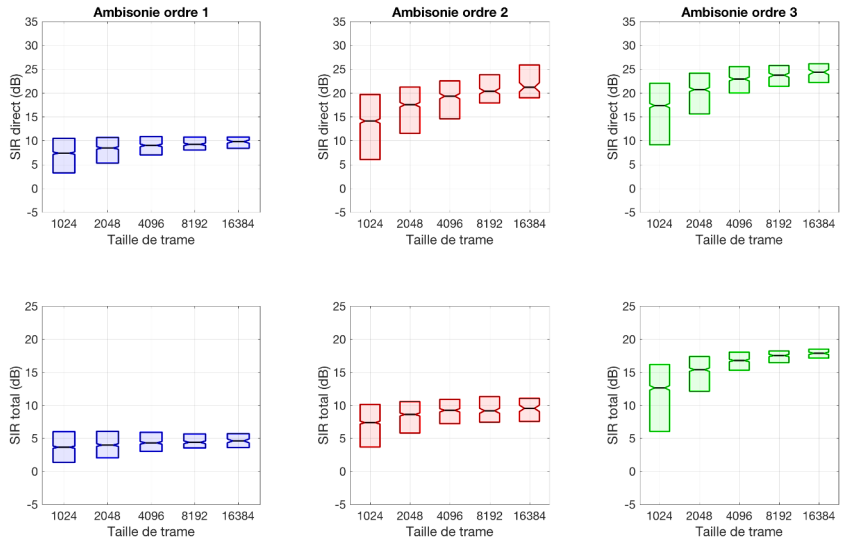

FIGURE 4.13. – Salle mate : SIR des sources extraites (quartiles, dB) pour deux sources espacées de 20°, en fonction de la taille de trame et de l'ordre ambisonique. En haut : SIR des signaux directs. En bas : SIR total.

Salle réverbérante En environnement réverbérant (figure 4.14), l'ordre 1 est clairement insuffisant pour séparer deux sources distantes de 20° seulement, et ce quelle que soit la longueur de trame utilisée. L'ordre 2 à partir de 4096 points remplit cette mission correctement, avec néanmoins un biais de respectivement 5° et −10° sur les sources 1 et 2. Enfin, l'ordre 3 ne semble pas affecté par la proximité des sources dès 2048 points, avec un biais quasi-nul et un écart-type inférieur à 5°.
La séparation reflète les différences de localisation entre les différents ordres ambisoniques (figure 4.15), avec un SIR direct stagnant entre 0 et 5 dB à l'ordre 1, des valeurs médianes qui dépassent difficilement 10 dB à l'ordre 2 à partir de 4096 points, mais des valeurs supérieures à 20 dB à l'ordre 3 pour des trames de 4096 points. Le SIR total est également quasi-nul à l'ordre 1 et plafonne à respectivement 5 dB et 10 dB aux ordres 2 et 3, soit des valeurs inférieures d'environ 5 dB à celles obtenues lorsque les sources étaient plus espacées.
On a constaté que la discrimination des sources se faisait souvent au prix d'un biais de localisation, en tout cas aux ordres 1 et 2. Ce biais, présent sur la localisation d'une seule ou des deux sources, tend généralement à augmenter la distance entre les sources estimées. Ce phénomène s'explique par le fait que plus les sources sont proches, plus la matrice de mélange ambisonique est mal conditionnée. L'inversion de celle-ci pour la séparation entraîne alors la formation de directivités avec des lobes principaux et secondaires de forte amplitude. Le gain restant unitaire dans la direction de la source ciblée, l'amplification des lobes entraîne une baisse du facteur de directivité et donc une augmentation du taux de réverbération captée (voir la section 2.4.3 consacrée à la formation de voies pour la

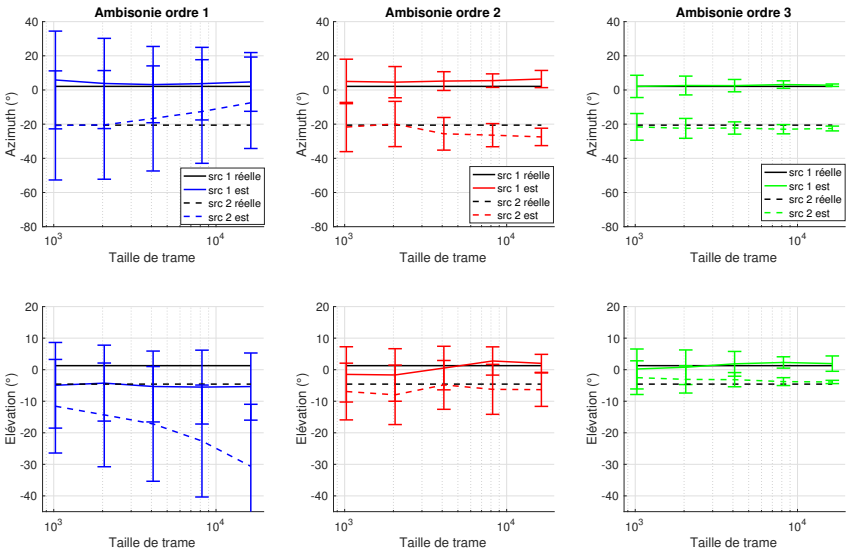

FIGURE 4.14. – Salle réverbérante : localisation (moyenne/écart-type) de deux sources espacées de 20°, en fonction de la taille de trame et de l'ordre ambisonique.

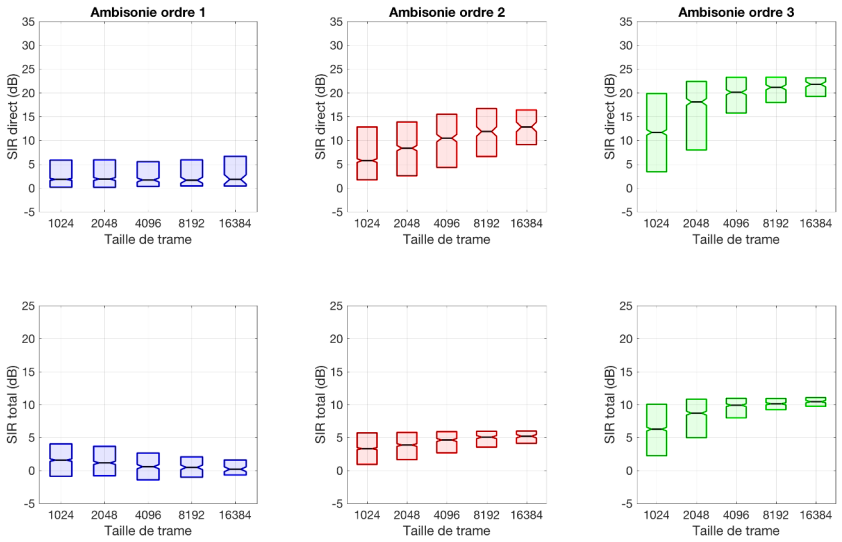

FIGURE 4.15. – Salle réverbérante : SIR des sources extraites (quartiles, dB) pour deux sources espacées de 20°, en fonction de la taille de trame et de l'ordre ambisonique. En haut : SIR des signaux directs. En bas : SIR total.

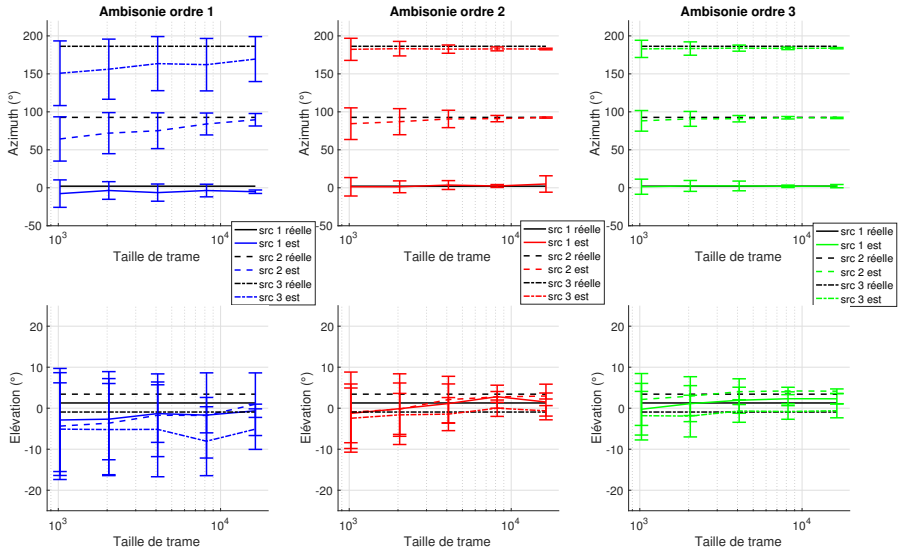

FIGURE 4.16. – Salle réverbérante : localisation (moyenne/écart-type) de trois sources, en fonction de la taille de trame et de l'ordre ambisonique.

séparation de sources, et plus particulièrement de la figure 2.4). Ainsi, on peut interpréter le biais de la localisation par ACI comme un compromis entre une bonne séparation des champs directs et un faible taux de réverbération résiduelle. Ce compromis, difficilement quantifiable, va dépendre en pratique de la distance entre les sources, du niveau de l'effet de salle et de l'ordre ambisonique pour la formation de voies.

4.3.3. Scénario 3 : trois sources simultanées

L'influence d'une source supplémentaire est étudiée en configuration réverbérante, avec les trois sources placées dans le plan azimutal à respectivement 0°, 90° et 180°.
Les performances de localisation, visibles sur la figure 4.16, mettent en avant la dégradation de la localisation à l'ordre 1, par rapport au cas de deux sources éloignées captées dans les mêmes conditions. La source 3 n'est pas correctement localisée même pour des trames longues, et il subsiste un biais de $-20°$ en azimuth pour N=16384 échantillons ainsi qu'une variance importante. *A contrario*, l'ordre 2 et l'ordre 3 permettent de localiser sans biais les trois sources. Pour des trames de 2048 points, on obtient un écart-type sur la localisation azimutal compris entre 7 et 17° pour les trois sources à l'ordre 2, tandis qu'à l'ordre 3 l'écart-type passe sous les 10° pour l'ensemble des sources.
Les imprécisions de localisation et les contraintes importantes sur la formation de voies entraîne à l'ordre 1 un SIR direct compris entre 5 dB et 10 dB, tandis que le SIR total ne dépasse pas les 3 dB. En revanche, les SIR directs aux ordres supérieurs dépassent les 15 dB et atteignent 25 dB pour les trames les plus longues. Aux ordres 2 et 3, la localisation précise et le nombre plus important de degrés de liberté permet d'obtenir un

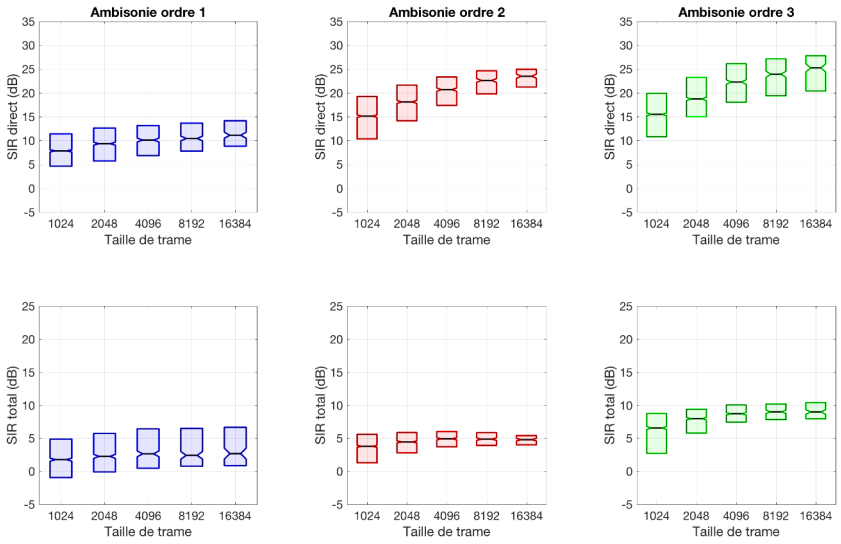

FIGURE 4.17. – Salle mate : SIR des sources extraites (quartiles, dB) pour trois sources simultanées, en fonction de la taille de trame et de l'ordre ambisonique. En haut : SIR des signaux directs. En bas : SIR total.

SIR total d'environ 5 dB à l'ordre 2 et de 8 à 10 dB à l'ordre 3.
Ce cas de figure montre les limites de l'ordre 1 à la fois pour la localisation et la séparation spatiale de trois sources en milieu réverbérant, à cause d'un nombre limité de canaux et une faible capacité de focalisation. En revanche, les ordres 2 et 3 sont peu affectés par l'augmentation du nombre de sources et permettent une localisation correcte et une séparation efficace des champs directs malgré une variabilité un peu plus importante des estimateurs.

4.4. Conclusion

Les performances de l'algorithme de localisation et séparation de sources basé sur ERBM ont ici été évalués sur des contenus ambisoniques réels. Les résultats obtenus en termes de localisation et de séparation de sources ont montré la pertinence de l'approche développée pour des contenus constitués de deux ou trois sources, en particulier lorsqu'il est fait appel aux ordres ambisoniques supérieurs. L'utilisation d'un contenu d'ordre 1 semble suffire dans des configurations favorables, à savoir des sources sonores peu nombreuses et suffisamment espacées. L'ajout des composantes d'ordres 2 et 3 augmente grandement la robustesse à la réverbération d'une part et permet d'autre part de discriminer des sources spatialement proches (20°). De plus, la séparation d'un nombre de sources supérieur à 3 grâce aux ordres supérieurs, qui n'est pas abordée ici, semble envisageable au vu de premiers résultats informels obtenus.

5. Comptage de sources dans un contenu réel

Dans les chapitres précédents, nous nous sommes attachés à évaluer la capacité de l'ACI à localiser et séparer des sources dans des contenus ambisoniques synthétiques ou réels. Néanmoins, la problématique du dénombrement et de l'identification des sources actives, essentielle à l'analyse de scène, n'a pas été abordée jusqu'à présent et fait donc l'objet de ce cinquième chapitre.
On dispose d'un contenu ambisonique réel $\mathbf{x}$ d'ordre m, possédant $M = (m+1)^2$ canaux, dans lequel un nombre inconnu de N sources sonores sont actives. Dans le cas d'un mélange instantané, le nombre de sources est directement déduit du rang de la matrice de covariance des observations. En revanche, dans le cas d'un mélange réel ou simplement réverbérant, la matrice de covariance est généralement de rang plein du fait de la contribution de l'effet de salle et ne permet pas d'en déduire directement le nombre de sources actives.
Ainsi, lorsque l'ACI est appliquée à un contenu ambisonique réel, on extrait un nombre de composantes s_j égal au nombre M d'observations. On rappelle ici que le vecteur des composantes indépendantes $\mathbf{s} = [s_1, \cdots, s_M]$ est obtenu par l'opération :

$$\mathbf{s} = \mathbf{B}\mathbf{x} \tag{5.1}$$

où $\mathbf{B}$ est la matrice de séparation estimée. Si les directions d'arrivée des sources sont correctement estimées, il en résultera N composantes correspondant aux champs directs des sources présentes (appelées par la suite composantes directes) ainsi que $M - N$ composantes restantes résultant uniquement de l'effet de salle (appelées par la suite composantes réverbérées).
Ces composantes supplémentaires posent plusieurs problèmes :

- Pour l'analyse de scène : on ne sait pas *a priori* quelles sont les composantes relatives aux sources et les composantes induites par l'effet de salle.
- Pour la séparation des sources par *beamforming* : chaque composante supplémentaire induit des contraintes sur les directivités formées et dégrade généralement le facteur de directivité (voir section 2.4.3) avec pour conséquence un réhaussement du niveau de réverbération dans les signaux extraits.

Plusieurs approches de comptage de sources pour des contenus multicanal ont été proposées dans la littérature, basées souvent sur une hypothèse de parcimonie dans le domaine temps-fréquence, avec un clustering des DOA pointées par chaque bin fréquentiel [86], une modélisation de la répartition des DOA par un mélange de gaussiennes [87], ou encore une approche bayésienne variationnelle [88].

Nous avons choisi de traiter cette problématique en mettant à profit les informations issues de l'ACI, à savoir les composantes extraites et les matrices de séparation/mélange estimées. Au vu de ce qui vient d'être dit, on peut naturellement envisager le problème de dénombrement et d'identification des sources comme un problème de classification supervisée des composantes extraites par ACI en deux classes : composantes directes (classe $\mathcal{C}^d$) et composantes réverbérées (classe $\mathcal{C}^r$).
Pour ce faire, un jeu de descripteurs d_k des composantes s_j est sélectionné, basés à la fois sur les signaux extraits et sur les coefficients de séparation/mélange estimés. Ces descripteurs, dont la distribution $p(d_k|\mathcal{C}^\alpha)$, $\alpha \in \{d, r\}$ est estimée pour chaque classe par une étape d'apprentissage, vont alors être utilisés au sein d'un classifieur naïf bayésien.
Ce chapitre détaille dans un premier temps les descripteurs choisis ainsi que les densités de probabilité associées, puis le formalisme du classifieur bayésien mis en œuvre. Dans un troisième temps, les performances de l'étape de classification sur des contenus ambisoniques réels sont discutés suivant différents paramètres comme le nombre de sources ou l'énergie relative des différentes sources concurrentes.

5.1. Descripteurs utilisés

Le choix des descripteurs est justifié par un ensemble d'hypothèses, liées à la fois à la physique du problème et au point de départ que l'on s'est fixé, à savoir un jeu de composantes extraites par ACI. Celles-ci sont les suivantes :

— Chaque composante dite "directe" est principalement constituée du champ direct d'une source, assimilable à une onde plane, auquel s'ajoute une réverbération résiduelle dont la contribution énergétique est néanmoins plus faible.

— Les sources sont indépendantes, il y a donc une faible corrélation entre les composantes directes extraites.

— Chaque composante réverbérée est constituée de premières réflexions, versions retardées et filtrées du ou des champs directs, et d'une réverbération tardive. Ainsi, les composantes réverbérées présentent une corrélation avec les composantes directes, et généralement un retard de groupe identifiable par rapport aux composantes directes.

5.1.1. Critère onde plane

D'après la première hypothèse énoncée, une composante directe est principalement composé d'une onde plane. On rappelle que l'encodage à l'ordre 1 d'une onde plane s_j d'incidence (θ_j, ϕ_j) au format ambisonique s'effectue suivant la formule :

$$\mathbf{x}_j = \mathbf{A}_j s_j \tag{5.2}$$

où

$$\mathbf{A}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ a_{3j} \\ a_{4j} \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{3}\cos\theta_j \cos\phi_j \\ \sqrt{3}\sin\theta_j \cos\phi_j \\ \sqrt{3}\sin\phi_j \end{bmatrix} \tag{5.3}$$

dépend de la direction d'arrivée de la source. D'après l'équation 5.3, on déduit une relation, que l'on nomme critère onde plane, liant les coefficients d'ordre 0 et d'ordre 1 :

$$c_{\mathrm{pw}}(0,1) = \sqrt{\frac{3a_{1j}^2}{a_{2j}^2 + a_{3j}^2 + a_{4j}^2}} \tag{5.4}$$

Cette équation, qui traduit une relation énergétique entre les contributions des différents ordres ambisoniques, peut être étendue aux ordres m et m', donnant la formulation générale du critère onde plane :

$$c_{\mathrm{pw}}(m,m') = \sqrt{\frac{(2m'+1)\cdot\sum_{n\sigma}(Y^{\sigma}_{mn,j})^2}{(2m+1)\cdot\sum_{n\sigma}(Y^{\sigma}_{m'n,j})^2}} \tag{5.5}$$

où $Y^{\sigma}_{mn,j}$ est la fonction harmonique sphérique qui dépend des coordonnées (θ_j, ϕ_j). On a donc dans le cas de l'encodage idéal d'une onde plane, l'égalité :

$$c_{\mathrm{pw}}(m,m') = 1, \quad \forall(m,m') \tag{5.6}$$

La matrice de séparation $\mathbf{B}$ obtenue par ACI permet de calculer une matrice de mélange estimée $\mathbf{A}$ par l'opération $\mathbf{A} = \mathbf{B}^{-1}$. On peut donc, pour chaque composante extraite, calculer le critère onde plane grâce aux coefficients de mélange estimés et évaluer ainsi la conformance avec un encodage onde plane. En présence d'un champ direct correctement identifié, on suppose que le critère onde plane restera très proche de la valeur 1. A l'inverse, dans le cas d'une composante réverbérée, la multitude des contributions (premières réflexions et réverbération tardive) avec des niveaux énergétiques équivalents vont généralement éloigner le critère onde plane de sa valeur idéale.
Cette remarque peut être nuancée par la capacité d'ERBM à identifier dans des conditions favorables une ou plusieurs premières réflexions, qui vont pouvoir présenter des caractéristiques d'ondes planes similaires. Néanmoins, la discrimination onde plane/champ diffus est un premier critère qui semble pertinent pour déterminer si une composante extraite correspond ou non à une source réelle. Ce critère peut être rapproché de celui de l'algorithme diRAC présenté en 2.3, qui calcule un facteur de diffusivité en fonction du ratio énergétique entre les composantes ambisoniques.
Le critère onde plane à l'ordre 1 est le premier descripteur utilisé. Pour ce descripteur comme pour les suivants, la distribution associée connaît une certaine variabilité, en fonction notamment du niveau de bruit présent dans les composantes extraites. Ce bruit est constitué principalement de la réverbération résiduelle et des contributions des sources interférentes qui n'auront pas été parfaitement annulées. Le choix a donc été fait, pour affiner l'analyse, d'estimer la distribution des descripteurs en fonction :

— de l'ordre ambisonique utilisé, qui influe sur la sélectivité du *beamforming* et donc sur le niveau de bruit résiduel,

— du nombre de sources contenues dans le mélange, dont l'augmentation entraine mécaniquement une hausse du niveau de bruit. Ce nombre de sources supposé est pris en compte dans le classifieur présenté dans la section suivante, au moment d'évaluer les classes de l'ensemble des composantes extraites.

La modélisation des densités de probabilité des descripteurs fait l'objet de l'annexe C. On peut observer sur la figure 5.1 les lois de probabilités associées à ce descripteur, en fonction du nombre de sources actives simultanément (1 ou 2) et de l'ordre ambisonique du contenu analysé (ordres 1 à 2). Conformément à l'hypothèse initiale, la valeur du critère onde plane est concentrée autour de la valeur 1 pour les composantes directes. Pour les composantes réverbérées, la distribution est plus uniforme, avec cependant une forme légèrement asymétrique, à cause du descripteur lui-même qui est asymétrique, avec une forme en $\frac{1}{x}$.
Par ailleurs, on observe pour les composantes directes une répartition du descripteur plus resserrée dans le cas d'un contenu ordre 2, conséquence d'une facteur de directivité plus élevé et donc d'une contribution plus importante du champ direct par rapport au champ réverbéré résiduel.

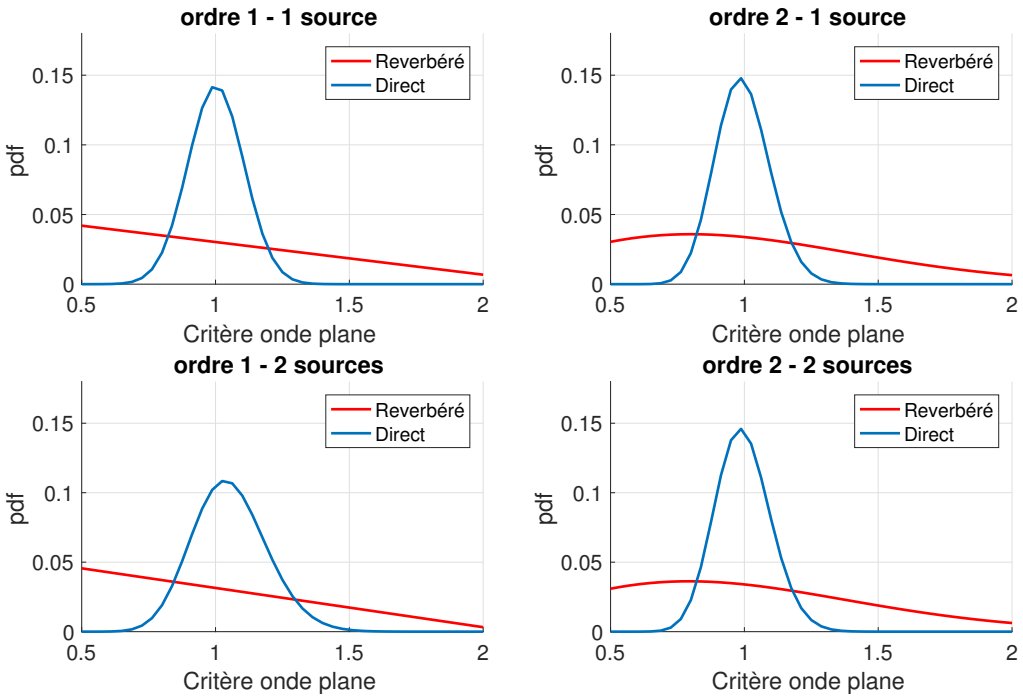

FIGURE 5.1. – Distribution du critère onde plane (équation 5.4) en fonction de la classe de la composante (directe ou réverbérée), de l'ordre ambisonique (gauche/droite) et du nombre de sources (haut/bas).

Au vu de cette répartition et pour simplifier la classification, le descripteur est uniquement pris en compte sur l'intervalle [0.5 2]. En dehors de cet intervalle de valeurs, les composantes sont considérées automatiquement comme réverbérées.
La distance entre les distributions des deux classes permet une discrimination assez fiable entre les composantes de type ondes planes et celles plus diffuses. Les descripteurs suivants vont permettre d'affiner l'identification des champs directs et d'identifier, parmi les composantes ondes planes, celles liées aux premiers fronts d'ondes et celles issues de premières réflexions.

5.1.2. Cohérence moyenne

Le second descripteur permet de conforter l'identification des ondes planes et de fournir une indication sur la présence de premières réflexions parmi celles-ci.
La fonction de cohérence γ^2_{jl} renseigne sur l'existence d'une corrélation entre deux signaux s_j et s_l et s'exprime suivant la formule :

$$\gamma^2_{jl}(f) = \frac{|S_{jl}(f)|^2}{S_{jj}(f)S_{ll}(f)} \tag{5.7}$$

où S_{jl} est l'interspectre entre s_j et s_l et S_{jj} et S_{ll} sont les autospectres respectifs de s_j et s_l.
la cohérence est idéalement nulle lorsque s_j et s_l sont les champs directs de sources indépendantes mais elle prend une valeur élevée lorsque s_j et s_l sont deux contributions d'une même source : le champ direct et une première réflexion ou bien deux réflexions.
Le descripteur utilisé est la moyenne sur l'ensemble des fréquences de la fonction de cohérence entre deux composantes, soit :

$$d^{\gamma}(s_j, s_l) = \mathrm{E}_f\{\gamma^2_{jl}(f)\} \tag{5.8}$$

La cohérence étant bornée entre 0 et 1, la cohérence moyenne sera également compris dans cet intervalle, tendant vers 0 pour des signaux parfaitement indépendants et vers 1 pour des signaux fortement corrélés. Ce descripteur bivarié est obtenu en pratique en calculant les estimateurs spectraux sur des trames de quelques dizaines de ms, uniquement sur la plage de fréquences utilisée pour l'analyse ambisonique, c'est-à-dire entre 200 Hz et 8 kHz à l'ordre 1 et au dessus de 800 Hz pour un contenu d'ordre 2. S'agissant d'un descripteur dépendant de deux composantes, trois cas de figures doivent être modélisés en fonction des classes respectives des composantes :

1. Les deux composantes sont des champs directs,
2. L'une des deux composantes est directe et l'autre est réverbérée,
3. Les deux composantes sont réverbérées.

La figure 5.2 donne un aperçu des valeurs de cohérence obtenues pour les cas 1 (deux signaux directs) et 2 (couple direct/réverbéré). Dans le premier cas, d^{γ} est inférieur à 0.3 alors que dans le second cas d^{γ} atteint 0.7 en présence d'une seule source active. Ces valeurs reflètent bien à la fois l'indépendance des signaux directs et la relation liant

un signal direct et le même signal réverbéré, en l'absence d'interférences. Cependant, en incorporant une seconde source active dans le mélange initial, la cohérence moyenne du cas direct/reverbéré descend à 0.55 et se retrouve fortement dépendante du contenu spectral et du niveau énergétique des différentes sources. Ici, la concurrence des différentes sources fait chuter la cohérence en basses fréquences, tandis que les valeurs sont plus élevées au-dessus de 5500 Hz en raison d'une plus faible contribution de la source interférente. Cet exemple met en avant la nécessité de prendre en compte le nombre de sources actives pour déterminer la distribution de la cohérence moyenne.

Les lois de probabilité associées de la cohérence moyenne sont présentées en fonction de

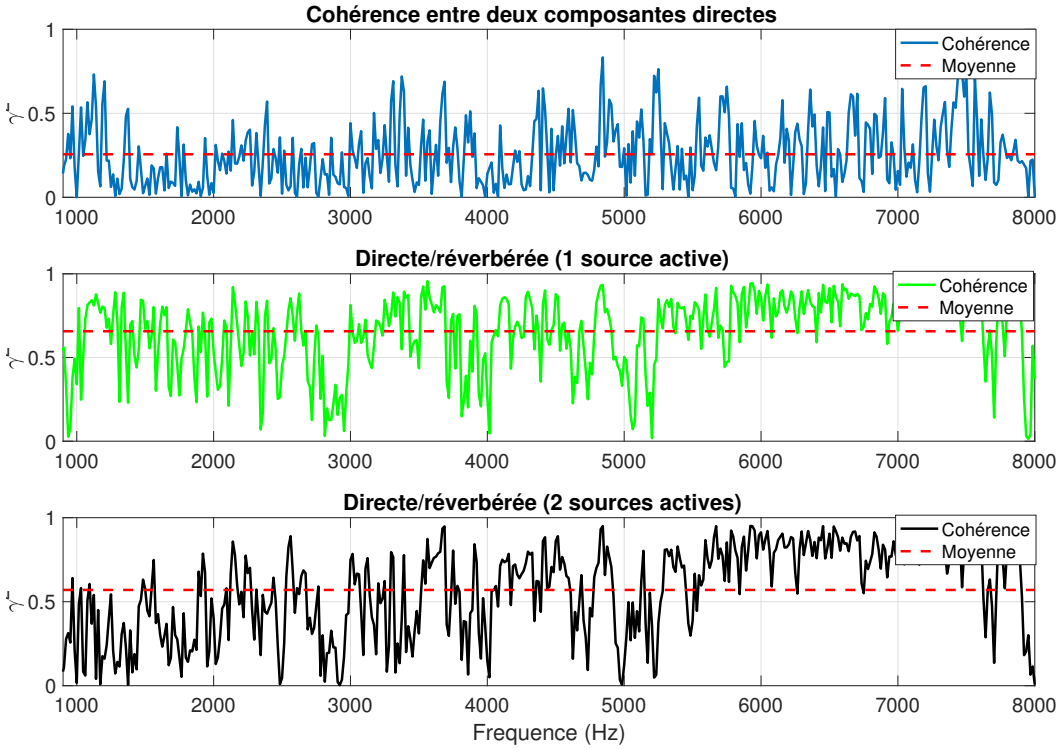

FIGURE 5.2. – Fonction de cohérence entre composantes extraites. En haut : deux composantes directes. Au centre : composantes directe et réverbérée avec une seule source active dans le mélange. En bas : composantes directe et réverbérée en présence de deux sources actives.

l'ordre ambisonique sur la figure 5.3, dans le cas d'une ou deux sources actives simultanément. On observe tout d'abord que d^{γ} prend des valeurs nettement plus faibles pour des couples de composantes directes par rapport aux cas où au moins une des composantes est réverbérée, et cette observation est d'autant plus marquée que l'ordre ambisonique est élevé. Cela paraît logique étant donné la sélectivité plus importante de la formation de voies.

On constate également qu'en présence de deux sources, les estimateurs de cohérence se dégradent, que ce soient les couples direct/réverbéré ou réverbéré/réverbéré (en présence d'une seule source, le couple direct/direct n'existe pas).
En définitive, il apparaît que les densités de probabilité dépendent fortement du nombre de sources dans le mélange, ce qui pourra être mis à profit par la suite. En revanche, identifier les composantes directes uniquement au vu des distributions de la figure 5.3 semble compliqué : même en présence d'une seule source, les couples de composantes réverbérées possède une distribution plus étalée mais finalement très proches des couples direct/réverbéré.

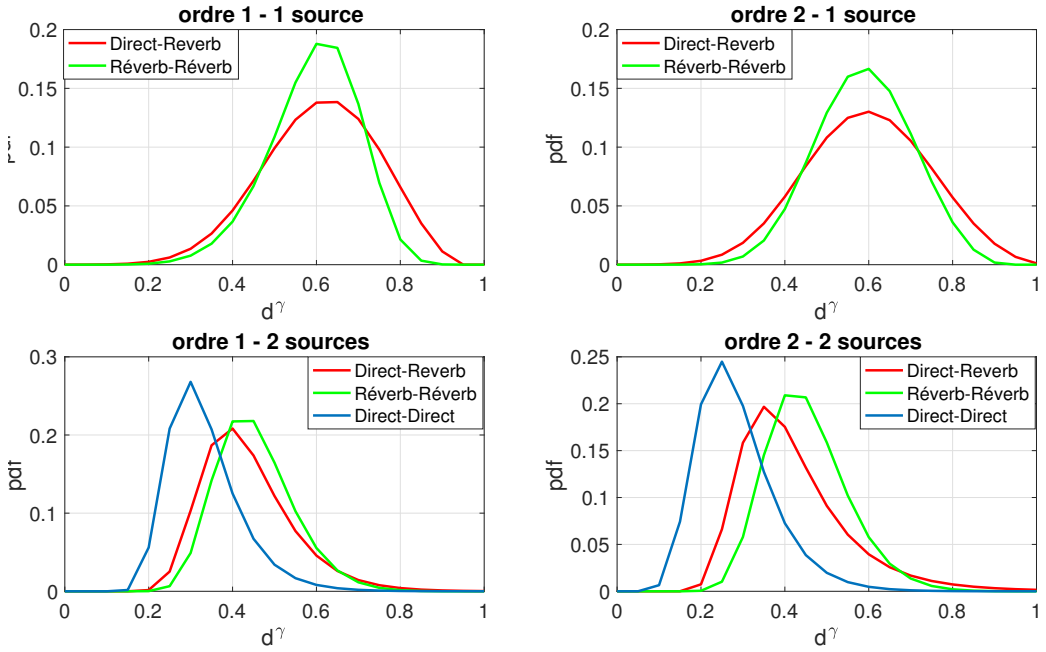

FIGURE 5.3. – Distribution de la cohérence moyenne d^{γ} entre deux composantes, en fonction de leur classe, de l'ordre ambisonique (gauche/droite) et du nombre de sources (haut/bas).

5.1.3. Délai de groupe

La cohérence moyenne permet de déterminer le degré de corrélation entre les couples de composantes, pour déterminer si celles-ci sont indépendantes (champs directs) ou plus corrélées, suggérant alors une relation signal direct/réflexion. Le dernier descripteur va permettre de déterminer dans ce cas de figure quelle composante est plus probablement le signal direct et laquelle correspond au signal réverbéré, en se basant sur l'hypothèse simple que les premières réflexions sont des versions retardées et atténuées du signal direct.
On définit le retard $\tau_{\max}$ comme le retard qui maximise la fonction d'intercorrélation

entre les signaux s_j et s_l :

$$\tau_{max} = \arg\max_{\tau} \mathrm{E}\{s_j(t)s_l(t-\tau)\} \tag{5.9}$$

Lorsque s_j est un signal direct et s_l une réflexion associée, le tracé de la fonction d'intercorrélation fera généralement apparaître un $\tau_{\max}$ négatif. On peut ainsi espérer discriminer les composantes directe et réverbérée grâce au signe de $\tau_{\max}$.
En pratique, l'estimation du signe de $\tau_{\max}$ est souvent très bruitée, voire même parfois inversée :

— Lorsque la scène est constituée d'une seule source, il n'y a pas forcément de délai de groupe qui émerge distinctement si le champ réverbéré est composé de multiples réflexions et de réverbération tardive. De plus les composantes directes extraites par ACI contiennent toujours un résidu d'effet de salle plus ou moins important, qui va bruiter la mesure du délai.

— Lorsque plusieurs sources sont présentes, les interférences viennent perturber la mesure, à plus forte raison si les trames d'analyse sont courtes et que tous les champs directs interférents n'ont pas été parfaitement localisés et séparés.

Pour ces raisons, le signe de $\tau_{\max}$ est utilisé comme descripteur, mais couplé avec une mesure de fiabilité. La cohérence moyenne entre les composantes permet d'évaluer la pertinence du couple direct/réverbéré. Si celle-ci est forte, on peut espérer que le délai de groupe sera un descripteur fiable. D'autre part, la valeur du pic d'intercorrélation $\tau_{\max}$ renseigne également sur la fiabilité du délai de groupe. On utilise donc comme second indicateur de fiabilité le rapport entre la valeur de l'intercorrélation à $\tau_{\max}$ et le maximum de corrélation de signe opposé. Ce ratio, que l'on nomme émergence, est un critère *ad hoc* dont la pertinence se vérifie en pratique.
La figure 5.4 illustre l'émergence pour une composante directe et une composante réverbérée, en présence d'une ou deux sources actives dans le mélange. On voit clairement émerger le maximum de corrélation dans le premier cas, alors que la valeur relative du maximum de corrélation est moindre en présence d'une source interférente.
On a donc un descripteur d^{τ} qui détermine, pour chaque couple supposé direct/réverbéré, la probabilité pour chaque composante du couple d'être la composante directe ou la composante réverbérée. Ce descripteur est fonction du signe de $\tau_{\max}$, de la cohérence moyenne entre les composantes et de l'émergence du maximum d'intercorrélation.
Pour s_j identifiée comme étant en avance sur s_l, on estime la probabilité que s_j soit directe et s_l réverbérée par une loi à deux dimensions, dont on visualise sur la figure 5.5 la modélisation par un polynôme d'ordre 5. Une modélisation idéale donnerait une fiabilité égale à 1 lorsque la cohérence et l'émergence sont maximales, et à l'opposé une fiabilité de 0.5 (une chance sur deux) lorsque la cohérence et l'émergence sont nulles, car il n'y a alors aucune relation entre les deux composantes analysées. Malgré la précision limitée de l'approximation polynomiale, notamment aux limites du domaine affiché, il apparaît nettement que le signe du retard est un indicateur fiable lorsqu'à la fois la cohérence et l'émergence ont des valeurs moyennes ou élevées. Une émergence faible ou une cohérence faible vont rendre les couples direct/réverbéré ou réverbéré/direct équiprobables.

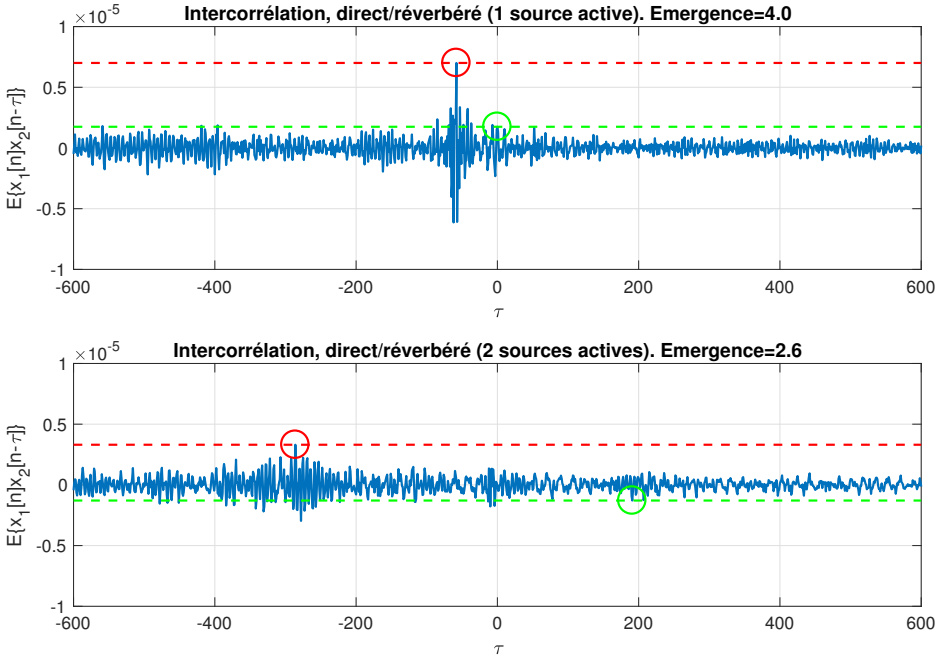

FIGURE 5.4. – Fonction d'intercorrélation entre deux composantes directe et réverbérée, en présence d'une seule source active (en haut) et de deux sources (en bas). En rouge : maximum d'intercorrélation. En vert : maximum de signe opposé servant au calcul de l'émergence.

Logiquement, on estime alors la probabilité que s_j soit réverbérée et s_l directe alors même que s_j est en avance de phase comme le complément à 1 du cas direct/réverbéré :

$$p(\mathcal{C}_j = \mathcal{C}^r, \mathcal{C}_l = \mathcal{C}^d | d^\tau) = 1 - p(\mathcal{C}_j = \mathcal{C}^d, \mathcal{C}_l = \mathcal{C}^r | d^\tau) \tag{5.10}$$

où $\mathcal{C}_j$ et $\mathcal{C}_l$ sont les classes respectives des composantes s_j et s_l. Ce descripteur n'est utilisable que pour les couples direct/réverbéré. Les couples direct/direct et réverbéré/réverbéré ne sont pas concernés par ce descripteur, on les considère donc comme équiprobables :

$$\begin{cases} p(\mathcal{C}_j = \mathcal{C}^d, \mathcal{C}_l = \mathcal{C}^d | d^\tau) = 0.5 \\ p(\mathcal{C}_j = \mathcal{C}^r, \mathcal{C}_l = \mathcal{C}^r | d^\tau) = 0.5 \end{cases} \tag{5.11}$$

5.1.4. Récapitulatif

Trois descripteurs ont été exposés ici, permettant *a priori* de discriminer à la fois les composantes extraites par ACI qui correspondent à des ondes planes ou du champ diffus, et les composantes qui relèvent du champ direct et de premières réflexions. Ces descripteurs sont basés à la fois sur les statistiques des composantes extraites (cohérence moyenne et retard de groupe) et sur la matrice de mélange estimée (critère onde plane). Il reste alors à mettre en place le classifieur intégrant ces trois descripteurs.

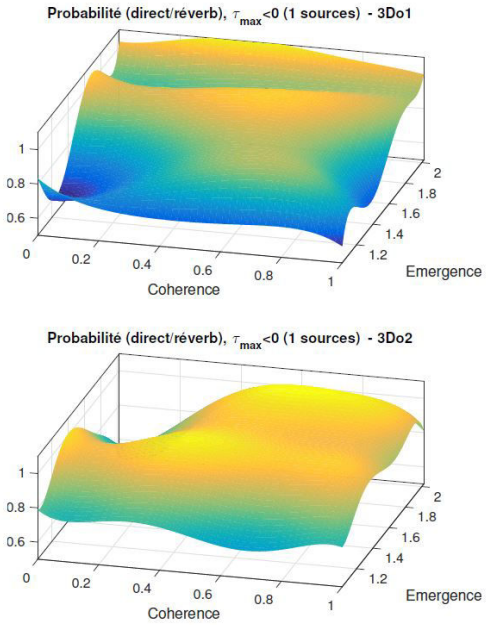

FIGURE 5.5. – Probabilité du couple (direct/réverbéré) lorsque la composante 1 est en avance de phase, en fonction de la cohérence moyenne et de l'émergence. Présence d'une seule source active. Haut (bas) : contenu ordre 1 (2).

5.2. Classifieur mis en oeuvre

5.2.1. Approche générale

Parmi les M composantes extraites par ACI se trouvent donc N composantes directes et $M - N$ composantes réverbérées qui appartiennent à l'une des deux classes :

— $\mathcal{C}^d$ pour les champs directs,

— $\mathcal{C}^r$ pour les champs réverbérés.

Pour une composante s_j, on note $\mathcal{C}_{s_j}$ ou simplement $\mathcal{C}_j$ la classe correspondante.
S'agissant de classer l'ensemble des M composantes extraites, on nomme "configuration" le vecteur des classes $\mathbf{C}$ de dimension $1 \times M$ tel que :

$$\mathbf{C} = [\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M], \quad \mathcal{C}_j \in \{\mathcal{C}^d, \mathcal{C}^r\} \tag{5.12}$$

Sachant qu'il existe deux classes possibles pour chaque composante, le problème revient finalement à choisir parmi un total de 2^M configurations potentielles supposées équiprobables. Pour ce faire, la règle du maximum *a posteriori* est appliquée : connaissant $L(\mathbf{C}_i)$ la vraisemblance de la i^{e} configuration, la configuration retenue sera celle possédant la vraisemblance maximale, c'est-à-dire :

$$\mathbf{C} = \arg\max_{\mathbf{C}_i} L(\mathbf{C}_i) \quad \forall\ 1 \leq i \leq 2^M \tag{5.13}$$

L'approche choisie est donc exhaustive et consiste à estimer la vraisemblance de toutes les configurations possibles, à partir des descripteurs présentés précédemment.

5.2.2. Estimation de la vraisemblance

Pour estimer la vraisemblance de chaque configuration à l'aide des descripteurs mesurés, une approche naïve bayésienne est utilisée. Dans ce type d'approche, on se donne un ensemble de descripteurs d_k pour chaque composante s_j. Pour chaque descripteur, on formule la probabilité pour la composante s_j d'appartenir à la classe $\mathcal{C}^\alpha$ ($\alpha = r$ ou d) grâce à la loi de Bayes :

$$p(\mathcal{C}_j = \mathcal{C}^\alpha | d_k) = \frac{p(\mathcal{C}_j = \mathcal{C}^\alpha) p(d_k | \mathcal{C}_j = \mathcal{C}^\alpha)}{p(d_k)} \tag{5.14}$$

Les deux classes $\mathcal{C}^r$ et $\mathcal{C}^d$ étant supposées équiprobables, il en découle :

$$p(\mathcal{C}_j = \mathcal{C}^\alpha) = \frac{1}{2} \quad \forall \alpha \tag{5.15}$$

ainsi que

$$p(d_k) = \frac{p(d_k | \mathcal{C} = \mathcal{C}^r) + p(d_k | \mathcal{C} = \mathcal{C}^d)}{2} \tag{5.16}$$

On obtient alors :

$$p(\mathcal{C}^\alpha | d_k) = \frac{p(d_k | \mathcal{C}^\alpha)}{p(d_k | \mathcal{C}^r) + p(d_k | \mathcal{C}^d)} \tag{5.17}$$

où le terme $\mathcal{C}_j = \mathcal{C}^\alpha$ est abrégé en $\mathcal{C}^\alpha$ pour alléger les notations. S'agissant ici de rechercher le maximum de vraisemblance, le terme au dénominateur de chaque probabilité conditionnelle est constant quelle que soit la configuration évaluée. Aussi, on peut par la suite en simplifier l'expression :

$$p(\mathcal{C}^\alpha | d_k) \propto p(d_k | \mathcal{C}^\alpha) \tag{5.18}$$

Pour un descripteur bivarié (comme par exemple la cohérence) faisant intervenir deux composantes s_j et s_l et leurs classes respectives supposées, on étend l'expression 5.18 :

$$p(\mathcal{C}_j = \mathcal{C}^\alpha, \mathcal{C}_l = \mathcal{C}^\beta | d_k) \propto p(d_k | \mathcal{C}^\alpha, \mathcal{C}^\beta) \tag{5.19}$$

et ainsi de suite.
La vraisemblance s'exprime comme le produit des probabilités conditionnelles associées à chacun des K descripteurs, si l'on suppose que ceux-ci sont indépendants :

$$L(\mathbf{C}) = p(\mathbf{d} | \mathbf{C}) = \prod_{k=1}^{K} p(d_k | \mathbf{C}) \tag{5.20}$$

où $\mathbf{d}$ est le vecteur des descripteurs et $\mathbf{C}$ un vecteur représentant une configuration, comme défini dans l'équation 5.12. Plus précisément, un nombre K_1 de descripteurs

univariés est mis à profit pour chacune des composantes, tandis qu'un nombre K_2 de descripteurs bivariés est utilisé pour chaque paire de composantes. Les lois de probabilités des descripteurs étant établies en fonction du nombre de sources supposé et de l'ordre ambisonique, on formule alors l'expression finale de la vraisemblance :

$$L(\mathbf{C}) = \prod_{j=1}^{M} \Big(\prod_{k=1}^{K_1} p(d_k(j)|\mathcal{C}_j, N, m) \prod_{l=j+1}^{M} \prod_{k=1}^{K_2} p(d_k(j,l)|\mathcal{C}_j, \mathcal{C}_l, N, m) \Big) \tag{5.21}$$

où :

— $d_k(j)$ est la valeur du descripteur d'indice k pour la composante s_j,

— $d_k(j,l)$ est la valeur du descripteur bivarié d'indice k pour les composantes s_j et s_l,

— $\mathcal{C}_j$ et $\mathcal{C}_l$ sont les classes supposées des composantes j et l,

— N est le nombre de sources actives associé à la configuration évaluée :

$$N = \sum_{j=1}^{M} (\mathcal{C}_j = \mathcal{C}^d) \tag{5.22}$$

Pour des raisons calculatoires, on préfère à la vraisemblance sa version logarithmique (log-vraisemblance) :

$$LL(\mathbf{C}) = \sum_{j=1}^{M} \Big(\sum_{k=1}^{K_1} \log p(d_k(j)|\mathcal{C}_j, N, m) + \sum_{l=j+1}^{M} \sum_{k=1}^{K_2} \log p(d_k(j,l)|\mathcal{C}_j, \mathcal{C}_l, N, m) \Big) \tag{5.23}$$

L'équation 5.23 est celle utilisée en définitive pour déterminer la configuration la plus vraisemblable dans le classifieur.

5.3. Procédure d'identification et de comptage

La procédure générale pour compter et identifier les sources actives, faisant intervenir les descripteurs et le classifieur bayésien exposés précédemment, se retrouve sur le schéma de la figure 5.6.

Pour chaque trame de contenu ambisonique, l'ACI est appliquée afin d'extraire à la fois les composantes et la matrice de mélange estimée. Si l'on souhaite identifier les sources actives avec une résolution temporelle plus fine que l'ACI, les composantes peuvent alors être subdivisées en sous-trames, en conservant néanmoins la même matrice de mélange pour chaque sous-trame.

Les descripteurs sont ensuite extraits afin d'évaluer la vraisemblance associée à chacune des 2^M configurations potentielles. Celle-ci est obtenue au moyen des lois de probabilité conditionnelles, en fonction de l'ordre ambisonique du contenu et du nombre de sources actives supposé de chaque configuration (équation 5.23). Au final, la configuration présentant le maximum de vraisemblance est retenue, indiquant la classe directe ou réverbérée associée à chacune des M composantes.

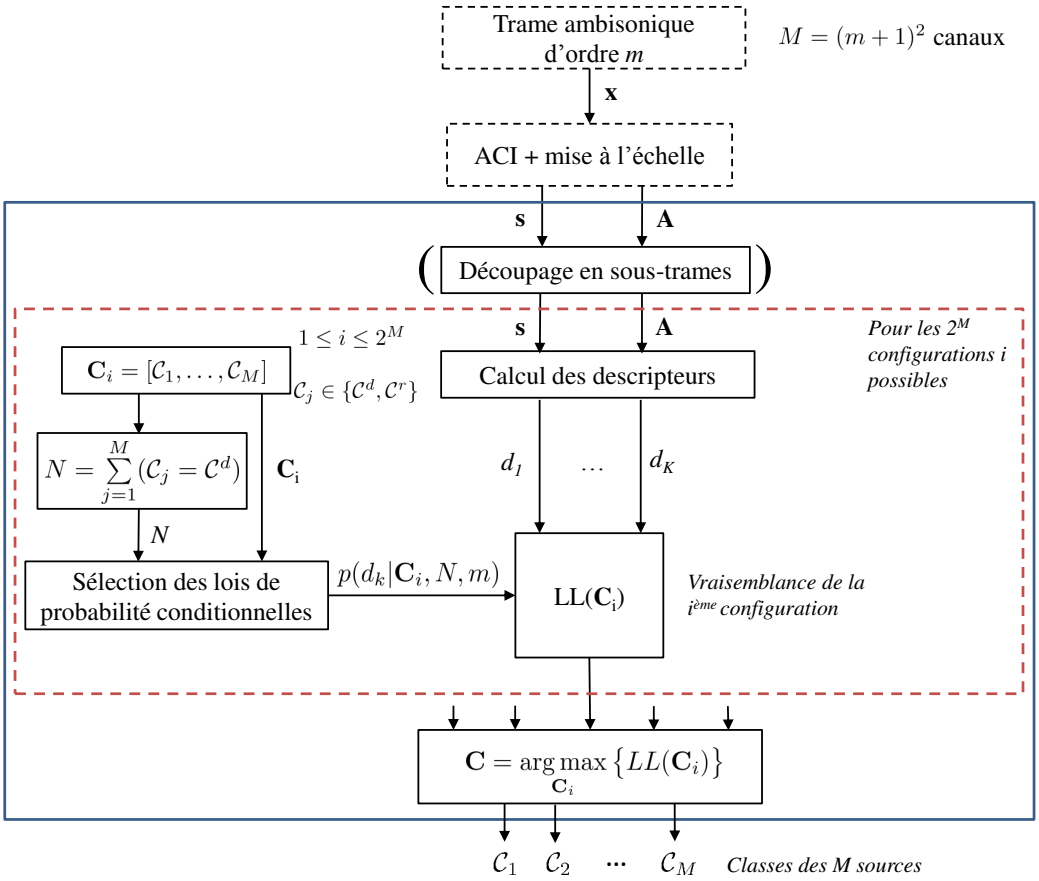

FIGURE 5.6. – Procédure de comptage des sources. N : nombre de sources actives dans la configuration $\mathbf{C}_i$. $LL(\mathbf{C}_i)$: log-vraisemblance de la configuration $\mathbf{C}_i$.

5.4. Conclusion

L'outil de classification bayésienne décrit dans ce chapitre a été mis au point afin de répartir les composantes extraites par ACI en deux classes : composantes directes et composantes réverbérés. Cette classification permet d'identifier le nombre de sources sonores actives dans un mélange réel, malgré la présence de l'effet de salle qui ajoute aux contributions directes des sources des premières réflexions et un champ sonore diffus.
Des descripteurs univariés et bivariés sont sélectionnés pour chaque composante ou chaque couple de composantes, des lois de probabilités conditionnelles fonction du nombre de sources supposées et de l'ordre ambisonique sont modélisées, puis l'ensemble des combinaisons directes/réverbérées sont évaluées afin de retenir pour chaque composante la classe la plus vraisemblable.
Des résultats de comptage obtenus grâce à cette méthode de classification sont abordés dans le chapitre suivant, à travers une utilisation pour une application de domotique.

6. Application de l'outil d'analyse de scène pour la commande vocale

L'algorithme de comptage présenté dans le chapitre précédent permet de dénombrer les sources actives dans le mélange à partir de la décomposition spatiale opérée par l'ACI. La granularité nécessaire à l'ACI, typiquement 500 ms pour une analyse robuste de contenus réels, peut par ailleurs être affinée pour le comptage des sources actives, en subdivisant les composantes extraites en trames de plus petite taille, typiquement 125 ms ou 250 ms. On obtient alors un système complet d'analyse de scène, permettant d'identifier et de localiser au cours du temps les sources actives dans un contenu ambisonique réel, dont les principales étapes sont rappelées par le schéma de la figure 6.1, qui reprend une partie des briques algorithmiques présentées dans le chapitre 3 pour l'analyse de contenus synthétiques. Les étapes 1 à 3 correspondent à la captation, le filtrage et le découpage en trames des observations, tandis que l'analyse s'effectue entre les étapes 4 et 6 (ACI, comptage des sources et localisation) et que les étapes 7 et 8 permettent la reconstruction des signaux à partir des trames d'analyse successives (uniquement dans la bande de fréquences utilisée pour l'ACI).
Un projet de recherche interne à Orange Labs a démarré récemment, portant sur un prototype d'assistant vocal avec prise de son multicanal en champ lointain pour une application de domotique. Le but de ce projet est de développer un dispositif de commande vocal pour des tâches domestiques : allumer la télévision, changer de chaîne, fermer les volets, allumer la lumière, etc...
Dans le cadre de ce projet, l'algorithme d'analyse de scène décrit dans les chapitres 3 et 4, associé à une captation ambisonique, a été implémenté comme un *front-end* acoustique en amont d'un moteur de transcription automatique. Cette expérimentation a eu pour objectif d'évaluer la valeur ajoutée d'un tel dispositif de prise de son par rapport à une captation omnidirectionnelle, notamment en présence de sources interférentes stationnaires (télévision allumée) ou intermittentes (sources de voix concurrentes).
Ce chapitre détaille les résultats obtenus :

— Tout d'abord en termes de détection d'activité des sources actives (source d'intérêt et sources interférentes), ce qui permet l'évaluation du module de comptage de sources décrit dans le chapitre précédent,

— Puis en termes de retranscription automatique, en comparaison avec une prise de son omnidirectionnelle.

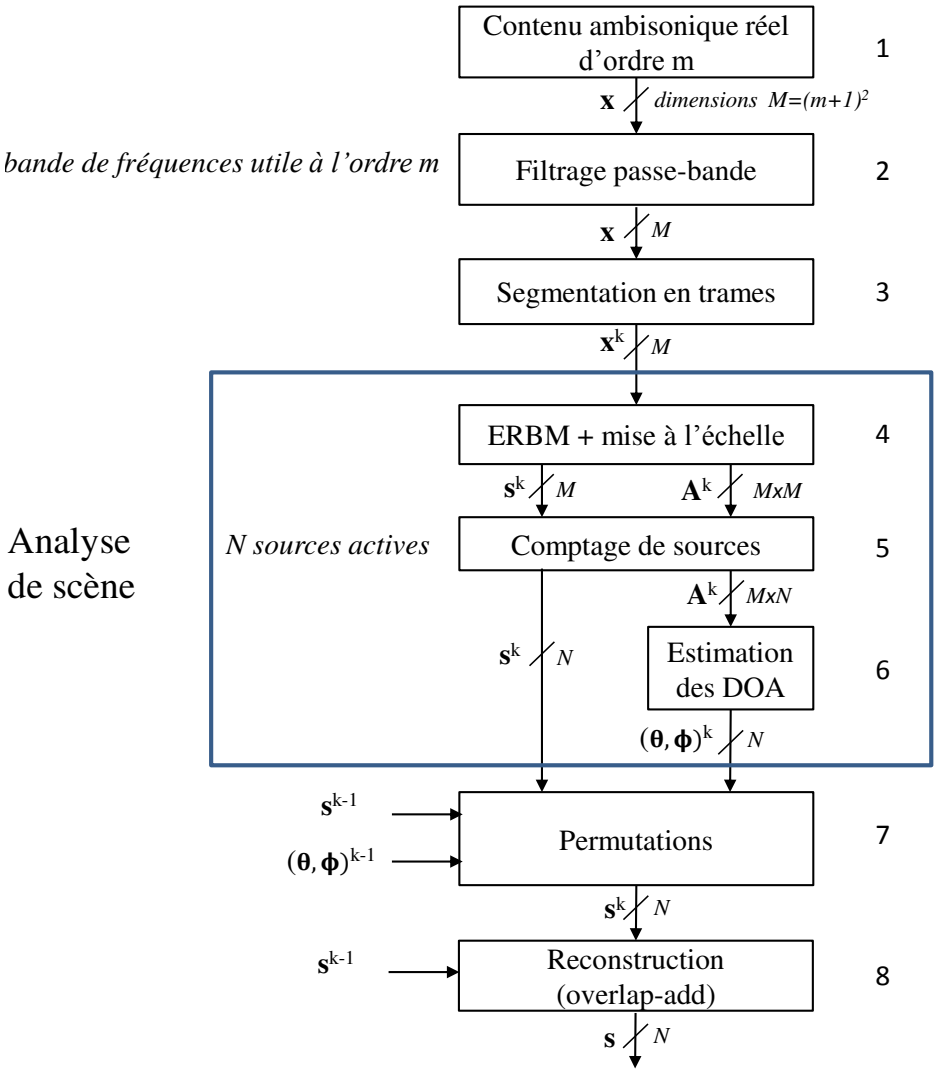

FIGURE 6.1. – Schéma-bloc de l'algorithme d'analyse de scène pour des contenus ambisoniques réels d'ordre m. Extraction des sources $\mathbf{s}$ uniquement dans la bande de fréquences utilisée pour l'analyse.

6.1. Protocole expérimental pour le comptage de sources

6.1.1. Contenus utilisés

Six scènes ambisoniques réelles d'ordre 2, d'une durée de 5 min chacune, sont générées pour évaluer la robustesse du module de comptage de sources. Parmi celles-ci, on trouve :

— Deux scènes contenant deux sources de voix intermittentes, de niveau sonore équivalent,

— Deux scènes contenant une source de voix intermittente et une source de bruit stationnaire (télévision) de niveau sonore équivalent,

— Deux scènes comportant deux sources de voix et une source de bruit (télévision), les trois sources étant de niveaux globalement identiques lorsque celles-ci sont actives simultanément.

Les sources de voix (hommes/femmes) sont en français et issues d'un corpus interne d'Orange Labs. Positionnées dans le plan horizontal à 2.30 m du microphone, les différentes sources sont à une distance variant entre 60° et 120° d'azimuth les unes des autres. Les scènes ont été générées à partir de sources monophoniques et de réponses impulsionnelles mesurées dans la salle d'écoute illustrée en figure 1.12, suivant le protocole décrit en annexe B. Cette salle, de dimensions 5 m × 5.50 m possède un TR_{60} d'environ 270 ms. Les contenus ainsi générés sont constitués d'une ou deux sources de voix concurrentes, additionnées de la source interférente (télévision).

6.1.2. Mesures de performances

L'ACI est appliquée aux différents contenus sur des trames de 500 ms tandis que la détection de sources est opérée sur des trames de 250 ms suivant le protocole de la figure 5.6. La subdivision des trames d'ACI pour le comptage est illustrée par la figure 6.2, donnant pour chaque trame $\mathbf{x}^k$ deux sous-trames $\mathbf{x}^{k,1}$ et $\mathbf{x}^{k,2}$. Par simplicité, c'est la notation $\mathbf{x}^k$ qui sera conservée par la suite, sauf indication contraire.

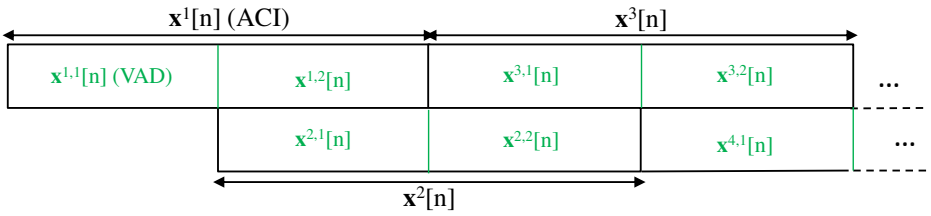

FIGURE 6.2. – Découpage du signal en sous-trames pour le comptage de sources. En noir : trames pour l'ACI avec un recouvrement de 50%. En vert : trames pour le comptage de sources, deux fois plus courtes.

Les vecteurs d'activités estimés $\tilde{\mathbf{C}} = [\tilde{\mathcal{C}}_1, \cdots, \tilde{\mathcal{C}}_M], \mathcal{C}_j \in \{\mathcal{C}^d, \mathcal{C}^r\}$ (voir équation 5.12) sont estimés trame à trame et comparés avec un vecteur d'activité de référence $\tilde{\mathbf{C}}$, obtenu grâce à une détection d'activité appliquée aux sources mono initiales et dont le calcul est détaillé dans la section suivante.
Trois critères sont utilisés pour l'évaluation du comptage de sources sur l'ensemble des trames k :

— Le taux de bonne détection, qui correspond au pourcentage de composantes directes

correctement identifiées :

$$\tau_{\text{detection}} = \frac{\sum\limits_{k,j} \tilde{\mathcal{C}}_j = \mathcal{C}^d}{\sum\limits_{k,j} \mathcal{C}_j} \cdot 100 \quad \forall\, \mathcal{C}_j = \mathcal{C}^d \tag{6.1}$$

— Le taux de fausses alarmes, correspondant au pourcentage de composantes réverbérées identifiées comme directes par l'algorithme :

$$\tau_{\text{alarm}} = \frac{\sum\limits_{k,j} \tilde{\mathcal{C}}_j = \mathcal{C}^d}{\sum\limits_{k,j} \mathcal{C}_j} \cdot 100 \quad \forall\, \mathcal{C}_j = \mathcal{C}^r \tag{6.2}$$

— Le taux de composantes correctement classées, directes ou réverbérées :

$$\tau_{\text{overall}} = \frac{\sum\limits_{k,j} \tilde{\mathcal{C}}_j = \mathcal{C}_j}{\sum\limits_{k,j} \mathcal{C}_j} \cdot 100 \quad \forall\, \mathcal{C}_j \tag{6.3}$$

Le recouvrement de 50% des trames d'ACI, elles-mêmes divisées en deux pour le comptage, amène à effectuer deux fois la détection sur le même segment temporel, aux trames k et $k+1$ (figure 6.2). Le résultat va possiblement différer si les composantes en sortie de l'ACI diffèrent d'une trame à l'autre. Pour le calcul des performances de comptage, les indicateurs de chaque sous-trame sont ici considérés indépendamment les uns des autres.

6.1.3. Références pour le comptage de sources

Pour évaluer la performance du comptage de sources actives, il est nécessaire de disposer, pour chaque trame, d'une information d'activité de référence pour chacune des sources. Pour cela, on calcule une SAD (pour *Source Activity Detection*) pour chaque source utilisée (signal de parole ou télévision) de la manière suivante :

1. On mesure sur la source mono un niveau de bruit, par sélection « à la main » des phases de bruit seul (source éteinte).
2. Le niveau RMS de la source est calculé sur des trames de 12.5 ms et celle-ci est classée comme active si son énergie est supérieure de 30 dB à celle du bruit. Elle est considérée comme inactive dans le cas contraire.
3. Chaque trame de comptage de 250 ms est ensuite classée comme active si elle possède au moins 50 ms d'activité, soit 4 trames de 12.5 ms.

Le choix de 50 ms est arbitraire, il est cependant justifié par le fait qu'il s'agit de l'ordre de grandeur pour lequel une non-détection de la source va être dommageable pour l'algorithme de transcription automatique de la parole.
La figure 6.3 illustre un exemple de SAD sur les trames de 12.5 ms. Sur un signal de parole (haut), on visualise parfaitement les différentes phases d'activité détectées par la SAD. Sur le signal de télévision (bas), la SAD détecte une activité quasi-permanente (flux continu), hormis certaines phases de transition (coupure publicité, changement de séquence...).

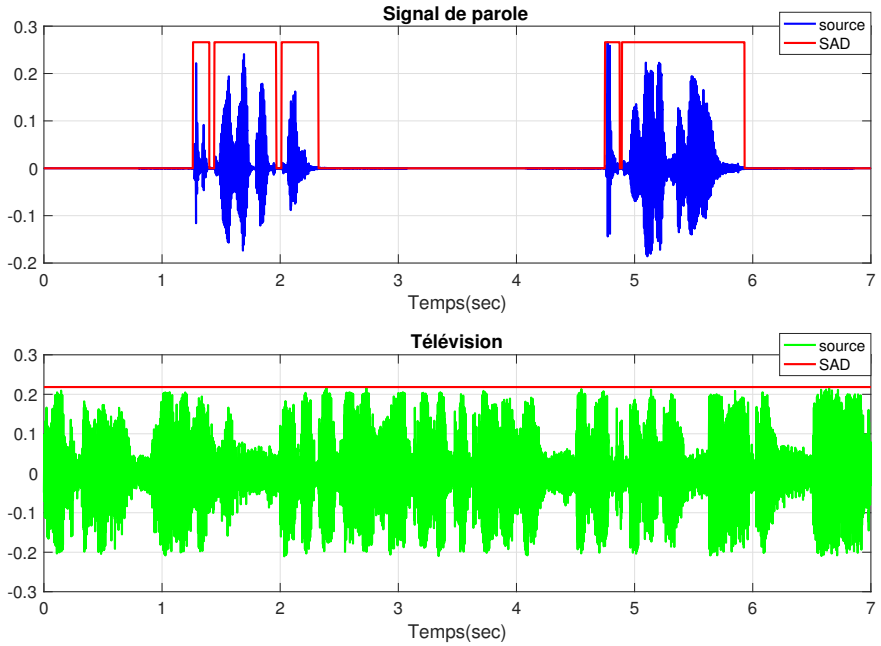

FIGURE 6.3. – Détection d'activité vocale de référence sur les sources mono. En haut : signal de parole intermittent. En bas : source interférente (télévision). En rouge : détection d'activité sur des trames de 12.5 ms.

6.2. Résultats du comptage de sources

La figure 6.4 présente les performances de l'algorithme de comptage de sources en fonction du nombre de sources concurrentes, pour l'ensemble des scènes décrites en 6.1.1. Le pourcentage de sources actives détectées ($\tau_{\text{detection}}$, équation 6.1) est respectivement de 88% et 84% en présence de deux ou trois sources sonores. Le taux de fausses alarmes (τ_{alarm}, équation 6.2) est très faible dans les deux cas, avec une valeur d'environ 4%. S'agissant de contenus ambisoniques d'ordre 2, il y a entre 0 et 3 sources actives par trame, et donc entre 6 et 9 composantes réverbérées, soit plus du double. Le faible taux de fausses alarmes a pour conséquence un taux élevé de composantes correctement identifiées (τ_{overall}, équation 6.1), atteignant respectivement 92% et 94% en présence de deux ou trois sources.
Ces scores reflètent les performances globales de la méthode, mais ils ne renseignent pas sur la distribution des erreurs commises, en particulier pour des sources non stationnaires comme les signaux de parole. La figure 6.4 montre qu'il existe une corrélation entre la durée d'activité d'une source et la probabilité que celle-ci soit correctement détectée. Ainsi, pour un signal actif sur seulement 100 ms, le taux de bonne détection est situé entre 60 et 70% suivant le nombre de sources simultanées. En revanche, lorsqu'une source est active au moins 200 ms par trame, celle-ci est détectée correctement dans au moins 90% des cas, pour deux ou trois sources simultanées. Plus la durée d'activité croît et

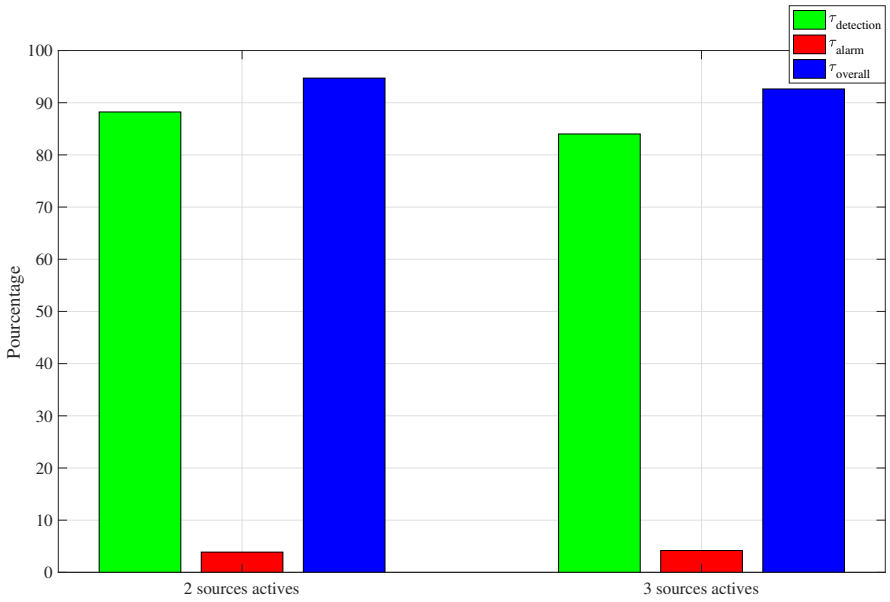

FIGURE 6.4. – Performances de l'algorithme de comptage de sources (trames de 250 ms). En vert : composantes actives (activité $>$ 50 ms) correctement détectées (%). En rouge : composantes réverbérées détectées comme des sources actives (%). En bleu : composantes correctement identifiées (%).

plus les sources sont donc détectées de façon systématique. Pour les signaux de parole, les zones de faible durée d'activité correspondent en général aux périodes transitoires : attaque et fin de phrase ou d'instruction.
Le faible taux de détection sur les débuts et fins de phrase peut être nuancé par le fait que la détection est également fortement liée au rapport signal-sur-bruit. Les résultats présentés ici ont été obtenus sur des contenus dont les sources possèdent un niveau énergétique sensiblement équivalent, ce qui semble particulièrement ardu pour un cas d'usage réel : cela suppose que le locuteur ne hausse pas son niveau de voix en présence de bruit, ou que la source de bruit est elle-même très proche de l'assistant vocal. La figure 6.6 donne le taux de bonne détection en fonction du rapport signal à bruit dans la scène initiale, le bruit étant une source sonore parasite (TV) ou une source de parole concurrente. On observe que le taux de bonne détection atteint 98% dès lors que le SNR initial est nul ou positif. A l'inverse, celui-ci tombe à 91% pour un SNR initial de -5 dB. Cela signifie que les non-détections, en dehors de quelques passages transitoires, concernent la plupart du temps des trames où le niveau de la source d'intérêt est nettement plus faible que celui des sources interférentes.
Afin d'avoir une vision plus synthétique des traitements réalisés, on visualise sur la figure 6.7 les signaux obtenus aux différentes étapes de l'algorithme pour extraire une source de parole d'un mélange voix + télévision. On visualise d'abord la source de parole que l'on cherche à isoler et sa SAD, puis le mélange (canal omni), et le signal extrait par ACI. Le signal obtenu après l'étape de dénombrement (courbe du bas) permet de visua-

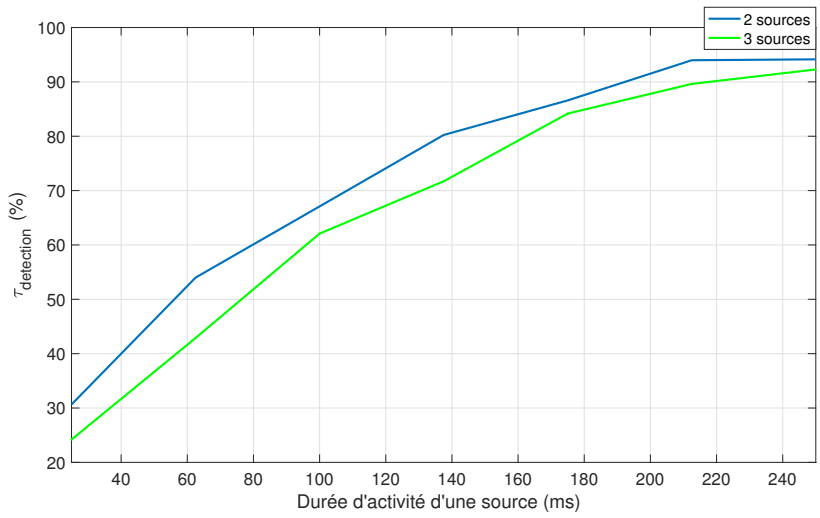

FIGURE 6.5. – Pourcentage de détection des sources actives, en fonction de la durée d'activité de celles-ci. Trames d'analyse de 250 ms. En bleu (vert) : deux (trois) sources actives concurrentes.

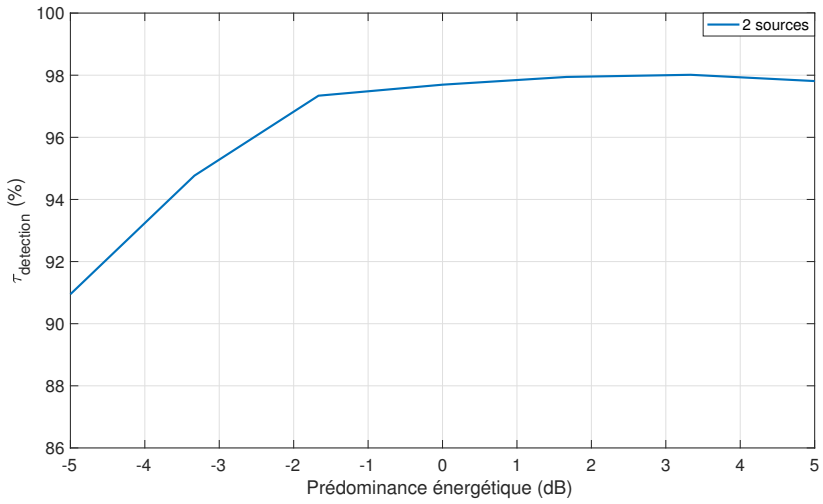

FIGURE 6.6. – Pourcentage de détection des sources actives, en fonction de la prédominance énergétique de la source. Deux sources actives concurrentes. Trames d'analyse de 250 ms, sources actives sur au moins 200 ms.

liser la SAD estimée sur les périodes d'activité de chaque source estimée et illustre bien une partie des observations précédentes. La détection d'activité est globalement performante, avec les zones d'activité continue parfaitement identifiées et un nombre faible de fausses alarmes, cependant les périodes de transition sont moins bien détectées et il arrive fréquemment que quelques dizaines de ms soient oubliées en début et fin de phrases.

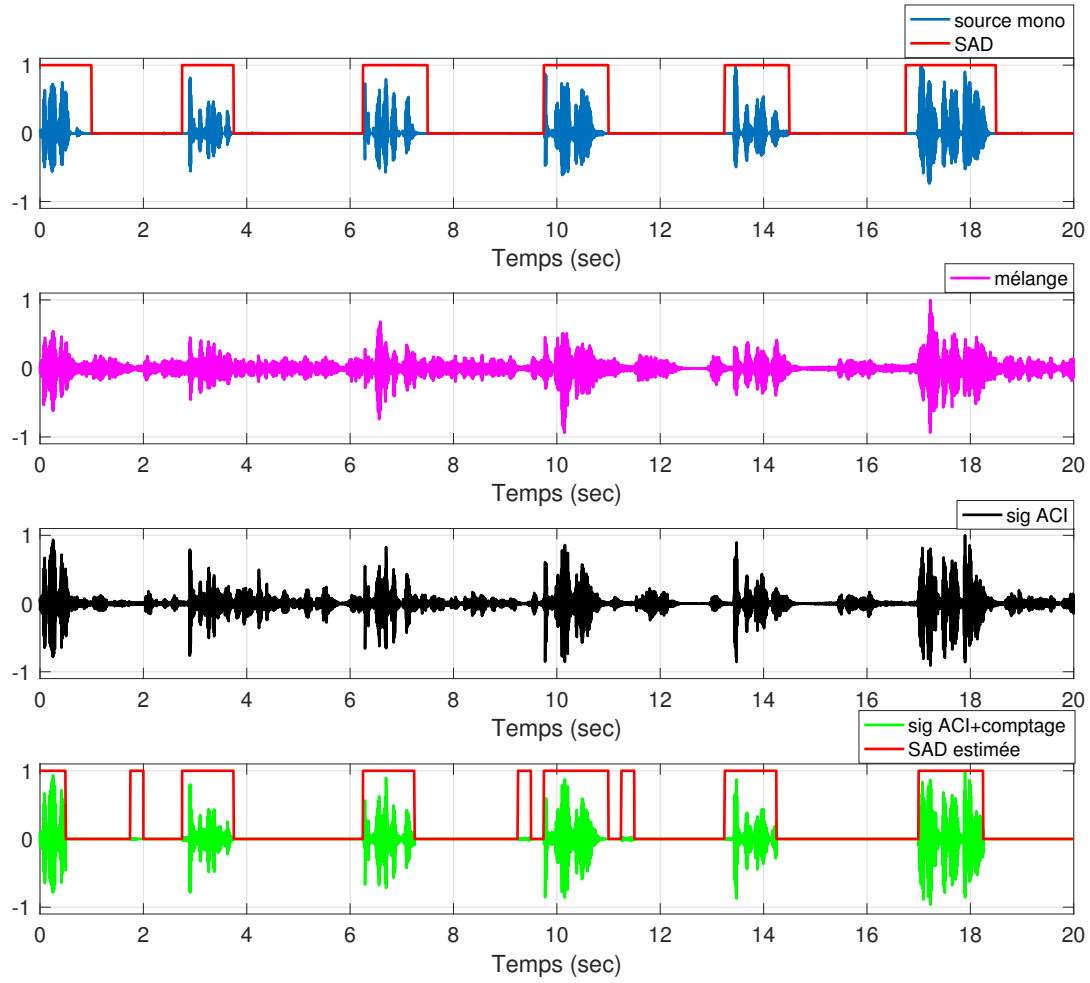

FIGURE 6.7. – Illustration du résultat obtenu avec le comptage de sources, en présence de deux sources (voix+tv). De haut en bas : 1) source de voix initiale. 2) Mélange (canal omnidirectionnel). 3) source de voix extraite par ACI. 4) signal extrait après détection d'activité. ACI sur 500 ms, comptage sur 250 ms.

6.3. Discussion sur le procédé de comptage

Nous avons montré ici que la procédure de comptage de sources décrite dans le chapitre 5 pouvait être appliquée à des contenus ambisoniques réels d'ordre 2 pour traiter des mélanges de signaux de parole intermittents et de signaux de bruit (télévision). Pour des trames d'analyse de 250 ms, nous avons obtenu de bonnes performances générales : 88%/84% de bonne détection en présence de 2/3 sources dans des conditions plutôt adverses (sources concurrentes de niveaux sonores sensiblement équivalents), tout en conservant un très faible taux de fausses alarmes : environ 4%.
Dans des configurations plus favorables, mais néanmoins réalistes - source d'intérêt active

sur la majeure partie de la trame d'analyse, avec un SNR nul ou légèrement positif - on obtient même un taux de bonne détection de quasiment 100%.
Une inflexion des performances est cependant à noter lorsqu'il s'agit de détecter des sources actives sur des durées nettement inférieures à la trame d'analyse (250 ms), ce qui correspond à la détection des débuts et fins de phrases. Le temps a manqué pour améliorer la robustesse dans ces cas de figure, mais plusieurs pistes sont déjà envisageables :

— Une première solution rapide à mettre en œuvre serait de réduire la taille de trame de comptage, à la fois pour l'entraînement de la base de données et pour l'analyse. Les descripteurs auraient probablement une variabilité plus grande, mais cela permettrait sûrement de mieux gérer les zones de transition.

— Pour une application de type assistance vocale, la contrainte de causalité peut être relâchée, car la transcription automatique nécessite de toute manière d'attendre la fin de la phrase prononcée. En pratique, cela signifie que l'on peut exploiter le futur dans une certaine limite, notamment pour mieux détecter les attaques *a posteriori*.

— Au milieu d'une zone de signal actif, les non-détections sont rares, et ne concernent en général qu'une trame isolée. Un mécanisme simple de décision pourrait être mis en place, permettant d'aller récupérer l'information manquée lorsque l'on détecte une activité sur les trames précédant et suivant la trame "perdue".

De manière plus générale, on pourrait intégrer la vraisemblance estimée dans un algorithme de décision statistique, par exemple de type HMM (*Hidden Markov Model*) avec un apprentissage de type *forward-backward* (utilisation des états antérieurs et postérieurs) et une prise de décision avec un algorithme de Viterbi.
Pour finir, l'environnement acoustique dont sont issues les différentes scènes ne présente pas un niveau de réverbération très élevé, il faudra donc évaluer l'algorithme en présence d'un effet de salle plus important. De premières expérimentations à partir de sources captées dans une salle peu absorbante (figure 4.1) ont déjà mis en évidence de bons résultats d'analyse, au prix cependant d'un SNR plus favorable.

6.4. Application à la transcription automatique

Grâce à l'étape de comptage des sources couplée à l'algorithme d'ACI, on dispose à présent d'une chaîne de traitement complète, fonctionnant à l'aveugle, pour analyser un contenu ambisonique réel et en séparer les principales sources. Cette section présente un cas d'application concret de cet outil.
Des équipes d'Orange Labs travaillent actuellement sur la problématique de la retranscription automatique de la parole pour différentes applications en lien avec les télécommunications ou la domotique. L'amélioration des performances de retranscription dans des conditions acoustiques peu favorables - captation en champ lointain, présence de bruit ou de sources interférentes - est une des applications directes de l'outil d'analyse de scène mis en place. Aussi, les contenus sonores détaillés dans la première section de ce chapitre ont servi à évaluer l'apport de l'analyse de scène pour une application de commande vocale à usage domestique.

En pratique, l'algorithme d'analyse est inséré entre la captation et la retranscription pour former l'outil de bout-en-bout, schématisé sur la figure 6.8. Comme on peut le voir, l'outil de reconnaissance peut potentiellement être appliqué à l'ensemble des sources identifiées, ce qui permet théoriquement de traiter plusieurs instructions simultanément si les sources ont été au préalable convenablement séparées.

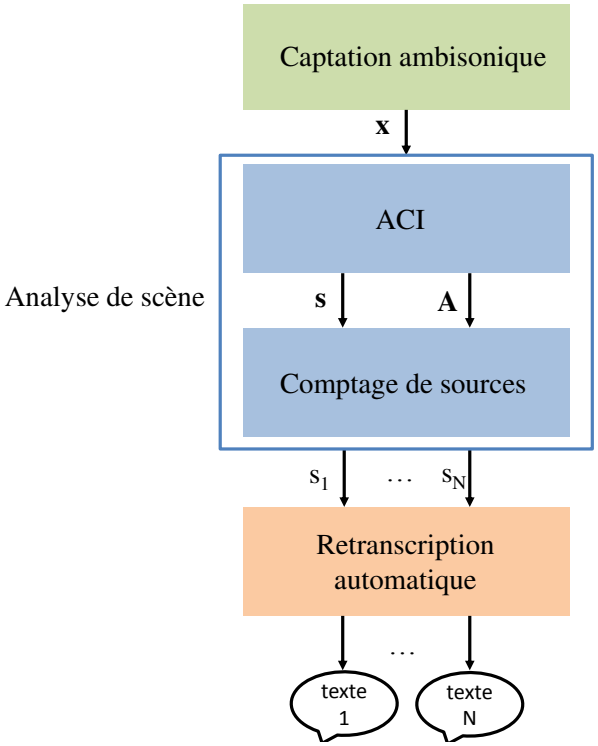

FIGURE 6.8. – Chaîne de traitements de l'outil de reconnaissance vocale : association d'une captation ambisonique, d'un outil d'analyse de scène et d'un moteur de transcription automatique

La reconnaissance vocale est effectuée sur la bande de fréquences 800 Hz-8 kHz (bande de validité de l'ordre 2) grâce à un moteur de retranscription automatique interne à Orange Labs, basé sur l'architecture libre Kaldi [89]. Un modèle acoustique basé sur un réseau de neurones profonds (DNN *Deep Neural Network*) est couplé avec un modèle de langage statistique dit *n-gramme*, permettant d'identifier par maximum de vraisemblance les phonèmes prononcés en fonction du signal acoustique associé mais également des n phonèmes adjacents.

6.4.1. Protocole expérimental

Les scènes sonores décrites en 6.1.1 sont d'abord analysées pour en extraire les différentes sources sonores (schéma de l'algorithme présenté en figure 6.1). On rappelle qu'il s'agit de contenus ambisoniques d'ordre 2 constitués de deux ou trois sources sonores : un ou

deux signaux de parole et un bruit stationnaire, en l'occurrence un contenu télévisuel. Parmi les sources extraites, les sources d'intérêt (voix uniquement) sont identifiées à la main puis traitées par le module de reconnaissance vocale, qui retourne alors la transcription estimée de chacune d'elles. Les résultats issus de la retranscription sont ensuite comparés à la transcription exacte, fournie avec le corpus de voix utilisé.
Pour évaluer la qualité de la retranscription obtenue, une métrique issue du *Hidden Markov Model Toolkit* (HTK) [90] est utilisée. Celle-ci est basée sur un ensemble d'indicateurs :

- N le nombre de mots à reconnaître,
- H le nombre de mots correctement reconnus,
- S le nombre de substitutions (retranscription incorrecte),
- D le nombre de suppressions (*Deletions*), c'est-à-dire de mots non détectés,
- I le nombre d'insertions, c'est-à-dire le nombre de mots détectés alors même qu'il n'y a pas de mot prononcé,

avec la relation $N = H + S + D$. Ces indicateurs permettent d'obtenir les scores de performances suivant :

- le pourcentage de mots correctement reconnus :

$$\tau_{\text{correct}} = \frac{H}{N} \cdot 100 \tag{6.4}$$

- la précision :

$$\tau_{\text{accuracy}} = \frac{H - I}{N} \cdot 100 \tag{6.5}$$

- le taux d'erreurs-mot (WER-*Word Error Rate*) :

$$\tau_{\text{wer}} = \frac{S + D + I}{N} \cdot 100 = 100 - \tau_{\text{accuracy}} \tag{6.6}$$

On déduit de leurs formules respectives que le pourcentage de mots correctement reconnus est compris entre 0 et 100, mais que la précision peut prendre des valeurs négatives, lorsque le nombre d'insertions est supérieur au nombre de mots correctement reconnus. En pratique, cela se produit lorsqu'une source est intermittente et que d'autres sources de voix ou de bruit viennent polluer le signal pendant les périodes d'inactivité. Mécaniquement, le taux d'erreurs-mot dépasse alors 100, ce qui signifie qu'il y a plus d'erreurs que de mots correctement reconnus. Le WER est un critère que l'on retrouve souvent dans la littérature, néanmoins il se déduit directement de la précision, nous avons donc choisi d'exposer les résultats uniquement en termes de pourcentage de mots reconnus et de précision.
Pour juger de l'intérêt de l'outil d'analyse, il est nécessaire de comparer les scores obtenus à des scores de référence. Pour cela, l'algorithme de transcription automatique est également appliqué à :

- la scène complète, sans séparation de sources (captation omnidirectionnelle),

— chaque source seule dans son environnement acoustique (captation omnidirectionnelle). Cette valeur sera prise comme valeur de référence.

On obtient ainsi une valeur plancher avec l'analyse de la scène complète sans traitement qui est le point de départ que l'on cherche à améliorer et une valeur idéale à atteindre avec la source seule. On peut toutefois remarquer que ce cas est idéal uniquement en ce qui concerne l'absence de bruit et d'interférences, mais que la présence d'effet de salle dans la captation omnidirectionnelle dégrade possiblement les résultats obtenus.

6.4.2. Performances

Les taux de mots correctement reconnus et la précision en présence de deux sources - une source de voix intermittente mélangée à une source de bruit stationnaire (télévision) - sont présentés sur la figure 6.9. La première observation est que les performances de la retranscription sur la source extraite (en rouge) sont largement supérieures à celles de la scène complète (en vert), que ce soit en termes de précision ou de taux de mots correctement reconnus. Alors qu'il est de seulement 14% pour la scène bruitée, le taux de mots correctement reconnus monte ainsi à 63% après extraction de la source d'intérêt, soit quasiment 2 mots sur 3, même si l'on perd quand-même 20% par rapport au signal de référence. La précision, qui est négative dans le mélange initial à cause des insertions de bruits entre chaque instruction, remonte également à 57% après séparation, ce qui montre que l'outil de comptage de sources a bien segmenté le signal de parole en identifiant correctement les périodes d'activité.

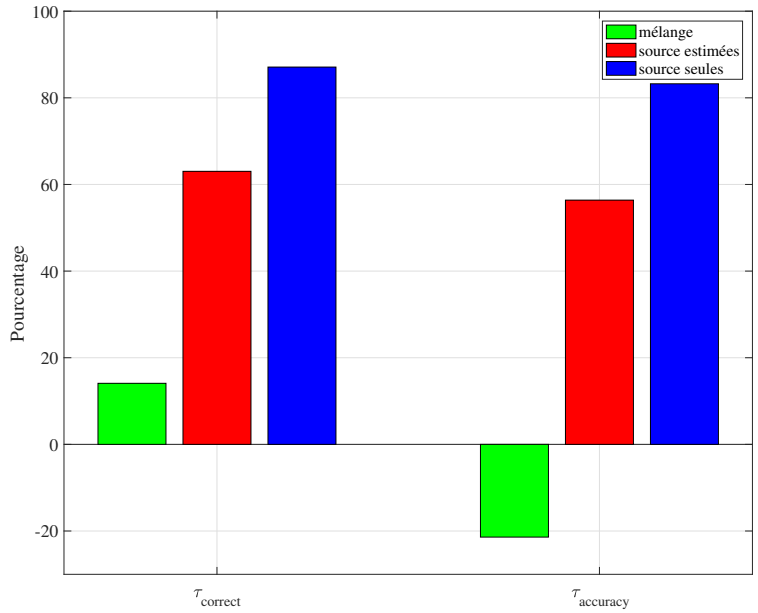

FIGURE 6.9. – Performances de retranscription automatique (taux de mots corrects et précision). Scène captée : parole intermittente + TV.

En présence cette fois de deux sources de voix concurrentes sans bruit additionnel, les observations sont similaires même si les scores sont un peu plus favorables, à la fois pour le mélange et pour les sources séparées (figure 6.10). Les scores présentés sont ici une moyenne des scores respectifs des deux sources. Le taux de mots correctement reconnus dans le mélange est relativement haut (40%) car les deux sources sont intermittentes. Il y a donc des portions de signal pour lesquels la source d'intérêt se retrouve seule active, même si les deux voix se superposent régulièrement. En revanche, ce même taux n'est pas beaucoup plus élevé que dans le premier scénario (67% contre 63% précédemment). Cela suggère donc qu'une source de voix interférente est plus gênante pour la retranscription qu'une source de bruit qui ne contient pas systématiquement d'information vocale. Cette légère baisse peut aussi être interprétée au regard des résultats de comptage exposés précédemment, où une dégradation de la détection est observée sur les attaques et fins de phrases. Le taux de précision après séparation est quant à lui sensiblement équivalent au signal extrait du mélange voix-TV.

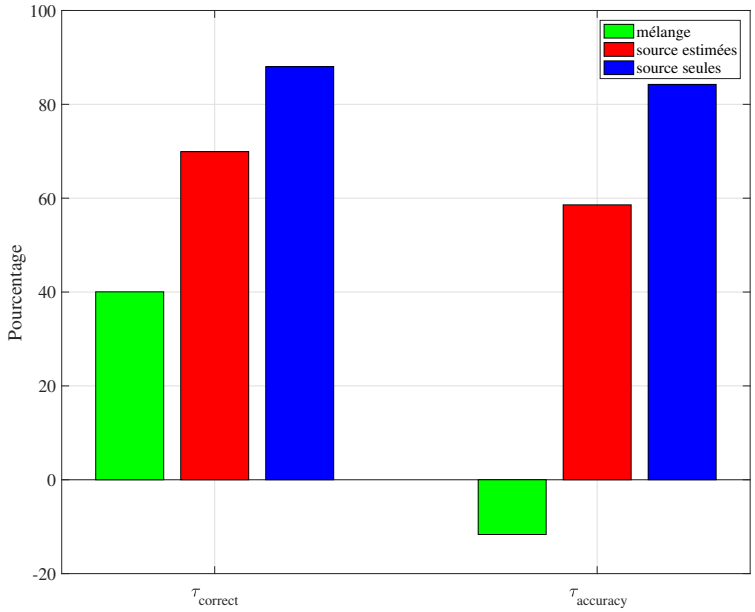

FIGURE 6.10. – Performances de retranscription automatique (taux de mots corrects et précision). Scène captée : 2 sources de voix.

Le troisième cas, le plus ardu à traiter, concerne les scènes constituées de deux sources de parole et d'une source de bruit stationnaire (TV). Comme on peut s'y attendre, les performances sont légèrement dégradées (figure 6.11) par rapport au cas de deux sources de parole seules, mais il faut prendre en compte le point de départ qui est lui aussi nettement dégradé. Ainsi, le taux de mots correctement reconnus est à 63% après séparation alors qu'il est de 83% avec les sources isolées, mais celui-ci tombe à moins de 10% pour une transcription appliquée au mélange. Dans ce cas de figure, la captation

omnidirectionnelle de la scène est donc inutilisable en l'état, et l'apport de l'algorithme d'analyse/séparation apparaît clairement. De plus, la précision est à 50% après séparation alors qu'elle est de -25% pour le mélange total, cela montre que l'étape de comptage des sources actives a supprimé la majorité des insertions malgré la présence de plusieurs sources interférentes.
Enfin, on peut remarquer que la présence de trois sources concurrentes, de niveaux sonores équivalents, est un cas d'usage nettement défavorable qui ne va pas refléter la majorité des usages rencontrés. Néanmoins, l'outil d'analyse permet de reconnaître près de 2 mots sur 3 même dans ce contexte difficile, alors qu'à peine 1 mot sur 10 était reconnu en transcrivant le mélange non traité.

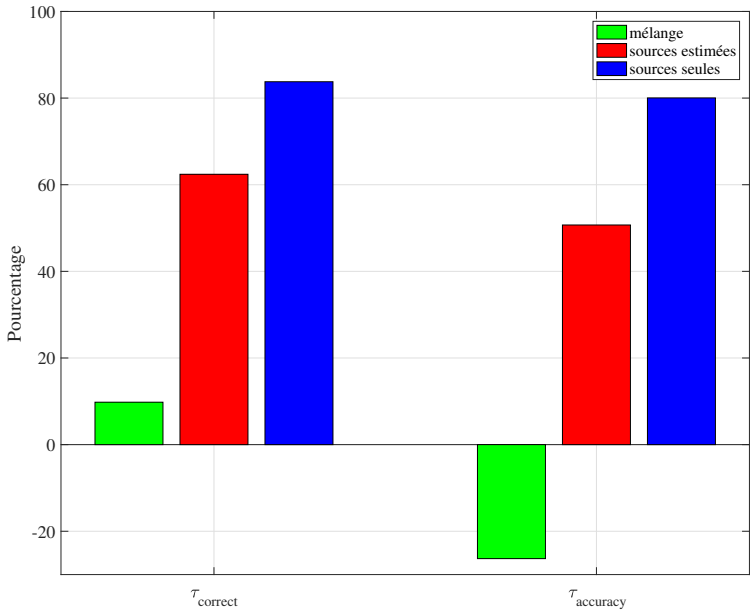

FIGURE 6.11. – Performances de retranscription automatique (taux de mots corrects et précision). Scène captée : 2 sources de voix et une source de bruit (TV).

6.5. Conclusion

Malgré un nombre restreint de configurations évaluées, les résultats présentés dans ce chapitre ont montré l'intérêt de l'outil d'analyse pour un cas d'usage concret, à savoir la reconnaissance vocale pour la domotique, en permettant l'amélioration des performances de reconnaissance en milieu fortement bruité ou en présence de plusieurs sources de voix concurrentes.
Les performances de retranscription automatique obtenues grâce à l'analyse de scène ont été comparées à celles obtenues en applicant la retranscription à la seule captation omnidirectionnelle de la scène sonore. L'étape de séparation des sources par formation de

voies entraîne une hausse significative du taux de mots reconnus en atténuant fortement le bruit et les interférences. De plus, la robustesse de l'algorithme de comptage de sources, évaluée dans la section 6.1.1, permet de baisser très fortement le taux d'insertions dues, c'est-à-dire le nombre de mots détectés sans qu'il n'y ait réellement d'instructions prononcées grâce à une segmentation des commandes vocales reçues.
Pour compléter ces résultats, il aurait été intéressant d'évaluer également l'influence de l'ordre ambisonique utilisé pour l'analyse, notamment en se restreignant à l'utilisation de l'ordre 1.
Finalement, les performances décrites dans les sections 6.1.1 et 6.4 montrent la capacité de l'algorithme à analyser et traiter de façon fiable des contenus ambisoniques réels d'ordre 2 contenant jusqu'à 3 sources sonores simultanées, même si l'étape de séparation ne permet pas de s'affranchir complètement des artefacts engendrés par les sources de bruit.

Conclusion et perspectives

Synthèse des travaux réalisés Cette thèse a permis de mettre au point un outil d'analyse pour dénombrer, localiser et extraire des sources sonores, essentiellement des signaux de parole, dans une captation ambisonique réelle en configuration sur-déterminée. L'approche choisie, complètement aveugle, est basée sur une décomposition du champ sonore par analyse en composantes indépendantes (ACI), permettant d'extraire un certain nombre de composantes sonores dont certaines sont liées à des sources d'intérêt et d'autres à de la réverbération. Une deuxième étape de classification, appliquée aux composantes extraites, identifie celles correspondant aux sources, permettant finalement de dénombrer et de localiser les sources actives de manière robuste.
L'algorithme complet a été étudié de manière exhaustive et validé sur des contenus ambisoniques réels provenant de différents types d'environnements acoustiques. L' étape d'ACI se base sur l'algorithme ERBM qui permet une localisation robuste des sources, y compris dans des conditions réverbérantes, alors que la plupart des algorithmes d'ACI de la littérature voient leurs performances se dégrader fortement en présence d'effet de salle (chapitre 3). Les performances d'ERBM ont été validées par une évaluation portant sur des contenus ambisoniques synthétiques et réels. Nous avons mis en évidence que la robustesse d'ERBM à la réverbération était due à une étape de décorrélation temporelle, initialement prévue pour tirer parti de la structure temporelle des signaux, mais qui a finalement pour conséquence de permettre la séparation du champ direct d'une source et de ses premières réflexions pourtant fortement corrélées (section 3.5). L'utilisation d'ERBM a montré des performances de localisation et également de séparation de sources intéressantes pour des contenus réels contenant jusqu'à trois sources sonores simultanées, et ce même dans des environnements assez réverbérants (chapitre 4).
Par ailleurs, il a été mis en avant l'intérêt d'utiliser les ordres ambisoniques supérieurs pour améliorer à la fois la localisation et la séparation de sources, grâce à un nombre d'observations plus important et donc une sélectivité spatiale accrue. On peut ajouter que des résultats préliminaires encourageants ont été obtenus pour des scènes sonores HOA constituées d'un nombre de sources plus important.
La seconde étape, dite de classification, est fondée sur une approche bayésienne (chapitre 5). Pour réaliser cette classification, des descripteurs ont été proposés afin de discriminer les composantes extraites par ACI : champs directs (sources) ou champ réverbéré. Cette seconde étape a été évaluée sur des contenus ambisoniques réels d'ordre 2 et a montré un taux de bonne détection des sources supérieur à 90% avec une résolution temporelle de 250 ms.
Au final, l'association de ces deux étapes - qui ont fait l'objet de deux brevets - forme un algorithme d'analyse et de séparation aveugle de sources applicable à des contenus ambisoniques réels. Il est montré dans le chapitre 6 que cet outil présente un intérêt

pour effectuer du réhaussement de la voix pour de la reconnaissance vocale et permet d'augmenter significativement les performances de la retranscription automatique pour une prise de son en champ lointain et en milieu fortement bruité (une ou deux sources interférentes de niveaux sonores équivalents).
La segmentation temporelle, bien que pour l'instant limitée à des trames de 250 ms, permet néanmoins de délimiter les débuts et fins de phrases, diminuant ainsi le taux d'insertions lors de la retranscription. L'approche de séparation des sources utilisée, essentiellement spatiale, ne permet pas de débruiter complètement les sources notamment en milieu réverbérant, mais elle permet une dégradation *a minima* du contenu spectral des sources et n'introduit donc pas d'artefacts qui perturberaient le moteur de transcription automatique.
Un prototype de captation multicanal est actuellement mis au point par les équipes d'Orange Labs, intégrant les algorithmes conçus durant la thèse sous forme d'un prétraitement pour la commande vocale à distance.

Limites identifiées L'approche choisie pour traiter le problème de localisation est essentiellement spatiale - une analyse en composantes indépendantes appliquée dans le domaine temporel - et les conditions nécessaires pour une utilisation performante peuvent parfois être difficiles à réunir, en particulier pour une captation ambisonique d'ordre 1. Le nombre de sources possiblement identifiables est inférieur ou égal au nombre de canaux, ce qui correspond à 4 sources à l'ordre 1 et 9 sources à l'ordre 2. La proximité des différentes sources est également un facteur limitant. Un contenu d'ordre 1 possède par exemple une faible capacité de séparation spatiale et ne permet pas d'identifier clairement des sources distantes de seulement 20° lorsque l'effet de salle devient important. Dans ce cas de figure, l'utilisation des ordres ambisoniques supérieurs permet d'augmenter nettement la discrimination spatiale des sources et la précision de la localisation (chapitre 4).
De plus, même si ERBM possède une robustesse plus grande et converge plus rapidement que les autres méthodes d'ACI évaluées, il nécessite quelques centaines de millisecondes de données pour arriver à localiser précisément plusieurs sources dans un contenu réel. Cela peut poser problème si la scène sonore évolue rapidement (déplacement, apparition/disparition d'une source). Par ailleurs ERBM possède une certaine complexité algorithmique et il reste encore des optimisations à réaliser pour que celui-ci soit implémenté pour une utilisation "live".
Par ailleurs, des expérimentations préliminaires ont mis en évidence un manque de robustesse de l'algorithme ERBM lorsque les sources étaient trop périodiques, typiquement en présence de signaux musicaux. L'étape de blanchiment temporel, bien qu'améliorant la robustesse à la localisation, n'est pertinente qu'en présence de signaux à large spectre se prêtant bien à ce type de modélisation.
Concernant la séparation des sources, le choix d'une formation de voies sous contraintes implique des limites identiques à celles de la localisation, à savoir un nombre de sources séparables au plus égal au nombre de canaux. De plus, le système de captation utilisé pendant la thèse ne permet pas d'exploiter les ordres ambisoniques supérieurs à 1 en dessous de 800 Hz, ce qui réduit à 4 le nombre de sources séparables en basses fréquences,

quand bien même le nombre d'observations initiales est largement supérieur. Les ordres supérieurs conservent néanmoins une utilité pour améliorer la localisation et détecter un nombre de sources plus important mais également pour affiner la séparation dans les bandes de fréquences où ceux-ci sont exploitables, selon le procédé décrit dans la section 4.2.
Dernier point, la déréverbération obtenue avec une séparation spatiale n'est que partielle, et le taux de réverbération résiduelle, dépendant du facteur de directivité des directivités formées, peut être assez important lorsque l'ordre 1 est utilisé seul.

Perspectives Les résultats obtenus ont essentiellement porté sur l'analyse de scène et la séparation des sources à l'aide d'indicateurs de performances objectifs. Dans l'optique de manipuler des contenus destinés à l'écoute, il est également nécessaire de prévoir une étape de validation subjective des traitements réalisés suivant des critères psychoacoustiques. Par manque de temps, cette tâche n'a pas pu être réalisée durant la thèse, on peut néanmoins déjà penser à plusieurs critères pour l'évaluation de la manipulation de contenus ambisoniques :

- pour la déréverbération : la qualité du timbre et l'intelligibilité (en particulier pour des signaux de parole),
- pour l'atténuation / le réhaussement d'une source dans une scène : le niveau relatif perçu des différentes sources,
- pour le déplacement d'une source dans une scène : la localisation perçue par l'auditeur pour la source manipulée et la cohésion générale de la scène (étalement de la source déplacée, cohérence entre la localisation de la source et l'effet de salle).

Même en l'absence de validation subjective formelle, les résultats obtenus en termes d'analyse et de séparation ont d'ores et déjà permis de mettre au point de premières expériences de manipulation de contenus 3D destinés à l'écoute, grâce à l'utilisation d'un *plugin* VST de spatialisation développé par les équipes d'Orange Labs. Une démonstration a été effectuée lors de la 60ème conférence de l'AES à Louvain, où les auditeurs pouvaient manipuler en temps-réel des scènes réelles contenant deux sources de voix simultanées enregistrées dans différents environnements acoustiques, avec une restitution binaurale du contenu manipulé.
Un certain nombre de pistes sont également envisageables pour lever les limitations énoncées précédemment. L'une d'entre elles concerne l'étape de séparation, car nous n'avons au cours de la thèse exploité que les informations de localisation pour procéder à l'extraction des sources. Les informations de localisation et d'activité fournies par l'outil d'analyse pourraient aider à identifier la réverbération associée à chaque source afin d'atténuer les interférences résiduelles liées à l'effet de salle, et dé-réverbérer même partiellement les sources extraites. Des approches fréquentielles basées un filtrage de Wiener multicanal, de rang 1 [91] voire de rang plein [92], pourraient exploiter avantageusement les informations de localisation et d'activité pour estimer les matrices de covariance spatiale nécessaires à l'estimation des filtres de séparation. Par ailleurs, la plupart des avancées récentes en termes de séparation de sources ou de réhaussement de voix sont basées sur

des traitements temps-fréquence, permettant de prendre en compte à la fois la parcimonie temporelle et spectrale des sources et de traiter également des mélanges convolutifs. Une approche purement spectrale du problème d'analyse/manipulation de scène serait envisageable, basée par exemple sur des réseaux de neurones profonds (*DNN Deep Neural Network*).
Pour déterminer le nombre de sources actives, la classification des composantes issues de l'ACI a été effectuée à l'aide d'un classifieur bayésien. Ce classifieur présente des performances intéressantes (voir chapitre 5) et il possède l'avantage d'être relativement simple à implémenter. Néanmoins, on peut envisager des performances supérieures en utilisant un autre type de classifieur, comme une machine à vecteurs de support ou un réseau de neurones. Une classification plus robuste permettrait de réduire la durée des trames d'analyse, actuellement de 250 ms, et donc d'obtenir une analyse de scène plus réactive.
Enfin, on peut espérer que les limites d'utilisation observées en basses fréquences pour l'encodage des ordres ambisoniques supérieurs pourront à l'avenir être levées grâce à l'amélioration des systèmes de prises de sons, l'augmentation du nombre de capsules microphoniques et notamment l'utilisation de *clusters* de microphones permettant d'augmenter le rapport signal-à-bruit.

A. Génération de contenus HOA synthétiques

Les contenus ambisoniques utilisés pour évaluer les méthodes de séparation de sources sont générés à partir de sources mono, encodées avec des coefficients de mélange dépendant de leur direction d'arrivée dans le cas instantané et à l'aide de réponses impulsionnelles synthétiques dans le cas réverbérant simulé.

A.1. Sources sonores

Deux types de sources sont utilisées : des voix parlées ou bien des voix chantées associées à de la musique. Les sources de voix parlées sont en français et en anglais. Une partie d'entre elles sont issues d'un corpus d'Orange Labs de phrases phonétiquement équilibrées, enregistrées suivant la norme ITU-T P.800 [93] relative à l'évaluation subjective de la qualité de transmission. L'autre partie provient de la base de données [79] du projet collaboratif SiSEC 2008 portant sur l'évaluation de méthodes de séparation de sources sonores [80]. Les sources musicales et les voix chantées proviennent exclusivement de la base SiSEC 2008.

A.2. Génération des réponses impulsionnelles

Des coordonnées sphériques (θ_j,ϕ_j,r_j) sont attribués à chaque source j. Dans le cas instantané, des coefficients d'encodage ambisoniques relatifs aux coordonnées (θ_j,ϕ_j) sont calculés afin de simuler la contribution de la source à chaque observation, selon l'équation 2.3.
Dans le cas d'un effet de salle simulé avec un encodage microphonique parfait, l'outil de sources images RoomSim [81], basé sur les travaux de Allen et Berkley [82], est utilisé pour la génération d'un peu plus de 10^6 réflexions pour chaque source, dans une salle rectangulaire de dimensions 7 m $\times$ 6 m $\times$ 4.5 m (figure A.1). Chaque réflexion, modélisée comme une version atténuée, retardée et filtrée du signal direct, est encodée comme une onde plane au format ambisonique suivant sa direction d'arrivée.
Des coefficients d'absorption sont fixés pour chaque paroi du parallélépipède par bande d'octave, la première centrée à 125 Hz et la dernière centrée à 4 kHz, les valeurs d'absorption étant maintenues constantes pour les fréquences plus élevées (tableau A.1). Il en résulte un jeu de réponses impulsionnelles de salle spatiales (SRIR pour *Spatial Room Impulse Responses*), dépendant de la position de la source et des caractéristiques de la pièce (dimensions, absorption des parois...). Les réponses des quatre premières compo-

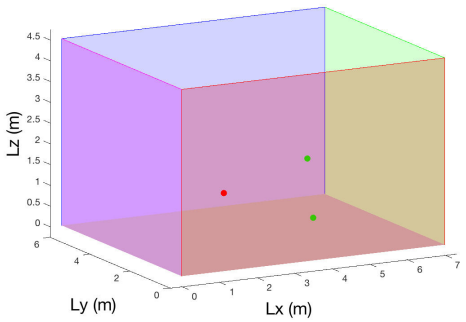

FIGURE A.1. – Aperçu de la boite à chaussures utilisée pour générer un effet de salle par tracé de rayons avec l'outil l'outil RoomSim. En rouge : microphone. En vert : Deux sources, positionnées à 2.3 m du microphone.

	Coefficients d'absorption					
Freq. (Hz)	125	250	500	1000	2000	4000
Sx_0	0.3	0.1	0.05	0.04	0.07	0.1
Sx_{Lx}	0.07	0.3	0.5	0.7	0.7	0.6
Sy_0	0.3	0.1	0.05	0.04	0.07	0.1
Sy_{Ly}	0.7	0.5	0.7	0.7	0.7	0.7
Sz_0	0.4	0.3	0.2	0.2	0.15	0.1
Sz_{Lz}	0.3	0.4	0.8	0.85	0.75	0.7

TABLEAU A.1. – Coefficients d'absorption utilisés pour la simulation d'effet de salle par tracé de rayons. Valeurs par bandes d'octaves. Au-dessus de 4 kHz : valeurs constantes.

santes ambisoniques sont représentées sur la figure A.2. Le TR_{60} est ici de 350 ms. Cette valeur apparaît comme assez faible par rapport au volume de la pièce (189 m^3) et aux coefficients d'absorption appliqués, cela s'explique par un nombre malgré tout limité de réflexions générées et par les limites du modèle utilisé pour la simulation de l'effet de salle.

A.2.1. Création des scènes sonores

La contribution de chaque source s_j au signal multicanal capté, appelée source-image et notée $\mathbf{x}_j$, est calculée en convoluant, pour chaque canal m, la source mono par la réponse impulsionnelle de salle $\mathrm{SRIR}_{j,m}$:

$$x_{j,m}(t) = s_j * \mathrm{SRIR}_{j,m} = \sum_{\tau=0}^{T-1} \mathrm{SRIR}_{j,m}(\tau) \cdot s_j(t-\tau) \tag{A.1}$$

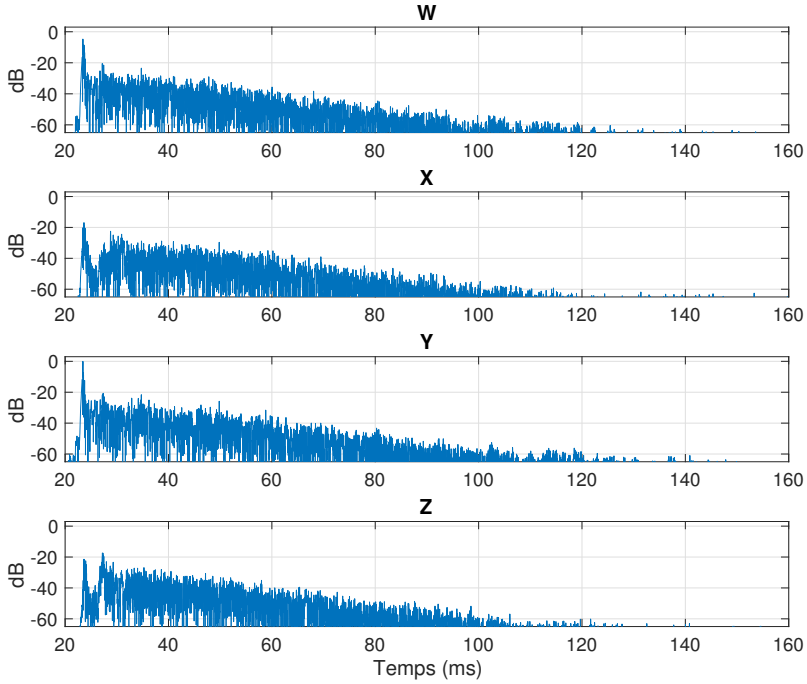

FIGURE A.2. – SRIR ambisonique d'ordre 1 simulée à partir de l'outil RoomSim pour une source positionnée $(90°, 0°, 2\ m)$ par rapport au microphone (amplitude normalisée, en dB)

.

La source-image est donc obtenue par la formule :

$$\mathbf{x}_j(t) = s_j * \mathbf{SRIR}_j = \begin{bmatrix} s_j * \mathrm{SRIR}_{j,1} \\ s_j * \mathrm{SRIR}_{j,2} \\ \dots \\ s_j * \mathrm{SRIR}_{j,M} \end{bmatrix} \tag{A.2}$$

Dans le cas d'un mélange instantané, $\mathrm{SRIR}_{j,m}$ est un dirac dont l'amplitude correspond à la valeur de la $m^{\text{ième}}$ harmonique sphérique dans la direction (θ_j, ϕ_j).
Pour construire une scène $\mathbf{x}$ constituée de N sources sans bruit additionnel, les sources-images sont sommées, soit :

$$\mathbf{x} = \sum_{j=1}^{N} \mathbf{x}_j \tag{A.3}$$

B. Acquisition des réponses impulsionnelles de salles réelles

La création de contenus ambisoniques réels se déroule suivant un procédé analogue au cas synthétique décrit en annexe A, par la génération des contributions ambisoniques de chaque source grâce à des réponses impulsionnelles, puis l'addition des contributions pour générer la scène complète. On décrit ici le processus d'acquisition des réponses impulsionnelles réelles et les différents environnements acoustiques utilisés.

B.1. Environnements acoustiques

Une salle, dont les murs sont équipés de panneaux absorbants amovibles, est exploitée pour générer les réponses impulsionnelles de salles. La pièce possède une forme parallélépipédique de dimensions 5.15 m × 4.08 m × 2.30 m, correspondant à un volume d'approximativement 48 m^3. Deux configurations acoustiques ont été retenues pour la génération des contenus :

— Une salle absorbante (figure B.1) dans laquelle les parois sont intégralement recouvertes de blocs de mousse, dont le TR_{60} est d'environ 120 ms.

— Une configuration de salle réverbérante (figure B.2), pour laquelle les panneaux absorbants sont enlevés dans leur intégralité. Le TR_{60} associé est alors de 350 ms.

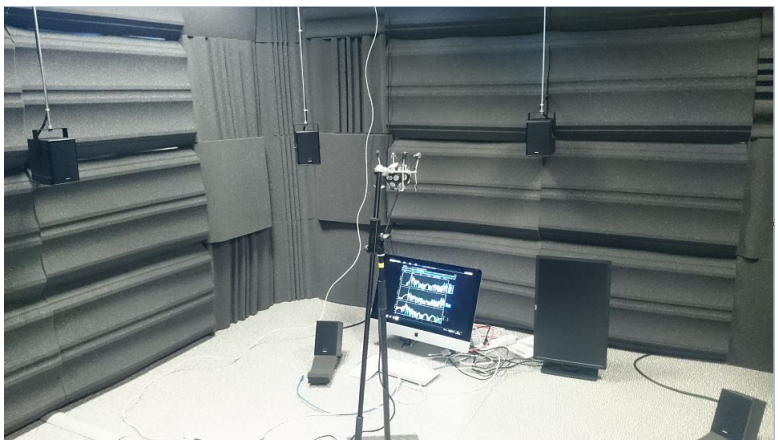

FIGURE B.1. – Salle munie de panneaux absorbants.

Sur la figure B.3, on visualise une réponse impulsionnelle omnidirectionnelle dans les deux configurations de salle, avec dans les deux cas un premier front d'onde aux alentours de

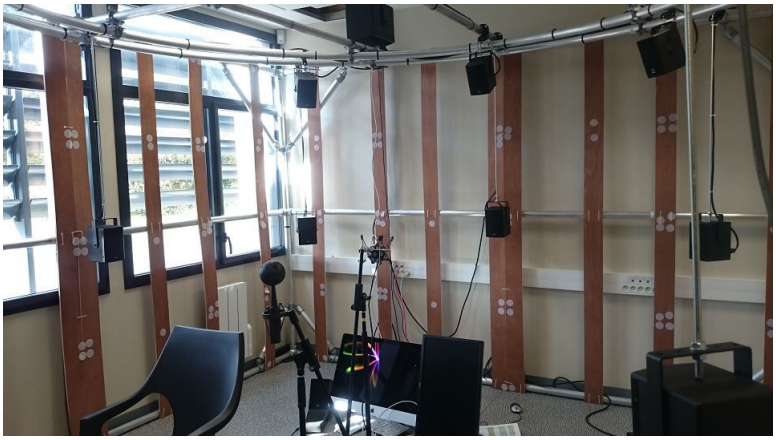

FIGURE B.2. – Salle rectangulaire possédant des parois peu absorbantes.

23 ms, suivi de la première réflexion à 27 ms causée par le sol de la pièce (différence de marche d'environ 1.1 m), suivie d'une décroissante de l'énergie à peu près linéaire jusqu'au niveau de bruit ambiant.

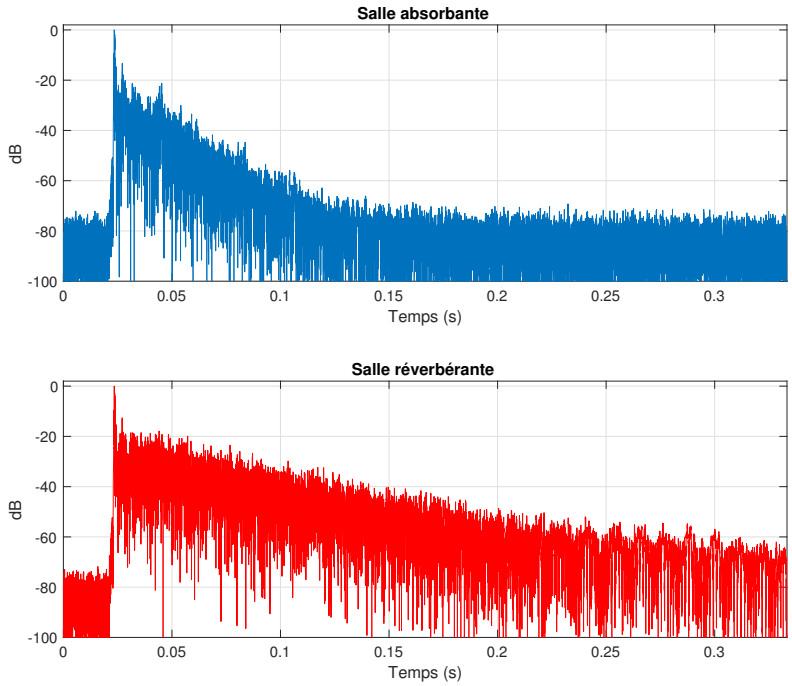

FIGURE B.3. – Réponses impulsionnelles omnidirectionnelles (en dB normalisés). En haut : salle absorbante. En bas : salle réverbérante.

B.2. Acquisitions et génération des réponses impulsionnelles

Le schéma de la figure B.4 représente l'ensemble de la chaîne d'émission/acquisition pour la mesure des SRIR.
Un ensemble de 14 enceintes Amadeus PMX4 sont fixées à une armature métallique circulaire visible sur la figure B.2. Huit d'entre elles sont disposées régulièrement tous les 45° dans ce que l'on considère comme le plan horizontal à une hauteur de 1.10 m tandis que 5 autres sont suspendues à une hauteur de 1.92 m, soit 30° d'élévation par rapport au centre du dispositif horizontal. La dernière enceinte est suspendue à la verticale du dispositif. L'ensemble est alimenté par un jeu d'amplificateurs Lab Gruppen, pilotés par une carte son 32 canaux Antelope Orion 32 (figure B.4) relié à un MacBook Pro. Un dispositif d'égalisation Trinnov est placé entre l'ordinateur et la carte son pour le système d'émission le plus transparent possible acoustiquement. Un microphone 32 capsules Eigenmike identique à celui-ci décrit dans la section 1.2.4 est placé au centre du dispositif pour l'acquisition des mesures. Il est relié au MacBook réalisant simultanément l'émission et l'acquisition à l'aide du logiciel d'édition multi-pistes Reaper.

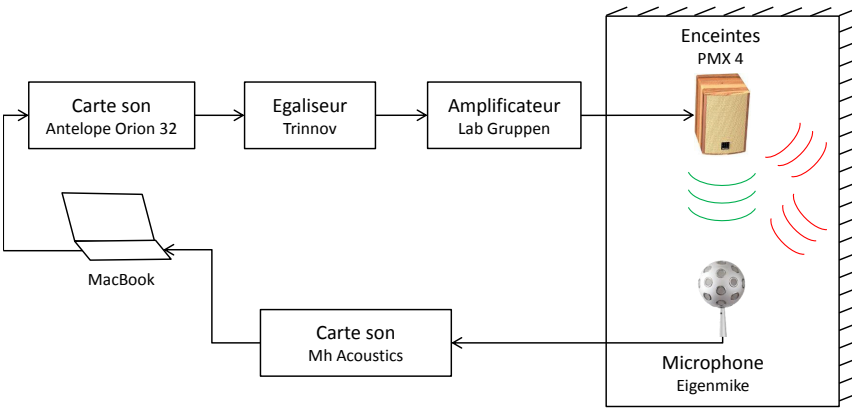

FIGURE B.4. – Chaine de génération/acquisition pour la mesure de réponses impulsionnelles de salles en conditions réelles.

Une fois le dispositif mis en place, des sinus glissants sont émis entre 0.1 et 24 kHz, enregistrés par le microphone et déconvolués par le signal inverse pour de générer les réponses impulsionnelles microphoniques (format A). Une matrice de filtres est appliquée en post-traitement pour convertir les réponses des capsules en réponses ambisoniques à l'ordre 4. La structure métallique permet de déplacer les enceintes sur le pourtour du cercle, aussi les enceintes horizontales ont pu être translatées le long de la structure métallique de façon à mailler le plan horizontal avec une résolution angulaire comprise entre 5° et 10°. Par ailleurs, les mêmes mesures ont été effectuées en déportant le microphone de 0.9 m par rapport au centre de la scène.
En résumé, un ensemble de 46 mesures ont été réalisées, dont 40 dans le plan horizontal,

pour chaque position de microphone (centré/déporté) et pour chaque configuration de salle (absorbante/réverbérante), permettant des configurations géométriques et acoustiques variées. Les scènes complètes sont la résultante de l'addition des contributions de chaque source, obtenues en convoluant un signal de parole avec les SRIR d'un couple de positions source/micro.

C. Modélisation de la distribution des descripteurs

Pour chaque descripteur d_k et chaque classe $\mathcal{C}^\alpha$, une loi de probabilité conditionnelle $p(d_k|\mathcal{C}^\alpha, N, m)$ dépendant du nombre de sources actives et de l'ordre ambisonique est modélisée grâce à une étape d'apprentissage supervisé. Celle-ci consiste à calculer les descripteurs de signaux extraits dont la classe est connue *a priori*. Pour cela, une base d'apprentissage est constituée à partir de contenus ambisoniques réels. Un grand nombre de configurations différentes doivent être évaluées, en termes de nombre de sources et d'environnements acoustiques de façon à estimer les lois de probabilité les plus consistantes et les moins biaisées possibles.

C.1. Contenus sonores utilisés

Le corpus d'apprentissage est constitué de la façon suivante :

— Des scènes sonores sont générées à partir de réponses impulsionnelles et de signaux de voix mono différents de ceux utilisés pour la validation (chapitre 6-, selon le procédé décrit dans l'annexe B.

— Trois configurations de salle sont utilisées : configuration mate avec les murs couverts de panneaux absorbants, configuration semi-réverbérante pour laquelle la moitié des panneaux absorbants est enlevée, et configuration réverbérante sans aucun panneau absorbant.

— 12 scènes contenant entre 0 et 4 sources de voix simultanées sont ainsi générées, à l'ordre 1 et à l'ordre 2. Le cas de figure ne contenant aucune source active correspond aux trames pour lesquelles on dispose seulement d'un champ réverbérant résiduel.

— Pour chaque scène, différents scénarios sont élaborés, en positionnant les sources suivant 12 jeux de positions différentes.

C.2. Procédure mise en place

Le protocole pour générer les lois de probabilité pour chaque descripteur est décrit en figure C.1. Tandis que l'ACI est appliquée trame à trame sur les différents contenus et que les composantes extraites sont permutées par rapport aux positions de référence des sources encodées, une détection d'activité sur les sources mono nous renseigne sur le nombre de sources réellement actives dans le mélange à chaque trame, et donc sur la classe directe ($\mathcal{C}^d$) ou réverbérée ($\mathcal{C}^r$) de chaque composante extraite par ACI. Pour

l'ensemble des trames et des configurations, les descripteurs sont calculés, puis regroupés selon les classes $\mathcal{C}^d$ ou$\mathcal{C}^r$ des composantes dont il dépend.
Par ailleurs, les descripteurs peuvent être calculés sur des trames plus courtes, si l'on souhaite procéder au comptage des sources sur des intervalles de temps plus fins que pour l'ACI.

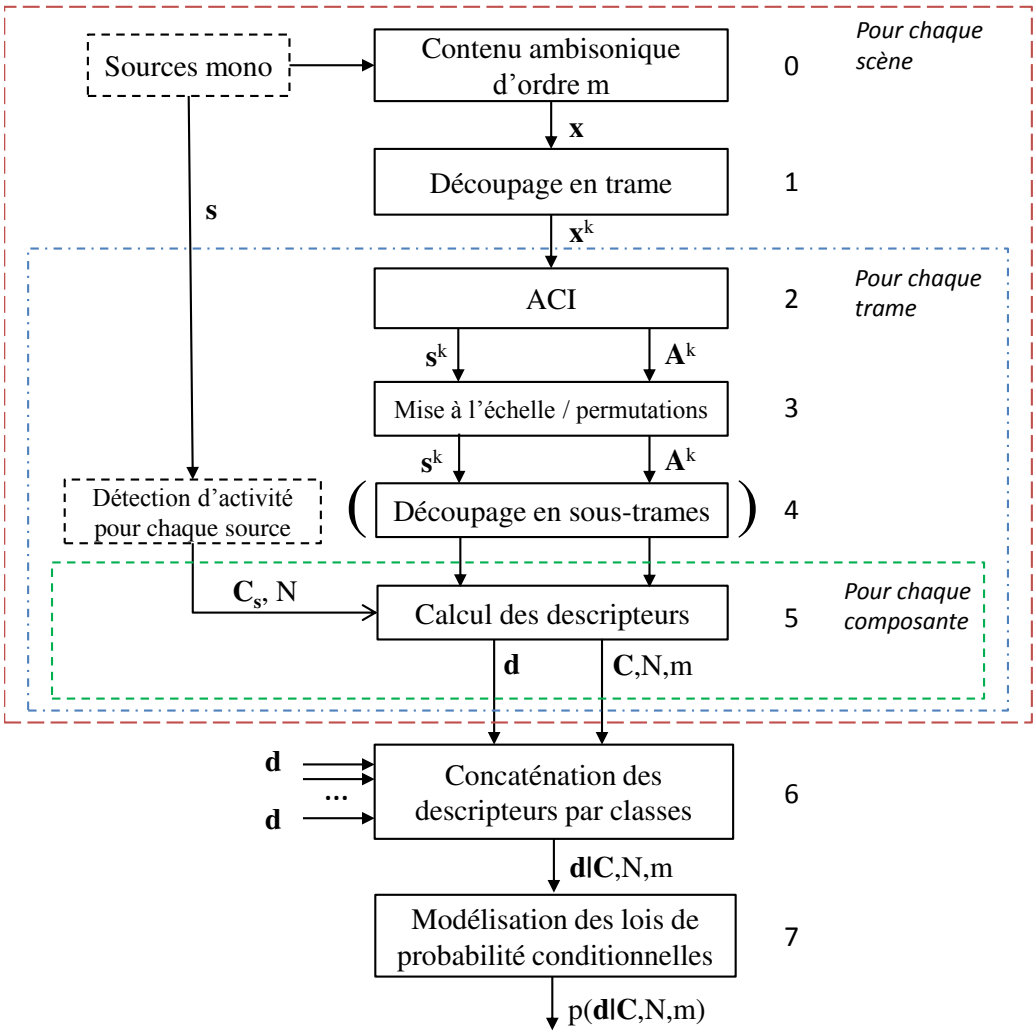

FIGURE C.1. – Génération des lois de probabilités pour la classification bayésienne des composantes extraites par ACI. $\mathbf{d}$: vecteur des descripteur $\mathbf{C}_s$: vecteur indiquant l'activité ou non des sources initiales pour chaque trame de comptage. $\mathbf{C}$: vecteur des classes des composantes extraites (directes ou réverbérées) de dimensions $1 \times M$.

C.3. Modélisation des densités de probabilité

A partir des valeurs du descripteur d_k dont on souhaite modéliser la distribution, un histogramme est généré dont la valeur à l'abscisse i est notée $p[i]$. Celui-ci est normalisé par l'opération

$$p[i] = \frac{p[i]}{\sum_i p[i]} \tag{C.1}$$

afin d'obtenir $\sum_i p[i] = 1$. On a alors une représentation discrète de la densité de probabilité du descripteur, sur laquelle sont projetées un nombre L de lois de probabilités $p^\ell(i)$, parmi lesquelles est sélectionnée la loi minimisant la divergence de Kullback-Leibler (KL) entre $p[i]$ et $p^\ell(i)$, soit :

$$\begin{aligned} \ell &= \arg\min_{1\leq\ell\leq L} \mathrm{D}\big(p[i] \;||\; p^\ell(i)\big) \\ &= \arg\min_{1\leq\ell\leq L} \Big\{ \sum_i p[i] \log \tfrac{p[i]}{p^\ell(i)} \Big\} \end{aligned} \tag{C.2}$$

La figure C.2 illustre l'histogramme normalisé du critère onde plane (voir 5.1.1), pour la classe $\mathcal{C}^d$ pour N=1 dans un contenu ambisonique à l'ordre 2. La loi minimisant la divergence de KL est ici une loi gamma.

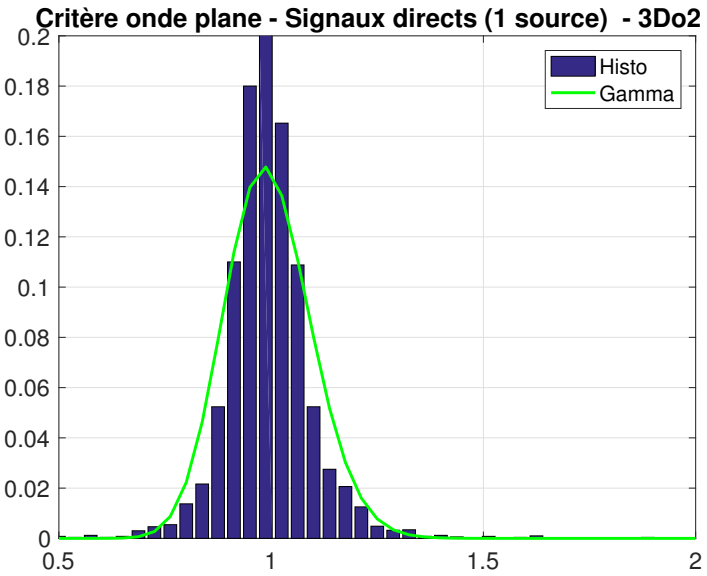

FIGURE C.2. – Histogramme normalisé du critère onde plane (équation 5.4) pour les composantes directes, dans un contenu extrait à l'ordre 2 et contenant une seule source active. En vert : Loi gamma modélisant la densité de probabilité du descripteur.

Bibliographie

[1] RadioFrance. http ://nouvoson.radiofrance.fr. Consulté le 20-09-2016.

[2] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka. A summary of the REVERB challenge : state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016(1), December 2016.

[3] Xmos. http ://www.xmos.com/products/voice/microphones/solutions. Consulté le 13-04-2017.

[4] Michael A. Gerzon. Periphony : With-height sound reproduction. *J. Audio Eng. Soc*, 21(1) :2–10, 1973.

[5] P.G. Craven and M.A. Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs, août 1977. US Patent.

[6] Jérôme Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Thèse de doctorat, Université Paris VI, France, 2000.

[7] 4ever2. http ://www.4ever-2.com/fr/accueil/. Consulté le 13-04-2017.

[8] Svein Berge and N. Barrett. High angular resolution planewave expansion. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, pages 6–7, Paris, 2010.

[9] Ville Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. In *Audio Engineering Society Conference : 28th International Conference : The Future of Audio Technology Surround and Beyond*, Pitea, Suède, 2006. Audio Engineering Society.

[10] Nicolas Epain, Craig Jin, and André Van Schaik. The application of compressive sampling to the analysis and synthesis of spatial sound fields. In *Audio Engineering Society Convention 127*, New-York, Etats-Unis, Oct 2009.

[11] A. Wabnitz, N. Epain, and C.T. Jin. A frequency-domain algorithm to upscale ambisonic sound scenes. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 385–388, Kyoto, Japon, March 2012.

[12] Nicolas Epain, Craig T. Jin, and André van Schaik. Blind source separation using independent component analysis in the spherical harmonic domain. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, 2010.

[13] Philip M Morse and K. Uno Ingard. *Theoretical acoustics*. Princeton University Press, Princeton, N.J., 1986.

[14] Michel Bruneau. *Fundamentals of acoustics*. ISTE Ltd., London ; Newport Beach, CA, 2006.

[15] Catherine Potel and Michel Bruneau. *Acoustique générale : équations différentielles et intégrales, solutions en milieux fluides et solides, applications*. Ellipses, Paris, 2006.

[16] TSL Products. http ://www.tslproducts.com/wp-content/uploads/2013/05/sps-software-mic-thumb.png. Consulté le 24-11-2016.

[17] Darren B. Ward and Thushara D. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on speech and audio processing*, 9(6) :697–707, 2001.

[18] Stéphanie Bertet. *Formats audio 3D hiérarchiques : caractérisation objective et perceptive des systèmes Ambisonics d'ordres supérieurs*. Thèse de doctorat, INSA Lyon, 2006.

[19] M. Chapman, T. Musil, H. Pomberger, W. Ritsch, A. Sontacchi, J. Zmölnig, and F. Zotter. A standard for interchange of ambisonic signal sets. In *Ambisonics Symposium*, Graz, Autriche, 2009.

[20] Sebastien Moreau. *Etude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle distance*. Thèse de doctorat, Université du Maine, le Mans, France, 2006.

[21] A Parthy, C. Jin, and A Van Schaik. Acoustic holography with a concentric rigid and open spherical microphone array. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 2173–2176, Taipei, Taiwan, April 2009.

[22] mh acoustics. https ://mhacoustics.com/products. Consulté le 20-09-2016.

[23] Sébastien Moreau, Jérôme Daniel, and Stephanie Bertet. 3d sound field recording with higher order ambisonics – objective measurements and validation of a 4th order spherical microphone. In *Audio Engineering Society 120th Convention*, Paris, 2006. Audio Engineering Society.

[24] Jérome Daniel, Jean-Bernard Rault, and Jean-Dominique Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*, San Francisco, Sep 1998.

[25] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2001.

[26] Tulay Adali, Matthew Anderson, and Geng-Shen Fu. Diversity in independent component and vector analyses : Identifiability, algorithms, and applications in medical imaging. *IEEE Signal Processing Magazine*, 31(3) :18–33, May 2014.

[27] Alexey Ozerov, Antoine Liutkus, Roland Badeau, and Gael Richard. Coding-Based Informed Source Separation : Nonnegative Tensor Factorization Approach. *IEEE*

Transactions on Audio, Speech, and Language Processing, 21(8) :1699–1712, August 2013.

[28] Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard. An overview of informed audio source separation. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, 2013.

[29] Shoji Makino, Hiroshi Sawada, and Shoko Araki. *Blind Speech Separation*. Springer Netherlands, Dordrecht, 2007.

[30] Nathan Souviraà-Labastie. *Détection de motifs audio pour la séparation de sources guidée. Application aux bandes-son de films.* Thèse de doctorat, Université de Rennes 1, 2015.

[31] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10), June 2016.

[32] J Herault and C Jutten. Space or time adaptive signal processing by neural network models. In *AIP Conference Proceedings 151 on Neural Networks for Computing*, pages 206–211, Woodbury, Etats-Unis, 1987. American Institute of Physics Inc.

[33] O. Shalvi and E. Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). *Information Theory, IEEE Transactions on*, 36(2) :312–321, Mar 1990.

[34] P. Loubaton and P. Regalia. Blind deconvolution of multivariate signals : A deflation approach. In *Communications, 1993. ICC '93 Geneva. Technical Program, Conference Record, IEEE International Conference on*, volume 2, pages 1160–1164 vol.2, May 1993.

[35] Andrew D. Back and Ah Chung Tsoi. Blind deconvolution of signals using a complex recurrent network. In *Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop*, pages 565–574, Ermioni,Grèce, 1994. IEEE.

[36] Gilles Chardon, Laurent Daudet, Antoine Peillot, François Ollivier, Nancy Bertin, and Rémi Gribonval. Nearfield Acoustic Holography using sparsity and compressive sampling principles. *Journal of the Acoustical Society of America*, 2012.

[37] Andrew Wabnitz, Nicolas Epain, Alistair McEwan, and Craig T. Jin. Upscaling ambisonic sound scenes using compressed sensing techniques. In *WASPAA*, pages 1–4, New Paltz, Etats-Unis, 2011. IEEE.

[38] Frédéric Abrard and Yannick Deville. From blind source separation to blind source cancellation in the underdetermined case : A new approach based on time-frequency analysis, 2001.

[39] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

[40] Noboru Murata, Shiro Ikeda, and Andreas Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41 :1–24, 1998.

[41] Alexander Jourjine, Scott Rickard, and Ozgur Yilmaz. Blind separation of disjoint orthogonal signals : Demixing n sources from 2 mixtures. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 5, pages 2985–2988, Istanbul, Turquie, 2000. IEEE.

[42] Simon Bimbot, Frédédic Arberet, and Rémi Gribonval. A robust method to count, locate and separate audio sources in a multichannel underdetermined mixture. 2008.

[43] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Third Edition*. Academic Press, Inc., Orlando, FL, USA, 2006.

[44] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, New York, 1991.

[45] D.R. Brillinger. *Time Series : Data Analysis an Theory*. Holden-Day Series in Time Series Analysis. Holden-Day, 1981.

[46] jean-Louis Lacoume, Pierre-Olivier Amblard, and Pierre Comon. *Statistiques d'ordres supérieurs pour le Traitement du Signal*. MASSON, masson edition, 1997.

[47] John Aldrich. R.a. fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3) :162–176, 09 1997.

[48] Ehud Weinstein, Meir Feder, and Alan V. Oppenheim. Multi-channel signal separation by decorrelation. *Speech and Audio Processing, IEEE Transactions on*, 1(4) :405–413, 1993.

[49] Dinh-Tuan Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *Signal Processing, IEEE Transactions on*, 49(9), 2001.

[50] Hakim Boumaraf. *Blind Source Separation (BSS) of Convolutive Mixtures of Sources*. Theses, Université Joseph-Fourier - Grenoble I ; Institut National Polytechnique de Grenoble - INPG, October 2005.

[51] Arie Yeredor. Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. *Signal Processing Letters, IEEE*, 7(7) :197–200, 2000.

[52] C. Jutten and J. Herault. Independent components analysis versus principal components analysis. In *European Conference on Signal Processing EUSIPCO*, pages 643–646, Grenoble, 1988.

[53] J.-F. Cardoso. Source separation using higher order moments. Ecosse, 1989.

[54] Pierre Comon. Separation of stochastic processes. In *Higher-Order Spectral Analysis, 1989. Workshop on*, pages 174–179, Vail, Etats-Unis, 1989. IEEE.

[55] Aapo Hyvärinen and Erkki Oja. Independent component analysis : algorithms and applications. 2000.

[56] Z. Koldovsky, P. Tichavsky, and E. Oja. Efficient variant of algorithm FastICA for independent component analysis attaining the cramer-rao lower bound. *IEEE Transactions on Neural Networks*, 17(5), September 2006.

[57] Shun-ichi Amari, Andrzej Cichocki, Howard Hua Yang, et al. A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, 1996.

[58] K.E. Hild, H.T. Attias, and S.S. Nagarajan. An expectation maximization method for spatio-temporal blind source separation using an AR-MOG source model. *IEEE Transactions on Neural Networks*, 19(3) :508–519, March 2008.

[59] Geng-Shen Fu, Ronald Phlypo, Matthew Anderson, Xi-Lin Li, and Tulay Adali. Algorithms for markovian source separation by entropy rate minimization. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3248–3252. IEEE, 2013.

[60] Xi-Lin Li and T. Adali. A novel entropy estimator and its application to ICA. In *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pages 1–6, Grenoble, 2009. IEEE.

[61] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4) :620–630, 1957.

[62] Xi-Lin Li and Tülay Adali. Blind spatiotemporal separation of second and/or higher-order correlated sources by entropy rate minimization. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1934–1937, Dallas, Etats-Unis, 2010. IEEE.

[63] Taesu Kim, Torbjørn Eltoft, and Te-Won Lee. Independent vector analysis : An extension of ica to multivariate components. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 165–172. Springer Berlin Heidelberg, 2006.

[64] Matthew Anderson, Geng-Shen Fu, Ronald Phlypo, and Tulay Adali. Independent Vector Analysis : Identification Conditions and Performance Bounds. *IEEE Transactions on Signal Processing*, 62(17) :4399–4410, September 2014.

[65] Xi-Lin Li, Tülay Adali, and Matthew Anderson. Joint blind source separation by generalized joint diagonalization of cumulant matrices. *Signal Process.*, 91(10) :2314–2322, October 2011.

[66] Koby Todros and Alfred O. Hero. On measure transformed canonical correlation analysis. *Signal Processing, IEEE Transactions on*, 60(9) :4570–4585, 2012.

[67] Xi-Lin Li, Matthew Anderson, and Tülay Adali. Second and higher-order correlation analysis of multiple multidimensional variables by joint diagonalization. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pages 197–204. Springer Berlin Heidelberg, 2010.

[68] Juha Merimaa and Ville Pulkki. Spatial impulse response rendering. *Proc. of the 7th Intl. Conf. on Digital Audio Effects (DAFX'04), Naples, Italy*, 2004.

[69] Christof Faller and Ville Pulkki. Directional audio coding : Filterbank and STFT-based design. In *Audio Engineering Society Convention 120*, Paris, 2006. Audio Engineering Society.

[70] Julian D. Palacino and Rozenn Nicol. Spatial sound pick-up with a low number of microphones. In *Proceedings of Meetings on Acoustics*, volume 19, pages 55–78, Indianapolis, Etats-Unis, 2013. Acoustical Society of America.

[71] Julian D. Palacino and Rozenn Nicol. Full 3d sound pick-up with a small microphone array : Prototype outline and preliminary assessment. In *Proc. of Conference on Acoustics AIA-DAGA*, Merano, Italie, 2013.

[72] Svein Berge and Natasha Barrett. A new method for b-format to binaural transcoding. In *Proc. 40th Intl. Conf. Audio Eng. Soc*, Tokyo, Japon, 2010.

[73] Svein Berge. Brevet HARPEX USA.

[74] P.K.T. Wu, N. Epain, and C. Jin. A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4053–4056, March 2012.

[75] P.K.T. Wu, N. Epain, and C. Jin. A super-resolution beamforming algorithm for spherical microphone arrays using a compressed sensing approach. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 649–653, Vancouver, Canada, May 2013.

[76] Nicolas Epain and Craig T. Jin. Independent component analysis using spherical microphone arrays. *Acta Acustica united with Acustica*, 98(1), 2012-01-01.

[77] Malham. Experience with large area 3-d ambisonic sound systems. volume 14, page 209–215, 1992.

[78] Boaz Rafaely. *Fundamentals of Spherical Array Processing*, volume 8 of *Springer Topics in Signal Processing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[79] SiSEC 2008. http ://sisec2008.wiki.irisa.fr/tiki-index.html. Consulté le 29-11-2016.

[80] Emmanuel Vincent, Shoko Araki, and Pau Bofill. The 2008 signal separation evaluation campaign : A community-based approach to large-scale evaluation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 734–741. Springer, 2009.

[81] D. R. Campbell, K. J. Palomäki, and G. Brown. A MATLAB simulation of "shoebox" room acoustics for use in research and teaching. *Computing and Information Systems Journal, ISSN 1352-9404*, 9, 2005.

[82] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4) :943–950, 1979.

[83] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4) :1462–1469, 2006.

[84] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent. BSS_eval toolbox user guide–Revision 2.0. 2005.

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

[86] Lukas Drude, Aleksej Chinaev, Dang Hai Tran Vu, and Reinhold Haeb-Umbach. Source counting in speech mixtures using a variational EM approach for complex Watson mixture models. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6834–6838. IEEE, 2014.

[87] Oliver Walter, Lukas Drude, and Reinhold Haeb-Umbach. Source counting in speech mixtures by nonparametric Bayesian estimation of an infinite Gaussian mixture model. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 459–463. IEEE, 2015.

[88] Jalil Taghia, Nasser Mohammadiha, and Arne Leijon. A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 253–256. IEEE, 2012.

[89] Kaldi. http ://kaldi-asr.org/doc/index.html. Consulté le 18-04-2017.

[90] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge university engineering department*, 3 :175, 2002.

[91] Romain Serizel, Marc Moonen, Bas Van Dijk, and Jan Wouters. Rank-1 approximation based multichannel wiener filtering algorithms for noise reduction in cochlear implants. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8634–8638. IEEE, 2013.

[92] Ngoc Q K Duong, Emmanuel Vincent, and Rémi Gribonval. Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1830–1840, September 2010.

[93] Union International des Télécommunications. P.800 - méthodes d'évaluation subjective de la qualité de transmission. 1996.