Décomposition de scènes sonores ambisoniques pour navigation en six degrés de liberté

Mohammed Hafsati

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

COMUE UNIVERSITÉ BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication

Spécialité : Signal, Image et Vision

Par

Mohammed HAFSATI

Higher Order Ambisonic sound scenes decomposition for six degree of freedom navigation

Thèse présentée et soutenue à Lieu, le November 23, 2020
Unité de recherche :
Thèse N°

Rapporteurs avant soutenance :

Rozenn NICOL Ingénieur de recherche, Orange Labs, Lannion
Laurent GIRIN Professeur, Grenoble INP

Composition du jury :

Président :	Prénom NOM	Fonction et établissement
Examinateurs :	Sylvain MARCHAND	Professeur, IUT de la Rochelle
	Laurent ALBERA	Maitre de conférences, Université de Rennes 1
Directeur :	Rémi GRIBONVAL	Directeur de recherche, ENS de Lyon
Co-directeur :	Nicolas EPAIN	Ingénieur de recherche, b<>com, Cesson sévigné
Co-encadrante (Invitée) :	Nancy BERTIN	Chargée de recherche, Université de Rennes 1

UNIVERSITE BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE RENNES 1

Intitulé de la thèse :

Décomposition de scènes sonores ambisoniques pour navigation en six degrés de liberté

-

Higher Order Ambisonic sound scenes decomposition for six degree of freedom navigation

Acknowledgements

First and foremost, I am incredibly grateful to my supervisors, Dr. Nancy BERTIN, Dr. Rémi GRIBONVAL and Dr. Nicolas EPAIN for their advice, continuous support, and extremely admirable patience during my Ph.D. study. Their immense knowledge and ample experience have encouraged me throughout my academic research and professional life. I want to thank all the audio team members in b-com, the entire PANAMA team in INRIA, and Mr. Jean-yves for being an incredible head of the AMC departement at b-com.

I want to express my deepest gratitude for both of my PH.D. referees, Dr. Laurent GIRIN and Dr. Rozenn NICOL, who agreed to read my manuscript and helped me improve it with their essential remarks and comments.

On a personal level, I'm very much thankful for my parents' support, my mother for being tremendous emotional support throughout my entire Ph.D. study, and my father for being important financial support at the end of my Ph.D. study. Finally, I am extremely grateful to my girlfriend for her help, patience, unconditional love, and always finding the right words to fire up my motivation during my downs and doubtful moments.

Résumé en Français

Ce résumé présente de manière concise les différents travaux abordés dans cette thèse. Les détails techniques concernant les outils utilisés et les méthodes proposées sont donnés dans la suite du manuscrit (en anglais).

Cette thèse s'inscrit dans le contexte multimedia dont le sujet technique est la navigation dans des champs sonores 3D. Contrairement aux contenus de réalité virtuelle, notre application vise les contenus issus de captations réelles. La technologie d'audio 3D (trois dimensions) choisie pour mener à bien nos recherches est la technologie ambisonique. Au moment de la captation, nous utiliserons une seule antenne ambisonique ainsi que ponctuellement des microphones d'appoint. Grâce à la technologie ambisonique, il est possible de capter un champ sonore, de le représenter dans un format dit pivot[1], et de le restituer en 3D à l'aide d'un dispositif d'écoute quelconque. Ces avantages vont être utiles pour notre application, afin d'avoir notamment la possibilité d'adapter le champ sonore par rapport aux pivotements de tête de l'utilisateur, permettant d'obtenir déjà trois degrés de liberté. Le problème apporté par ce type de représentation de champ sonore réside dans la difficulté d'avoir 6 degrés de liberté, avec la possibilité de changer de point de vue. Afin de contourner ce problème, nous recommandons de faire une décomposition du format ambisonique en ondes planes. Cela a été déjà proposé dans plusieurs contributions dans l'état de l'art en utilisant des techniques de formation de voies en pleine bande. La particularité d'une de nos méthodes est d'utiliser des techniques de séparations de sources sonores multicanale, avec lesquelles nous cherchons les contributions de chaque source dans chaque canal ambisonique. Cela n'a jamais été utilisé auparavant pour faire de la navigation dans des contenus ambisoniques. Nous avons fait une comparaison objective et nous avons obtenu de meilleurs résultats avec notre méthode. Nos méthodes dépendent énormément de la précision avec laquelle nous connaissons la direction d'arrivée des sources sonores. Pour cela nous avons établi un état de l'art général sur la localisation des sources dans les deux domaines à la fois microphonique et ambisonique, en adaptant à chaque fois les méthodes microphoniques. Nous avons validé ces techniques avec des simulations en ayant des résultats satisfaisants et comparables en ordre de grandeur à l'état de l'art. Pour la décomposition

[1] Un format permettant de tourner le champ avec une simple multiplication matricielle.

en ondes planes, il existe plusieurs techniques de formation de voies. Dans ce cadre, nous avons proposé 2 méthodes qui profitent de l'avantage du nombre de canaux ambisoniques. Nous avons comparé toutes les méthodes de formation de voies entre elles et nous avons obtenu de meilleurs résultats avec une des méthodes proposées, que nous recommanderons d'utiliser, si le nombre de sources est inférieur au nombre de canaux ambisoniques. La deuxième partie de nos recherches se concentre sur la séparation de sources multicanale dans le domaine ambisonique. Aux prémices de nos recherches, l'état de l'art manquait de contributions. L'idée est d'utiliser un filtre de Wiener multicanal. Cette catégorie de filtre est réputée pour être l'un des meilleurs filtres pour faire de la séparation de sources. Le problème réside dans la difficulté de chercher les coefficients de ce filtre. Dans le domaine microphonique, il existe une méthode dite séparation de sources multicanale en se basant sur le modèle gaussien local. Avec cette méthode, le filtre de Wiener se simplifie en faisant l'hypothèse que les contributions des sources sonores dans les microphones suivent une loi normale centrée. Cette méthode reste à nos jours une des meilleures méthodes de l'état de l'art. Nous avons dérivé les équations dans le domaine ambisonique en vérifiant que la méthode reste applicable. Nous avons aussi validé cela avec des simulations. Par la suite, en se basant toujours sur le modèle gaussien local, nous avons proposé d'utiliser des microphones d'appoint placés près des sources sonores ainsi qu'un algorithme, afin de guider la séparation de sources multicanale. Nous avons validé cette approche avec des simulations. Enfin, nous avons proposé de remplacer les microphones d'appoint et la méthode proposée avec des réseaux de neurones dans le cas des contenus musicaux contenant une voix et trois instruments, lesdits instruments étant une basse, une batterie, et un instrument quelconque. Nous avons validé cela avec des simulations. Dans la suite de ce résumé nous donnons plus de détails sur l'approche adoptée pour naviguer, ainsi que quelques contributions qu'ont marqué cette thèse.

1 Approche de navigation

La navigation avec l'approche que nous proposons repose sur la décomposition de la scène ambisonique à la position initiale (position de captation par l'antenne ambisonique) en objets audio. Ces derniers sont manipulés et utilisés pour recombiner la scène ambisonique à la position actuelle de l'utilisateur. Notre approche est illustrée dans la fig. 4.2. Notre approche est applicable sur des contenus ambisoniques captés et encodés.

Dans un premier temps, nous cherchons la direction d'arrivée de chaque source sonore ainsi que son signal. Ensuite nous manipulons le signal de chaque source sonore en fonction de la position de l'utilisateur afin d'appliquer une translation du point de vue. Cela est expliqué avec plus de détails avec des équations (de Eq. (4.2). à

Eq. (4.5)). Dans le cas où l'utilisateur tourne sa tête, une rotation est appliquée en utilisant Eq. (2.21).

Dans le cadre du schéma proposé dans la Fig. 4.2, nous proposons deux différentes manières pour décomposer la scène ambisonique en objets sonores :

- Une première décomposition qui repose sur une application de beamforming. Cette approche a été déjà proposée dans l'état de l'art dans plusieurs contributions. Contrairement à ce que l'état de l'art propose d'utiliser comme type de beamforming, nous recommandons d'utiliser un type de beamforming (PIV régularisée) qui profite de l'avantage du nombre de canaux offert par les mélanges ambisoniques. En effet, cela a été déduit d'une simulation dans laquelle nous avons comparé plusieurs types de beamforming et prouvé qu'avec la PIV régularisée nous obtenons les meilleures performances de séparation.

- Une deuxième décomposition qui repose dans un premier temps sur la recherche des contributions des sources sonores dans chaque canal suivi par une décomposition en utilisant des beamforming. La première opération est connue dans la littérature par la séparation multicanale. L'avantage d'une telle décomposition est le fait qu'elle repose à la fois sur des indices spectraux et spatiaux contrairement aux décompositions en beamforming qui en l'occurrence repose que sur des indices spatiaux. Un exemple de navigation avec cette approche est donné dans ma page web personnelle : **`https://hafsatimohammed.github.io/HTML_Files/Example_Navigation.html`**

Nous avons comparé les deux approches avec une des meilleures méthodes de navigation de l'état de l'art en utilisant une mesure objective. Afin de réaliser cela, nous avons considéré que les directions arrivées des sources étaient connues ainsi que les contributions des sources dans chaque canal (considération d'une séparation multicanale parfaite). Nous avons choisi d'utiliser le MOS_LQO score comme mesure objective. Cette mesure permet de quantifier la similarité entre deux contenus binauraux. Les résultats ont prouvé que la méthode qui repose sur la séparation multicanale donne de meilleurs résultats. Cela nous permet de valider le fonctionnement de notre approche, ainsi que de concentrer nos recherches sur les briques de localisation et de séparation nécessaires pour amener à bien notre algorithme.

2 Localisation de sources sonores dans le domaine ambisonique

La connaissance des directions d'arrivées des sources sonores est importante pour notre algorithme de navigation. Pour cela nous avons étudié différentes méthodes de localisa-

tion de sources sonores dans les deux domaines à la fois ambisonique et microphonique, tout en adaptant les méthodes du domaine microphonique au domaine ambisonique. Nous avons comparé ces méthodes entres elles avec des simulations et nous avons conclu sur la suffisance de certaines méthodes pour avoir des performances de localisation assez satisfaisantes pour notre application. Nous avons évalué les performances de localisation en calculant la précision et l'erreur moyenne sur plusieurs exemples pour des tolérances angulaires en azimut et en élévation qui sont égales soit à 5°, 10°, 15°. Avec Direction Estimation of Mixing Matrix DEMIX nous avons obtenu les meilleurs résultats avec une précision d'au moins 73% pour des scènes complexes avec 3 sources sonores et un temps de réverbération $RT_{60} = 0.7s$. Nous avons remarqué que les ordres de grandeur que nous avons obtenues restent en concurrence avec les méthodes de localisation récentes de l'état de l'art.

3 Séparation de sources multicanale dans le domaine ambisonique

Dans le cadre de la séparation de sources multicanale dans le domaine ambisonique nous avons investigué si le modèle Gaussien local pourrait être applicable aux mélanges ambisoniques. Pour cela nous avons dérivé les équations et prouvé que le formalisme mathématique reste le même que dans le domaine microphonique, et nous avons validé le fonctionnement de l'approche en comparant les résultats des performances dans le domaine ambisonique par les résultats des performances dans le domaine microphonique.

Nous avons proposé d'améliorer les performances de séparation avec le modèle Gaussien local en énonçant de rajouter des microphones d'appoint placées près des sources sonores. Nous avons proposé aussi un algorithme pour traiter l'information supplémentaire donnée par le microphone d'appoint. Nous avons validé le principe de l'approche avec des simulations numériques ainsi que l'amélioration en comparant l'approche à celle d'une application systématique du modèle Gaussien local.

Enfin nous avons proposé de remplacer les microphones d'appoint avec des réseaux de neurones. Pour preuve de concept, nous avons choisi d'étudier des contenus ambisoniques musicaux où nous cherchons les contributions des instruments dans chaque canal. Nous avons proposé 2 différentes architectures que nous avons entrainées et testées avec des simulations. A la fin nous avons comparé toutes les méthodes étudiées entre elles et nous avons conclu que la meilleure méthode en termes de performances reste la méthode avec les microphones d'appoint. L'utilisation des réseaux de neurones pour la séparation des sources sonores peut être assez performante, cependant l'utilisation reste assez contrainte aux nombres de sources sonores leurs types, leurs environnement, *etc.*

4 Conclusion

Dans le cadre de cette thèse, nous avons proposé des algorithmes qui permettent à un utilisateur de naviguer virtuellement dans des scènes sonores ambisoniques captées. Ces scènes pourraient être captées par une seule antenne ambisonique et occasionnellement nous rajouterons des microphones d'appoint pour une des méthodes proposées. Avec cette thèse nous avons pu déduire les contributions suivantes :

- Nous avons adapté des méthodes de localisation de sources sonores du domaine microphonique au domaine ambisonique. Avec ses méthodes adaptées nous avons obtenu de bonnes performances qui restent dans le même ordre de grandeur que des approches sophistiquées de l'état de l'art.
- Nous avons proposé de naviguer dans des champs sonores ambisoniques captés de deux différentes manières reposant sur la décomposition des scènes sonores. Nous avons validé objectivement les deux approches.
- Nous avons vérifié et validé la possibilité d'appliquer le modèle Gaussien local pour faire de la séparation de sources multicanale dans le domaine ambisonique.
- Nous avons proposé d'ajouter des microphones d'appoint pour améliorer la séparation de sources multicanale qui est basée sur le modèle Gaussien local.
- Nous avons proposé de remplacer les microphones d'appoint avec des réseaux de neurones pour guider la séparation de sources multicanale dans le cadre de contenus musicaux.

Deux exemples de separation de sources multicanales avec les approches proposées sont donnés dans les pages suivantes **`https://hafsatimohammed.github.io/HTML_Files/Example1.html`** et **`https://hafsatimohammed.github.io/HTML_Files/Example2.html`**.

Table of contents

List of figures

List of tables

Glossary

3-DoF	Three-Degrees-of-Freedom
3D	three-dimensional
6-DoF	Six-Degrees-of-Freedom
DirAC	Directional Audio Coding
DoA	Direction of Arrival
DoAs	Direction of Arrivals
DUET	Degenerate Unmixing Estimation Technique
EM	Expectation Maximization
FASST	Flexible Audio Source Separation Toolbox
HARPEX	High Angular Resolution Plane Wave Expansion
HOA	Higher Order Ambisonics
HRTF	Head Related Transfer Function
ILD	Ineraural Level Difference
IM	Ideal ratio Mask
ITD	Ineraural Time Difference
LCMV	Linearly-Constrained Minimum Variance
LGM	Local Gaussian model
MENUET	Multiple sENsor dUET

NMF None-negative Matrix Factorization

OBM Oracle Binary Mask
OSM Oracle Soft Mask

PCA Principal Component Analysis
PIV Pseudo-InVerse
PSD Power Spectral Densities

SAR Signal to Artifacts Ratio
SDR Signal to Distortion Ratio
SIR Signal to Interference Ratio
SMA Spherical microphone array
STFT Short-time Fourier transform
SV Steering Vector

TF Time-Frequency

VAS Virtual Auditory Spaces
ViSQOL Virtual Speech Quality Objective Listener
VR Virtual reality

WFS Wave Field Synthesis

Chapter 1

Introduction

1.1 Context and motivation

It would surely be interesting to visit new places from the comfort of one's home. Such an application might be possible with the arrival of the new forms of immersive media. Nowadays, users in their daily basis are getting more engaged with the new types of immersive media for entertainment purposes. For instance, among the most played video games today are the ones that propose a total immersion in the displayed environment such as Fortnite.[1] Another example that showcases the interest around immersive media is the enthusiasm around 360-video over the last decade. Various social media companies have quickly adopted 360-video format in their platforms, and they even encouraged their users to create their own. Moreover, the users have been appreciating the concept and found creative ways to share their videos. Beside entertainment uses, these new forms of immersive media have other exciting applications such as healthcare, education, sport, telecommunication, geolocation, *etc.*

Furthermore, the feeling of immersion in most of these applications can be remarkably improved if the content has three-dimensional (3D) audio. The user hears this 3D sound through a listening device and could perceive and locate the sound sources spatially as if they were physically present around him/her. Indeed three-dimensional sound would help the immersion by describing each visual event with a sound coming from its direction. Thus, the user would have a greater immersion in the displayed environment with the help of his/her two senses: sight and hearing.

When it comes to immerging the user in a recorded environment, the movements of the user are limited. For instance, with 360-video we can display an existing en-

[1]This game was considered one of the most played video games of 2018. In this game a new feature was added and was much appreciated, which locates enemies in the environment using 3D audio `https://www.vg247.com/2019/10/16/fortnite-new-3d-headphones-feature-makes-locating-enemies-much-easier/`.

vironment, and when the user is immerged in this kind of videos, he/she can only have three degrees of freedom (3-DoF) by the adaptation of the visuals and the audio to his/her head orientation. It would be more interesting to give the users complete flexibility with six degrees of freedom (6-DoF) in recorded environments similarly to video gaming and VR content.

On the one hand, in VR content and video gaming, the 3D audio is synthetic and not recorded, which is the exact reason that makes it easy to navigate with 6-DoF. Arguably, in such cases, the sound source signals are known. On the other hand, when it comes to recorded sound fields, it is challenging to navigate with 6-DoF. In such cases, the sound source signals are typically not known.

The objective of my work is to provide 6-DoF navigation in 3D sound fields that were acquired from a live recording. In other terms, we would like to have a similar experience as in VR with the ability to change the point of view as illustrated in Fig. 1.1, which will give the user 6-DoF as in VR content.

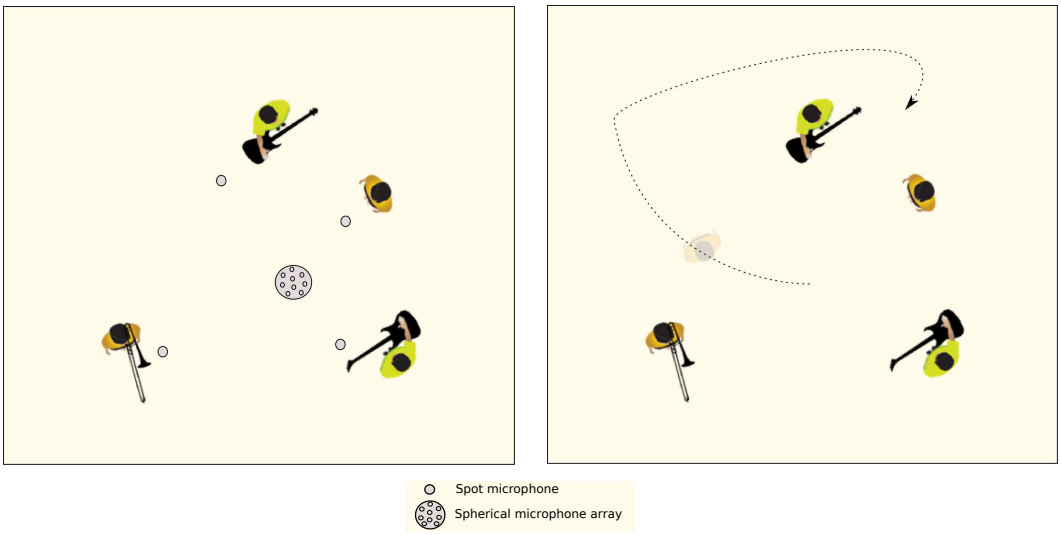


Fig. 1.1 Virtual navigation in a recorded sound scene. In the left a sound scene is recorded with some microphones. In the right a user is navigating virtually in the environment during the playback.

In order to carry out my Ph.D. work, we chose to work on a specific 3D audio technology called Higher Order Ambisonics (HOA). The reasons behind this choice will appear more clearly after discussing the possible technologies in Section 1.2. The motivation behind the selected technology is given in Section 1.3, as well as the potential challenges related to the main objective of my subject. In Section 1.4 I expose my contributions along the manuscript structure.

1.2 Three-dimensional audio technologies

Before surveying the different 3D audio technologies and pointing out their differences, we first explain how humans perceive sound.

A sound is a mechanical vibration (wave) that travels through a medium such as air. A person perceives a sound thanks to his ears that capture these vibrations. The waves vibrate the tympanic membrane, which results in the vibration of the middle ear's three bones. The waves are then transformed in the inner ear into electrical signals that travel to the brain. In an environment where different sources contribute in one sound field, a person with a healthy auditory system perceives the sound field with the ability to differentiate the sounds, locate each sound direction [52, 53] and have a rough idea of their distances[2] when they are in near field [28, 21, 73]. The auditory system of mammalians has been studied extensively. It appeared that their ability to distinguish a sound and its location in direction and distance (roughly) depends on several cues such as spectral information, correlation, and time and level differences between the ears. Indeed, the inter-aural time difference (ITD), and inter-aural level difference (ILD) are important cues that help humans with normal hearing to locate sound sources in the horizontal plane [60, 70] and play a critical role in speech recognition in complex listening environments [16, 112, 25]. In Fig 1.2 we illustrate how the ILD and ITD are identified in the case of a sound source that is closer to the right ear of the listener.

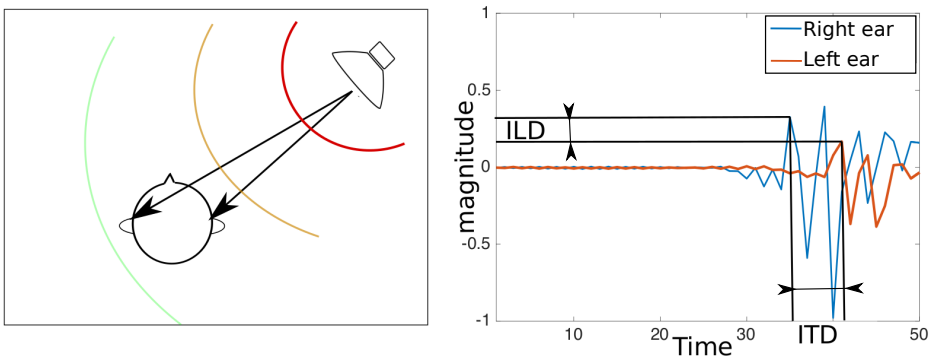


Fig. 1.2 Inter-aural level difference and inter-aural time difference for a sound source coming from the right. The sound source signal is perceived first and louder by the right ear ($ITD > 0, ILD > 0$).

"Spatial audio" or "three-dimensional audio" refers to an ensemble of techniques that allow the listener to perceive virtually a sound field in three dimensions as in real life. By this, we mean that at the time of the listening, the perceived sources

[2]The capacities of humans to distinguish sound distance is until nowadays, still doubtful. However, some studies proved that some sound cues vary with distance, which provides the listener with a possible basis for distance discrimination.

respect the localization cues of the auditory system. At the same time, these sources do not exist physically. However, they may have existed before if the sound scene was recorded. These virtual sound scenes are sometimes referred to as Virtual Auditory Spaces (VAS). Either they are synthesized, or created from recorded sound fields. In other words, one creates VAS by stimulating the human auditory system, which is done by processing the acoustical signals of the recorded/synthesized sound scenes (before playback). A VAS is played through a playback system such as a distribution of loudspeakers or headphones. There are several 3D audio technologies. They differ by their way to deal with the successive steps:

- The way the sound field is recorded. The recording process includes the used type of microphones and how they must be positioned in the sound field.
- The format in which we represent the sound field.
- The way the sound field is played back. The playback process includes the type of listening device and the way the device signals were derived from the representation format.

We give a brief survey of the main existing 3D sound technologies [84] in the following. We begin by describing their principles, followed by Table 1.1, in which we describe how these technologies handle the successive steps described above (recording system, representation format and rendering). The considered technologies are:

- Stereo [107]: This technology is based on the lateral time-spatial cues of the auditory system ILD and ITD. The perception of the sound sources is whether on the right or the left of the user. It requires to extract the difference of time and intensity between two points of the sound field.
- Multichannel surround systems X.Y (5.1, 6.1, 7.1, 10.2, 22.2, *etc.*)[110]: This technology is an extension of the stereo system. It requires to add more channels.
- Ambisonics and Higher Order Ambisonics (HOA) [41, 71, 29]: This technology is based on decomposing the sound field on the spherical harmonics basis. More details about this technology are given in Chapter 2.
- Binaural [77, 1]: This technology is based on mimicking the auditory system. The spatialization technique is binaural synthesis, which requires applying the filters that reproduce the transfer function between the sources and the ears, known as the head-related transfer function (HRTF).
- Wave field synthesis (WFS) [13, 129, 111, 17] : This technology is based on a decomposition of an acoustic wave into several "wavefronts" [14].

- Object-based audio [18]: This technology is based on describing the sound field in terms of sound objects which are sound source signals accompanied by metadata.[3]

Technology	Recording system	Representation format	Rendering
Stereo	Two microphones.	Two channels.	Two loudspeakers forming an equilateral triangle with the user.
Multichannel surround systems X.Y (5.1, 6.1, 7.1, 10.2, 22.2, etc.)	Multichannel trees such as INA 5, Fukada-Tree,OCT-Surround, IRT-Cross, Hamasaki-Square [113].	X+Y channels.	X loudspeakers and Y band-limited Low Frequency Effects (LFE)
Ambisonics and Higher Order Ambisonics (HOA)	Spherical microphone array (SMA).	$(L+1)^2$ channels.	Any loudspeaker distribution or headphones.
Binaural	Two microphones that are placed on the ears of a dummy head.	Two channels.	Headphones.
Wave field synthesis (WFS)	The recording process is in theory done by a set of extended microphones. [4]	Depends on the used microphone array. The number of channels equals the number of microphones.	Extended network of loudspeakers, which their signals are directly given by the microphone signals. Therefore, each microphone is replaced by a loudspeaker.
Object-based audio	Each sound source is recorded separately with a spot microphone.	Given the sound object, one can represent the sound field in any format, which makes this technology compatible with all the above-listed format.	Any loudspeaker distribution or headphones.

Table 1.1 3D sound technologies.

[3] A side information about the sound sources such as their position in the sound field.

[4] There are some WFS techniques to record sound fields. However, the recording requires a huge amount of microphones set linearly. For instance, in [62] the authors use a linear microphone array of 32 microphones set at intervals of 12 cm.

1.3 Scientific challenges and goals: navigation in three-dimensional recordings

To navigate in 3D sound fields with 6-DoF, one must have a representation of the sound field where the sound source signals are separated from each other. Such a representation is natural to obtain when the audio scene is synthetic. It is the case when it comes to pure VR, such as in video games. When it comes to recorded sound scenes with microphone arrays, it is challenging to navigate with 6-DoF, because the outcome of the recording is usually a panoramic description of the sound field at a specific point of the space such as with the ambisonic format.

There is a 3D sound technology with which it would be easy to navigate with 6-DoF. Indeed, since the sound sources signals are given with object-based audio, we can obtain the sound field at any representation format and in any position of the space. However, the main problem to use such a technology would be its time and resource-consumption when it comes to the recording process. As can be imagined, one must record each sound source signal separately as well as the room impulse responses corresponding to a recording at every point of space if the user wants an accurate representation of the ambiance at every point of space while navigating. The last operation is not possible. We can imagine several ways to overcome the physical limits of this operation. For instance, one can sample the environment, and have a limited number of points, and then somehow interpolate the representation of the reverberation to estimate the ambience at any point while navigating. A more realistic approach to handle reflections is using a reverberation model that is quite similar to the environment ambience using simulations. However, even with these practical approximations, it is still demanding in terms of time and resources to record each sound source signal separately and collect their metadata.

In contrast with object-based audio, with ambisonics, the recording process is more straightforward because one can record live (at the same time) the entire sound field at a single point. Note that with ambisonics, we can already perform 3-DoF navigation, with the ability to adapt the sound field to the user's head orientation. However, it is still difficult to perform free navigation with 6-DoF. To have complete navigation, the user must have the ability to change its position from the recording position to another one (translation).

When it comes to a translation unlike a rotation, each source will require a different process. Let us consider the sound field showcased in Fig. 1.4, the user's movement from point A to point B, would require to change the direction of arrivals and the magnitude of the guitar and the drums signals differently. To understand the challenge, let us imagine how the sources would be perceived at point A and point B and compare.

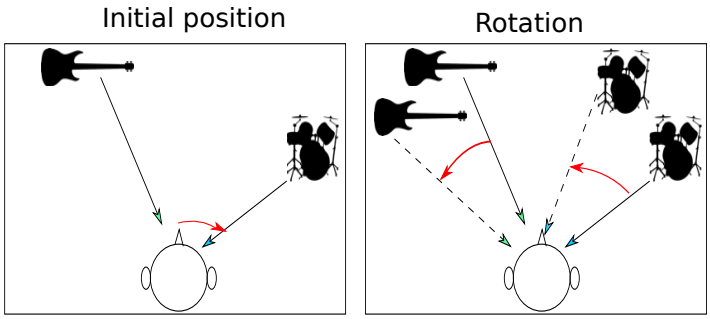


Fig. 1.3 Change of sound sources perception in the case of a head rotation.

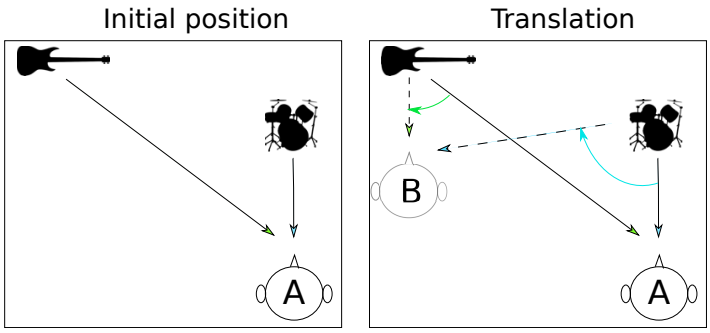


Fig. 1.4 Change of sound sources perception in the case of a head translation.

First, at point A, on the one hand, the drums would be perceived as coming from the front; on the other hand, the guitar would be perceived as coming from the left. Second, at point B, on the one hand, the drums would be perceived coming this time from the right and with less power compared to how it would be perceived from point A. On the other hand, the guitar would be perceived this time as coming from the front and with more power (louder) compared to how it is would be perceived from point A. Imagining the perception of the sound sources according to the example showed in Fig. 1.4, and compare, we can sense the scientific challenge behind translations. Unlike rotations where the transformation of the perception of each sound source is similar (the direction of all the sound sources rotate with the same degree) see Fig. 1.3, translations require a different type of processing for each sound source. Given only the representation format of the sound field at a particular point (mixture), we do not have control over each sound source separately.

In frame of my Ph.D. work, we took a couple of limitations and assets into account:

- We consider a single spherical microphone array since they are very expensive. For more information about the sound field, we could consider some spot microphones that would be close to the sound sources, as illustrated in Fig. 1.1.

- We assume the fact that we can have at our disposal some images about the

sound field since the main application demands to record the environment with cameras.[5]

- We consider the fact that we have a "time of flight type" camera, which allows us to have a map distance matrix.[6] This will help us to have an estimation of the distance between all sound sources and the main antenna.

- We consider fourth or lower-order ambisonic mixtures because, with commercially available spherical microphone arrays, we can have at best a fourth-order ambisonic signals.

In this Ph.D. manuscript, we tackle this problem by providing different approaches to manipulate ambisonic sound fields. This task will be carried out by decomposing the multichannel format into sound sources.

As further described in Chapter 2 ambisonic sound fields are recorded using spherical microphone arrays such as Ambeo VR Mic by Sennheiser,[7] ZM-1 Portable Recorder by Zylia,[8] and Eigenmike by MH Acoustics,[9] as shown in Fig. 1.5. Microphones that are dedicated to ambisonics are compact and usually come in the form of a sphere. The recording is done at a specific point, which will represent during the playback the position where the user's sound perception is virtually projected. In order to be able to navigate in the sound scene, the user's sound perception must be virtually projected from the recording position to the position where he/she would like to move. The main idea behind my Ph.D. subject is to synthesize from the recorded ambisonic sound scene the sound source signals and pan each one of them according to the current user position. In other words, we would like to decompose the 3D sound scene and reconstruct it according to the current user position.

1.4 Contributions and organization of the manuscript

To see through the main problem of my Ph.D. thesis, I organized this manuscript into eight chapters.

In **Chapter 2**, I will briefly recall the ambisonic formalism, which will be backed up with more details in the **first appendix**. I will explain with more information how an ambisonic sound field is recorded and played. Moreover, I will recall all its advantages,

[5] My Ph.D. work is attended to be used in an application in which we want to navigate visually and audibly in a recorded environment.

[6] With a map distance matrix, given the direction of arrival, we can have the distance from which a sound is coming.

[7] `https://fr-fr.sennheiser.com/microphone-3d-audio-ambeo-vr-mic`.

[8] `https://www.zylia.co/zylia-zm-1-microphone.html`.

[9] `https://mhacoustics.com/products#eigenmike1`.

Fig. 1.5 Commercially available coincident and spherical microphone arrays (middle and right). From left to right: Ambeo VR Mic by Sennheiser (coincident array), ZM-1 Portable Recorder by Zylia (spherical array), and Eigenmike by MH Acoustics (spherical array).

such as the ability to transform an ambisonic sound field. Furthermore, I will explain the difficulties to use such a technology for navigation with six-degrees-of-freedom (6-DoF) and I will survey some approaches about this subject (6-DoF navigation in ambisonic sound fields). Finally, I will conclude this chapter by explaining my strategy to perform navigation which is based on decomposing the ambisonic sound field.

In **Chapter 3**, I will recall a crucial mathematical tool, which is the short time Fourier transform (STFT). This tool will help us massively for locating and decomposing ambisonic sound fields. I will secondly gave a survey on sound localization in general, whether in the ambisonic domain or the microphone domain. I will adapt all the methods in the microphone domain to the ambisonic domain. I will finish this chapter by giving a little survey on multichannel sound source separation. This survey will be short for the lack of approaches in the ambisonic domain. However, I will survey a method that was initially proposed ten years ago for the microphone domain. The approach in question is known as multichannel sound source separation based on the local Gaussian model assumption. Finally, I will conclude this chapter by expressing first my interest in this approach, which will later be the center of some of my contributions (Chapter 5, Chapter 6, Chapter 7) and second the need to validate the surveyed localization approaches.

In **Chapter 4**, I will explain in detail my strategy to navigate in ambisonic sound fields and will propose two variants. My strategy is heavily dependent on both sound source localization and separation. Therefore, I will validate the surveyed approaches on sound source localization through some numerical experiments, which will end up giving us some satisfying results compared to state of the art. I will validate my strategy for navigation using an objective metric. I will finally conclude on the fact that I am able to locate sound sources in the ambisonic domain, which will allow us to concentrate my research on the multichannel sound source separation.

In **Chapter 5**, I will be interested in the local Gaussian model (LGM) approach for

multichannel sound source separation in the ambisonic domain. Therefore, I will derive the model equations for the first time from the microphone domain to the ambisonic domain, which will result in checking the adaptation of the approach in the ambisonic domain. I will validate the approach through some numerical experiments using an off-the-shelf toolbox that is based on the local Gaussian model approach.

In **Chapter 6**, I will propose a workflow that guides the multichannel sound source separation with the local Gaussian model approach. The proposed method will be based on adding some spot microphones along the ambisonic antenna. I will validate the approach and will study the efficiency of each block of the proposed workflow through some numerical experiments.

In **Chapter 7**, I will propose to replace the proposed workflow in the sixth chapter with neural networks in order to separate ambisonic musical content. In other words, I will propose in this chapter to perform for the first time a musical sound source separation in the ambisonic domain using neural networks. I will validate my approach through some numerical experiments.

In **Chapter 8** (the final chapter), I will recall my work, conclude on it, and finally, we will propose some perspectives regarding my Ph.D. work.

Chapter 2

Ambisonics and navigation in ambisonic sound scenes

2.1 Introduction and overview

Ambisonics is primarily a format that allows describing spatialized sound scenes. It has been introduced first by Gerzon [42, 40] as a spatialization technique for the first order $L = 1$, and he then named the representation format as the B-Format. This format was afterward extended to further orders $L > 1$ by Malham [71] and Daniel [31, 29] and referred to as the ambisonic format or higher-order ambisonic format (HOA).

One should know that in the literature, the words ambisonics and HOA are used to designate the technology, as well as the representation format. In this manuscript, we will specify if we are writing about the technology or the representation format. We will adopt in the rest of this manuscript the words "ambisonic format/technology" to designate both the "B-format/ambisonic technology" and the "HOA format/technology".

Despite being an old (in the late 70's) technology, ambisonic is nowadays very popular thanks to the new immersive video formats, and it has become the *de facto* standard for 360-degree video soundtracks. Indeed, the ambisonic technology offers:

- A panoramic description of sound fields, with a uniform resolution on the sphere.
- An inexpensive rotation of sound fields that are related to the user's head orientation.
- A convincing binaural rendering that requires relatively less computing compared to other 3D audio formats [76].

Moreover, the ambisonic format has other features:

- One can record any sound field using a compact spherical microphone array (SMA) and get its representation in the ambisonic format.

- The ambisonic format is hardware-independent. It can be played on any loudspeaker distribution.

Thanks to these advantages, ambisonic technology is becoming increasingly popular for 360 videos. Several companies such as Facebook [1] and Google (Youtube)[2] propose free plugins and platforms for their customers/users to create their content.

The ambisonic format is part of a complete framework (Fig. 2.1) as a 3D audio technology.

In the following, we will explain the ambisonic format by describing its mathematical formalism and its specificities. And why it is challenging to perform navigation when it comes to ambisonic sound scenes.

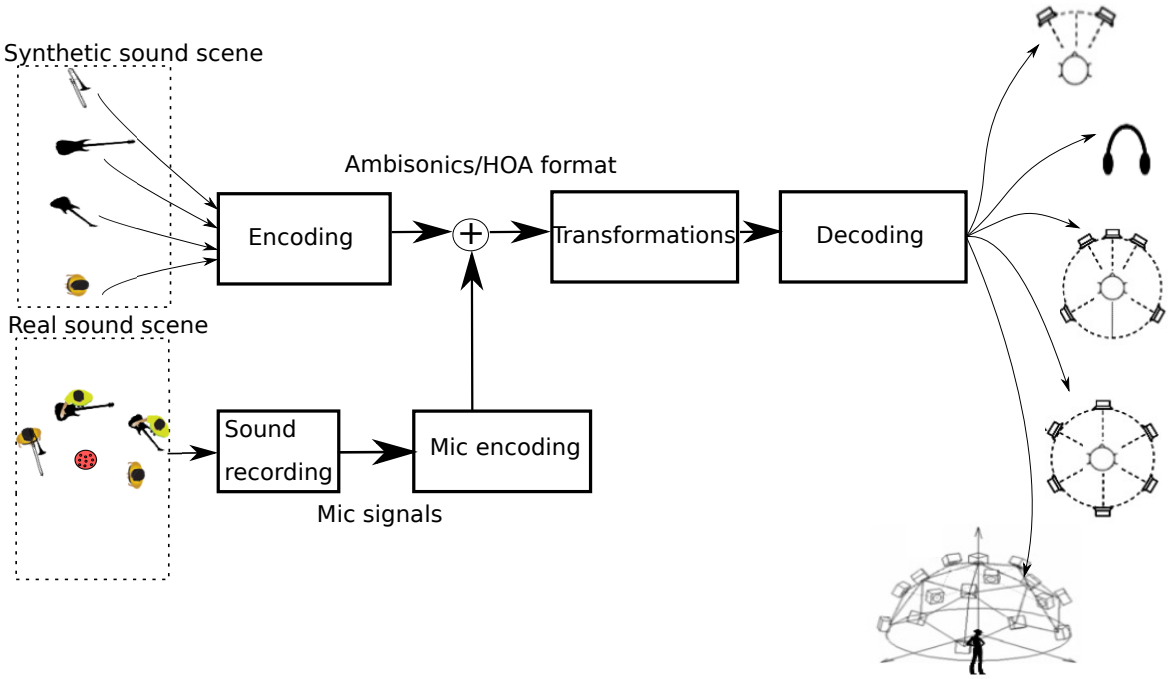


Fig. 2.1 Ambisonics framework adapted from [29].

2.2 Mathematical formalism

In this section, we describe the ambisonic format theoretically. First, we introduce the coordinate system in which we consider sound fields. The ambisonics approach bases the sound field description on the spherical coordinate system. Therefore, we consider the spherical coordinate system presented in Fig. 2.2, where a given point is defined

[1] https://facebookincubator.github.io/facebook-360-spatial-workstation/KB/CreatingVideosSpatialAudioFacebook360.html.

[2] https://support.google.com/youtube/answer/6395969.

by its radius r (distance from the origin), azimuth θ, and elevation ϕ. We can relate them to the Cartesian coordinates with the following system:

$$\begin{cases} x = r\cos(\theta)\cos(\phi) \\ y = r\sin(\theta)\cos(\phi) \\ z = r\sin(\phi) \end{cases} . \tag{2.1}$$

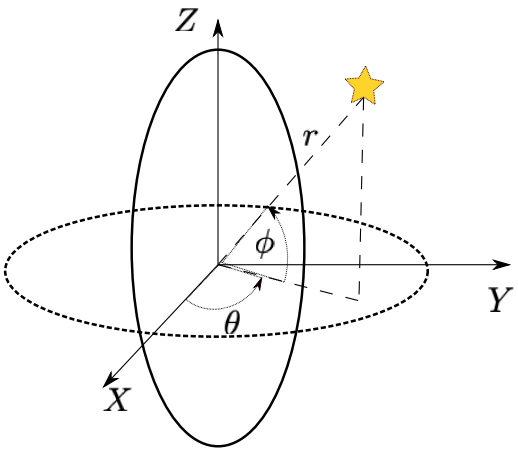


Fig. 2.2 Spherical coordinate system. A given point in space (yellow star) is described by its radius r, azimuth θ, and elevation ϕ.

2.2.1 Representation of a single plane wave

Mathematically, the ambisonic format is based on decomposing sound fields into a series of spherical harmonics. Indeed, the sound pressure at a specific point in space can be written as a weighted summation of spherical harmonic functions up to an infinite order. Let us first consider a harmonic plane wave (one frequency) coming from the direction (θ_p, ϕ_p), and that carries a signal with an amplitude of s_p:

$$p(k, r, \theta, \phi) = s_p e^{ikrcos(\gamma)}, \tag{2.2}$$

with γ being the angle difference between the observation direction (θ, ϕ) and the source direction (θ_p, ϕ_p), $k = 2\pi f = \omega/c$ denotes the wave number, ω denotes its pulsation and c is the speed of sound. From solving the wave propagation equation (see Appendix A),

the sound pressure can be decomposed on the spherical harmonic basis as follows:

$$p(k, r, \theta, \phi) = s_p \sum_{l=0}^{\infty} i^l j_l(kr) \sum_{e=-l}^{l} Y_{le}(\theta, \phi) Y_{le}(\theta_p, \phi_p), \tag{2.3}$$

where:

- Y_{le} represent the spherical harmonic functions
- j_l represents the spherical Bessel functions of the first kind
- $l \in \mathbb{N}$, and $e \in -l, \ldots, l$. For more information about these functions please refer to Appendix A.

In practice, the decomposition must be truncated to a certain order L:

$$p(k, r, \theta, \phi) \approx s_p \sum_{l=0}^{L} i^l j_l(kr) \sum_{e=-l}^{l} Y_{le}(\theta, \phi) Y_{le}(\theta_p, \phi_p). \tag{2.4}$$

The order of the decomposition is related to the spatial resolution: the higher the order is, the more accurate the decomposition will be. The ambisonic coefficients for order L are then identified from Eq (2.4). They are given by:

$$\forall l \in \{0, \ldots, L\}, \forall e \in \{-l, \ldots, l\} \qquad z_{le} = s_p Y_{le}(\theta_p, \phi_p). \tag{2.5}$$

Since the spherical harmonic basis is an orthonormal basis, the decomposition of an order L means the decomposition of the sound pressure on the first $M = (L+1)^2$ functions of the spherical harmonic basis. This decomposition leads to obtain $M = (L+1)^2$ ambisonic signals/channels.

For any plane wave, carrying a signal s_t at a time t and coming from the direction (θ_p, ϕ_p), the ambisonic coefficients can be written as follows [29]:

$$\mathbf{z}_t = \mathbf{y}(\theta_p, \phi_p) s_t, \tag{2.6}$$

where $\mathbf{z}_t \in \mathbb{R}^{M \times 1}$ are the ambisonic coefficients (referred to as "the ambisonic signals"), $\mathbf{y}(\theta_p, \phi_p)$ denotes the spherical harmonic vector corresponding to the direction (θ_p, ϕ_p), which coefficients are $Y_{le}(\theta_p, \phi_p)$, with $l, e, \in \{0, \ldots, L\}, \{-l, \ldots, l\}$, respectively. Note that in the ambisonic domain, the steering vector of a direction corresponds to its spherical harmonic vector.

2.2.2 Representation of multiple plane waves

In the case we have J harmonic plane waves with magnitude signal s_p, the ambisonic coefficients at the order L are given by:

$$\forall l \in \{0, \dots, L\}, \forall e \in \{0, \dots, l\} \qquad z_{le} = \sum_{p=1}^{J} s_p Y_{le}(\theta_p, \phi_p). \tag{2.7}$$

Similarly, to the one plane wave case, we can write the ambisonic coefficients as follows:

$$\mathbf{z}_t = \mathbf{Y}\mathbf{s}_t, \tag{2.8}$$

where $\mathbf{s}_t \in \mathbb{R}^{J\times 1}$ contains the magnitude of the plane wave signals at the time t, and $\mathbf{Y} \in \mathbb{R}^{M\times J}$ which columns are $\mathbf{y}(\theta_p, \phi_p)$ the spherical harmonic vectors corresponding to the plane wave directions of arrivals (DoA).

In the literature the step of presenting a sound field in the ambisonic format is referred to as encoding. With, Eq. (2.8), as one can imagine, we can create synthetic sound scenes in the ambisonic format. However, when it comes to record sound scenes and present them in the ambisonic format, this step is referred to as microphone encoding. This step is discussed in Section 2.3.

2.3 Recording ambisonic signals

In the literature, the spherical harmonic functions directivities are often compared to the directivities of coincident microphones. In Fig. 2.3, we plotted the first-order harmonic functions. As it is shown in the figure, the first spherical harmonic function corresponds to the directivity of an omnidirectional microphone, the second, the third, and the fourth, correspond to the directivity of figure-of-8 microphones oriented. We can deduce that it is possible to record ambisonic signals using microphones with the same directivity as the spherical harmonic function, and they must all be placed at the same point. However, there are two physical problems; the first one deals with the complexity of the spherical harmonic functions when the order increases. The second problem is that the microphones must be placed at the same point, which is physically impossible.

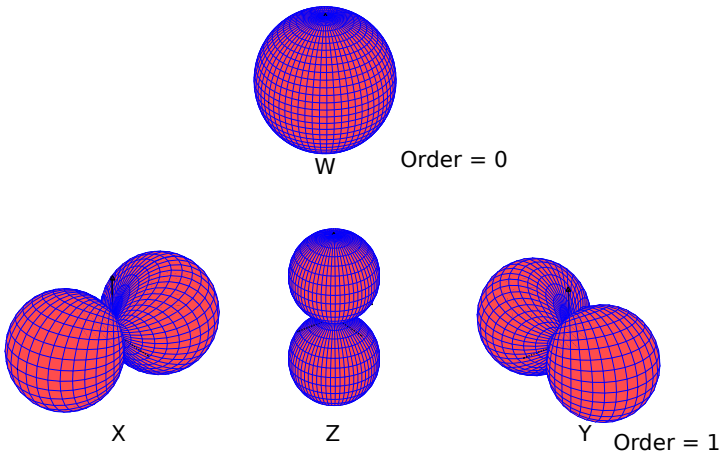


Fig. 2.3 3D polar pattern of the supposed microphone for the first four channels, Also known as the first four spherical harmonic functions. Order 0 contains the omnidirectional function W, the 1st order contains: X, Y, and Z.

A more practical solution is to recover the ambisonics signals from measuring the pressure on a sphere of a fixed radius r_s [75]. Let us consider $p(k, r_s, \theta, \phi)$ as the pressure on point (r_s, θ, ϕ) of the sphere. The ambisonic signals are given by projecting the value of the pressure measured on the sphere surface on the spherical harmonic basis, which is given by [82]:

$$z_{le}(k) = EQ_l(kr_s) \int_{\theta=0}^{2\pi} \int_{\phi=-\frac{\pi}{2}}^{\frac{\pi}{2}} p(k, r_s, \theta, \phi) Y_{le}(\theta, \phi) \cos(\phi) \mathrm{d}\phi \mathrm{d}\theta, \tag{2.9}$$

where $EQ_l(kr_s)$ is called the equalization term and depends on the spherical Bessel functions j_l. It is given by:

$$EQ_l(kr_s) = \frac{1}{i^l j_l(kr_s)}. \tag{2.10}$$

With Eq. (2.9), it is clearly understable that spherical microphone arrays do not directly give ambisonic signals. A pre-processing must be done in order to obtain them. We can say that spherical microphone encoding step deals with two steps:

- Projecting the microphone signals on the spherical harmonic basis.
- Multiplying the outcome of the first step by a frequency-dependent function $EQ_l(kr_s)$.

However, this approach yields two problems. The first problem is the extreme amplification of the signals when the Bessel functions vanish at some frequencies. A solution to this problem is to consider cardioid microphones as sensors instead of omnidirectional microphones [74]. With this solution the signals provided by the cardioid microphones depend on the pressure and its gradient [58] which results in adding a term in the denominator of $EQ_l(kr_s)$, and prevents it from vanishing and therefore preventing an extreme amplification of microphone noises. The equalization term is then given by [74]:

$$EQ_l(kr_s) = \frac{1}{i^l[j_l(kr_s) + kj_l'(kr_s)]}, \tag{2.11}$$

where j_l' are the derivative spherical Bessel functions j_l. The second problem is with the measurement of the pressure on the whole surface of the sphere. It is indeed physically impossible. In practice a limited number I of sensors are used.

Considering a limited number of cardioid microphones I_s on the sphere, and using the expression of the pressure at a point of a sphere $(r_s, \theta_s(q), \phi_s(q))$ in terms of the ambisonic signals, the signal $x(q,k)$ at sensor $q = 1, ..., I_s$ of the spherical array microphone is written as follows:

$$x(q,k) = \sum_{l=0}^{L} i^l[i^l j_l(kr_s) + kj_l'(kr_s)] \sum_{e=-l}^{l} z_{le}(k) Y_{le}(\theta(q), \phi(q)), \tag{2.12}$$

which can be seen as I_s equations for $M = (L+1)^2$ unknown ambisonics signals z_{le}. Indeed Eq. (2.12) can be seen as a linear system given by:

$$\mathbf{x}(k) = \mathbf{Y}_s \mathbf{W_s}(k) \mathbf{z}(k), \tag{2.13}$$

where the vectors $\mathbf{x} \in \mathbb{R}^{I_s}$ and $\mathbf{z} \in \mathbb{R}^M$ contain the signals from the spherical microphone array and the ambisonic signals, respectively. The matrices $\mathbf{Y}_s \in \mathbb{R}^{I_s \times M}$ and $\mathbf{W_s} \in \mathbb{R}^{M \times M}$ are given by:

$$\mathbf{Y}_s = [\mathbf{y}^\top(\theta_s(1), \phi_s(1)), \dots \mathbf{y}^\top(\theta_s(I_s), \phi_s(I_s))]^\top \tag{2.14}$$

$$\mathbf{W_s} = \mathrm{diag}([EQ_0(kr_s), \dots, EQ_L(kr_s)]^\top), \tag{2.15}$$

where $(.)^\top$ represents the transpose operator, andEQ_l is given by Eq. (2.11).

In order for Eq. (2.13) to have a unique solution the number of microphones I_s must be at least equal to the number of ambisonics signals $M = (L+1)^2$, *i.e*, in order to avoid an underdetermined problem the number of channels must be larger or equal

to the number of the ambisonics signals. In our case, the Eigenmike has 32 sensors, by encoding the 32 signals we can have up to 4^{th} order ambisonic signals (*i.e* 25 channels). The ambisonic signals are estimated by solving Eq. (2.13), they are given by:

$$\hat{\mathbf{z}}(k) = \mathbf{E_s}(k)\mathbf{Y}_s^{\dagger}\mathbf{x}(k) \tag{2.16}$$

$$\mathbf{E_s} = \mathbf{W_s}^{-1} = \text{diag}([\frac{1}{EQ_0(kr_s)}, \ldots, \frac{1}{EQ_L(kr_s)}]), \tag{2.17}$$

the notation $(.)^{\dagger}$ represents the pseudo-inverse operation. In [74] the author pointed out several problems concerning the reliability of the ambisonic signals estimation, mainly caused by the sensors noise and the errors about their positions. These errors may get amplified by the matrix $\mathbf{E_s}$. The author recommends to use a regularized version of the matrix $\mathbf{E_s}$, in order to reduce the risks of instabilities. It is recommended to replace the coefficients in diagonal of the matrix terms in $\mathbf{E_s}$ by the following coefficients F_l [74, 76]:

$$F_l(kr_s) = \frac{|i^l[j_l(kr_s) + kj_l'(kr_s)]|^2}{|i^l[j_l(kr_s) + kj_l'(kr_s)]|^2 + \lambda^2}, \tag{2.18}$$

where λ is a regularization parameter that needs to be adjusted.

Commercially available microphone arrays presented in the introduction allow to produce ambisonic signals of order 1 (Ambeo), 3 (ZM-1), and 4 (Eigenmike). Other prototypes of spherical microphone arrays were proposed in the literature such as the Orange Labs Prototype in [74], the University of Maryland prototype in [133], and the CNAM prototypes MemsBedev and SpherBedev in [67].

For more information about encoding microphone signals please refer to [30]. Acquiring ambisonic signals with microphone arrays results in two main issues:

- At high frequencies, the distance separating the microphone capsules induces spatial aliasing.
- At low frequencies, the small dimension of the array concerning the wavelength makes it very difficult to acquire the signals corresponding to the higher-order spherical harmonics.

This means that the effective order of ambisonic scenes recorded with microphone arrays varies as a function of the frequency. For example, using the mh Acoustics' Eigenmike to acquire fourth-order ambisonic signals, we obtain ambisonic signals with the following effective orders $\hat{L}$ (with typical encoding filters) [24]:

- $\hat{L} = 1$ below $300Hz$.
- $\hat{L} = 2$ between 300 and $1300Hz$.

- $\hat{L} = 3$ between 1300 and $2200Hz$.
- $\hat{L} = 4$ above $2200Hz$.

We can see the first issue in Fig. 2.4, where the real directivity of the first and the second channel are plotted against their frequencies. We can observe that some problems are present for high frequencies. Indeed, the first and second channel's directivity in really high frequencies does not match the theoretical ones. These problems are discussed in more detail in [76].

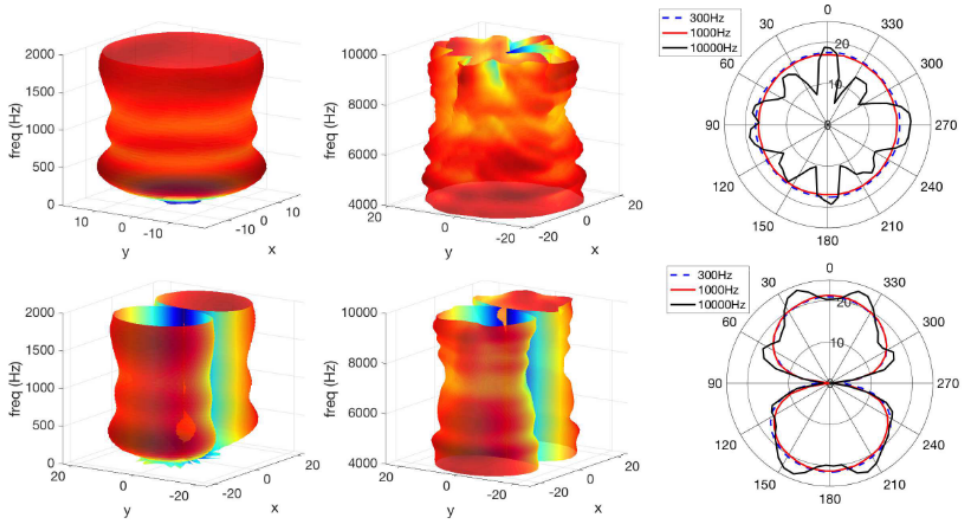


Fig. 2.4 Real directivity of the first and the second channels in regards to the frequency. Figure from [10]. The ambisonic signals in this case are acquired from encoding Eigenmike signals.

Usually these issues are under estimated and not talked about. This will help to understand Section 5.5.3 in Chapter 5.

2.4 Transformations

The ambisonic format is eligible to be manipulated [64, 63]. These transformations are linear, which means they can be applied by simple matrix multiplication.

2.4.1 Rotations

Rotations consist in pivoting the entire sound scene around an axis [29, 64]. We recall that the representation of the sound field in the ambisonic format is given at a particular point whether the sound field was recorded using a SMA and encoded to the ambisonic format, or synthesized. With ambisonics, sound scenes can be rotated around any

axis that goes through the representation point. This can be seen as a combination of rotations around the axes of the basis (Ox, Oy, Oz), with O being the representation point. Performing a rotation requires a simple matrix multiplication, which is one of the main reasons ambisonics has become the *de-facto* 3D audio standard for 360-video.

One should know first that the spherical harmonic functions can be rotated around the axes (Ox, Oy, Oz) by applying spherical harmonic rotation matrices. These matrices have a specific design; they are diagonal by blocks. Indeed, to rotate the basis, one must rotate each function of the basis. Note that each block of the matrix is dedicated to a given function of the basis. Rotation matrices $\mathbf{R}^i \in \mathbb{R}^{M\times M}$ are shaped as follows [29, 64]:

$$\mathbf{R}^i = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \mathbf{R}_1^i & \cdots & 0 \\ \vdots & 0 & \mathbf{R}_2^i & \vdots \\ \vdots & \vdots & \cdots & \mathbf{R}_L^i \end{pmatrix}. \tag{2.19}$$

The analytical expression of the matrices $\mathbf{R}_l^i \in \mathbb{R}^{(2l+1)\times(2l+1)}$ are listed in [29], the index i refers to an axis from the basis (Ox, Oy, Oz). All the blocks in rotation matrix $\mathbf{R}^i$ must describe the same rotation at a different order L. It is possible to perform a combination of the elementary rotations (pich, roll, yaw) [29, 64]:

$$\mathbf{R} = \mathbf{R}^x(\psi)\mathbf{R}^y(\phi)\mathbf{R}^z(\theta). \tag{2.20}$$

The new ambisonic signals are given by:

$$\mathbf{z}' = \mathbf{R}\mathbf{z}. \tag{2.21}$$

For our application, such a feature is going to be used. With rotations, we already have three degrees of freedom (3-DoF). Indeed, we can already adapt the sound field to the user's head orientation. In order to have six degrees of freedom (6-DoF), we need the ability to compute the sound field in the ambisonic format at another point in space from the ambisonic signals at the recording position.

For this a possible solution already discussed in the introduction is the decomposition of the sound field into plane waves. In the next subsection we will discuss beamforming approaches which are known methods for decomposing ambisonic sound scenes into plane waves.

2.4.2 Focus on a direction using beamforming techniques

In this section we discuss a technique that can be applied on the ambisonic format with which we can focus on a given direction [69, 91, 92].

Beamforming is a spatial filtering of a mixture signal made that enhance a source signal coming from a particular direction (θ, ϕ). In other words, with beamforming, we can simulate a mono-signal recording in a given direction. It is based on combining the ambisonics signals in a specific way so that it provides a similar effect of a recording by a cardioid microphone. This transformation is none else than a focus towards a direction. Different approaches exist; basic projection, max-Re projection, and in-Phase projection. The last two projection were proposed to be used for decoding[3] purpose.

An estimation of the signal s_j coming from the direction (θ_j, ϕ_j) by a projection type beamformer is given by [29]:

$$\hat{s}_j(t) = \frac{\mathbf{y}^\top(\theta_i, \phi_i).\mathrm{diag}(\mathbf{g}_l)}{||\mathrm{diag}(\mathbf{g}_l).\mathbf{y}^\top(\theta_i, \phi_i)||}\mathbf{z}(t), \tag{2.22}$$

where $\mathbf{y}(\theta_i, \phi_i) \in \mathbb{R}^{M\times 1}$ is the spherical harmonic vector of the direction that we would like to focus on, $\mathbf{g}_l$ is the vector that determines the used approach. For the matched filter or what we refer to as basic projection, the vector $\mathbf{g}_l = [1, \ldots, 1]^\top$. The max-Re and In-Phase approaches were suggested by Daniel in [29]. These methods offer a better focus towards a direction, because they extremely attenuate the opposite direction sounds, in return, closest sounds are accentuated. The analytical expression for both Max-Re and in-Phase approaches are presented in [29, 74]. Fig. 2.5 gives an idea of the form of the beamforming for each approach at different orders, the beams were formed towards the direction $(\theta = 0°, \phi = 0°)$. Beamfoming techniques are discussed in more detail in Chapter 3 Section 3.4.4.

In Section 2.5, we will be interested by how the ambisonic format can be played back in 3D.

2.5 Playback

This step is the last one in the ambisonics framework. It concerns the playback process. The main goal is to reproduce the captured or the synthesized scene by playing it in 3D on a loudspeaker distribution or on headphones.

[3]The decoding step aim to play back the sound field on loudspeaker distribution. A brief discussion about this subject is given in Section 2.5.

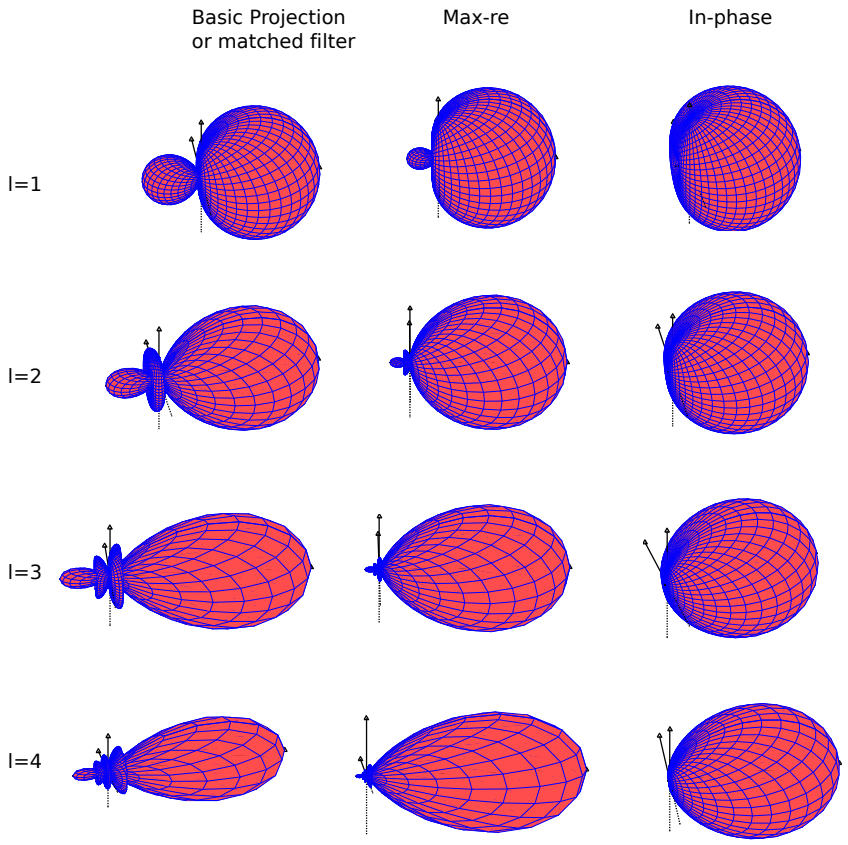


Fig. 2.5 Example of the beam shape for each approach at different order. The beams were formed towards the direction $(\theta = 0, \phi = 0)$.

2.5.1 Loudspeaker distribution

Ambisonic technology is independent of the playback system, although, many playback systems exist. This is possible thanks to the decoding step, in which the adaptation of the ambisonic signals to the playback configuration is made. Although the ambisonic technology is known for being independent of the restitution system, the playback configuration must follow several rules so the auditory rendering will be more realistic and exploit the maximum of information in the ambisonic signals [74].

There are different approaches for adapting the ambisonic signals to the playback system. The first one is called the basic approach, originally introduced by Gerzon [41] for the first order, Malham [71] for the second and the third order and then by Daniel [29] for further orders L. The idea behind these approaches is to find the signals of the loudspeakers. Given the direction of the loudspeakers, the problem is solved by looking at it backwards. Let us consider the sound field containing N_{sp}(represents the number of loudspeakers) sound sources, each one of them is coming from the direction $(\theta_{sp_j}, \phi_{sp_j})$ and emitting a signal s_{sp_j} at the time t. The ambisonic coefficients at the time t of the described sound field are given by:

$$\mathbf{z}_{sp,t} = \mathbf{Y}_{sp}\mathbf{s}_{sp,t}, \tag{2.23}$$

where $\mathbf{Y}_{sp} \in \mathbb{R}^{M\times N_{sp}}$ is a matrix that contains the spherical harmonic vector corresponding to the DoAs of the loudspeakers(θ_{sp}, ϕ_{sp}), which are the direction of the loudspeakers and $\mathbf{s}_{sp,t} \in \mathbb{R}^{N_{sp}\times 1}$ is vector that contains the magnitude signals s_{sp_j} of the loudspeakers.

Any ambisonic content $\mathbf{z}_t$ played on the loudspeaker system described above respects the following equation:

$$\mathbf{z}_t = \mathbf{z}_{sp,t} = \mathbf{Y}_{sp}\mathbf{s}_{sp,t}. \tag{2.24}$$

From Eq. (2.24), knowing in which directions the loudspeakers are placed (θ_{sp}, ϕ_{sp}), we can deduce the loudspeakers signals $\mathbf{s}_{sp,t}$:

$$\mathbf{s}_{sp,t} = \mathbf{D}.\mathbf{z}_t \tag{2.25}$$

$$\mathbf{D} = \mathbf{Y}_{sp}^{\dagger}. \tag{2.26}$$

The solution presented in Eq. (2.26) is a general solution that is adjustable to any kind of loudspeaker distribution if their direction (from which the sound is coming) are known. For systems with regularly distributed loudspeakers [30] a particular solution

based on projecting the ambisonic signals on the encoding vector of each loudspeaker is possible since:

$$\mathbf{Y}_{sp}^{\top}\mathbf{Y}_{sp} = N_{sp}.\mathbf{I}_{N_{sp}}, \tag{2.27}$$

where N_{sp} denotes the number of loudspeakers of the playback system. Therefore for a regular loudspeakers distribution the decoding matrix $\mathbf{D}_{basic} \in \mathbb{R}^{N_{sp}\times M}$ could be presented as follows:

$$\mathbf{D}_{basic} = \frac{1}{N_{sp}}.\mathbf{Y}_{sp}^{\top}. \tag{2.28}$$

In order to improve some concepts for the playback such as widening of the listening area for example, improvements of the so-called basic solution were proposed by Daniel in [29], which are the Max-re and In-phase, they are given by:

$$\mathbf{D} = \mathbf{D}_{\text{basic}}.\mathbf{\Gamma}_{M}, \tag{2.29}$$

where $\mathbf{\Gamma}_{M} \in \mathbb{R}^{M,M}$ is a diagonal matrix, its analytical expression (whether Max-re or In-phase) is presented in [29, 74].

The loudspeakers signals are then given by:

$$\mathbf{s}_{sp,t} \quad = \quad \mathbf{D}.\mathbf{z}_{t}. \tag{2.30}$$

2.5.2 Headphones

A possible strategy is to decode the ambisonic signals on a virtual loudspeaker distribution, and apply two (left and right) head-related transfer functions (HRTFs). Thereby, the signal of each headphone ear is given by [74]:

$$s_{ear}(\omega) = \sum_{sp=1}^{N_{sp}} h_{ear}(\omega, \theta_{sp}, \phi_{sp})\, s_{sp}(\omega), \tag{2.31}$$

where $h_{ear}(w, \theta_{sp}, \phi_{sp})$ is the HRTF related to whether the left ear or right ear and the loudspeaker sp.

2.6 Survey on navigation in ambisonic recordings

In VR contents, the immersion consists in giving the listener the ability to move freely in the environment. If accompanied with ambisonics technology, the framework must have the ability to perform navigation in six-degrees-of-freedom (6-DOF) *i.e.*, a combination of the three rotational degrees and a translationnel degree. Such operation is still difficult to achieve. Indeed, the DoA and the distance from each source to the newest listener's position aren't similar for the sources, unlike rotations in which the angle of rotation is identical for all the sources. Therefore, given only the mixture, the rotations are more comfortable to perform. However, a transition from point A to point B requires to take into consideration both the DoA and distance of the listener's position for each source in the sound field, which represents the main difficulty of my Ph.D. work.

Different 6-DoF approaches exist and can be categorized into three kinds:

- A synthetic rendering in which the whole sound scene is synthetic with known object sounds, such rendering is commonly used in video games. In this case, it is not difficult to perform 6-DOF navigation. Indeed, with known sound object, the simulation of a new recording position is done by a simple encoding, *i.e*, by taking into consideration the new DoA and distance of each sound object. An artificial reverberation can be added for each sound object using a room impulse response simulator.

- A rendering from multiple ambisonics recordings. In this case, several SMAs are used and positioned in different points of the sound field, the simulation of a movement is done by interpolating the ambisonic signals of the different used SMAs [109, 116, 72, 114, 104, 116].

- A rendering from a single ambisonic recording. This requires to decompose the ambisonic scene into directional components.

We are more interested in the third category. Therefore, for the next of this section, we survey strategies from this category.

In [115], a comparison between three different approaches was conducted:

- The first approach is called virtual higher-order ambisonic loudspeakers [86]. This method deals with simulating ambisonic playback over a virtual array of loudspeakers, the binaural navigation of this method requires to apply HRTFs to each loudspeaker signal that depends on the relative position of the listener to each loudspeaker.

- The second approach is a plane wave expansion method [105], which consists of a primary decomposition of the ambisonic sound field into a limited number of plane waves and taking into consideration the head translation for each plane wave.

- The third method is based on a sound field expansion, in which new ambisonic signals corresponding to new positions are computed by re-expanding the sound field using frequency-domain translation coefficients [115].

In terms of the sound field reconstruction, results showed that virtual-HOA and plane-wave translation techniques create static sweet-spots[4] that restrict the listener's range of motion. In contrast, the sweet-spot created with sound-field re-expansion coincides with the listener's translated position. A notable issue with all listed methods is the inability to characterize the response of the room as a function of user location. All of these techniques are limited by the original expansion accuracy and the region of validity. Consequently, for low-order recordings and those containing sources very near to the microphone array, the range of motion allowed by any navigational technique is significantly limited.

Another approach in [97] deals mainly with first-order ambisonic sound scenes. The main idea is to decompose the sound field using a directional audio coding (DirAC) approach explained later in Chapter 3 Section 3.3.3.1. A popular strategy that decomposes a first-order ambisonic sound field at each time-frequency bin into the DoA of the dominant source and a diffuseness coefficient. Given the fact that the direction of the dominant signal is encoded as a three-dimensional vector of unit length, and that source positions are known, the authors propose to integrate the distance information (corresponding to the newest listener's position) in the encoded DoA, as well as the rotation transformation. The new ambisonic signals are given by using a DirAC decoder with the newest DoA, the same diffuseness coefficient, and the same first channel, Fig. 2.6 illustrates the principle of the approach in a diagram.

Another approach is "Ambisonic sound field navigation using directional decomposition and path distance estimation" [2]. This method is also based on a plane wave decomposition, in which a matching pursuit algorithm is used to extract a source corresponding to the direction with the maximum of power and, therefore, the one that minimizes the residual sound field. This algorithm is run until a desired number of sources is reached. The second step of this approach consists in estimating the distance between each extracted subspace and the listener's newest position. At the end the ambisonic sound scene will be described by the estimated sound objects. The desired ambisonic signals corresponding to the new listener's position is given by re-encoding

[4]The reconstruction region of a sound field, it is an area in which a normalized reconstructed error is smaller than 4%.

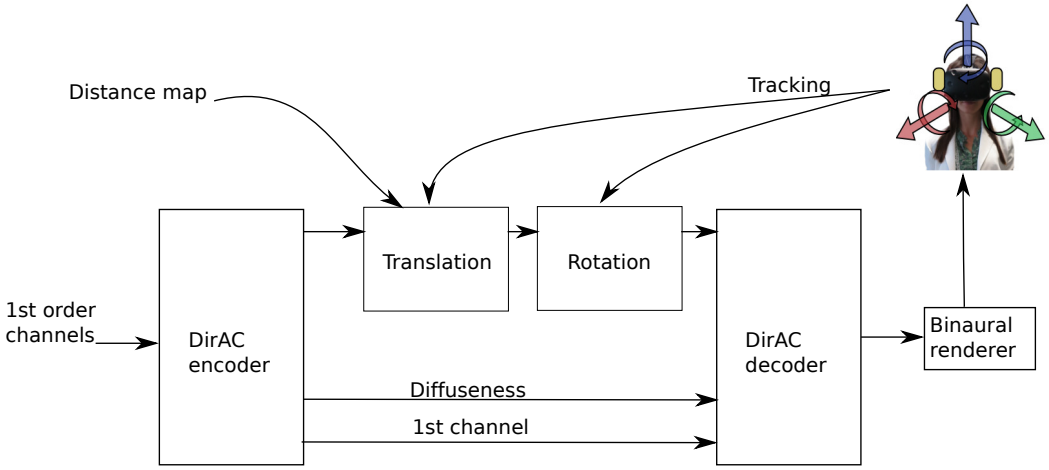


Fig. 2.6 Navigation with 6-DOF using DirAC approach [97].

each estimated source while tanking into consideration its estimated distance and add them to each other with the residual sound field.

As we can imagine, with the method in [2], it would be challenging to handle sound fields with sound sources that are close to each other. Indeed the accuracy of a plane wave decomposition with beamforming highly depends on the order of the mixture. For our application, we can have at best 4^{th} order ambisonic sound fields. This can be insufficient to separate close sources. With our work, we will propose a navigation approach base on multichannel sound source separation to overcome this problem (see Chapter 4).

2.7 Conclusion

In this chapter, we briefly recalled the ambisonic format (more information is given in Appendix A). We explained with more details how ambisonic sound fields are recorded using spherical microphone arrays. Knowing how the ambisonic format is acquired from microphone recordings will be an essential key to Chapter 5, and to the generation of our simulations (see Appendix B). We also discussed the advantages of ambisonic format with recalling ambisonic transformation. For our application, rotation will be heavily used. We recalled the playback process of ambisonic format because in our application playing the sound is essential. Moreover, binauralizing ambisonic sound fields will be used later for an objective evaluation in Chapter 4 Section 4.4. We closed this chapter by surveying some approaches for navigating in ambisonic sound fields from the third category (rendering from a single ambisonic recording, which requires to decompose the ambisonic scene into directional components). We concluded this chapter by discussing our interest in the third category, which deals with navigating

using one SMA, which requires to decompose the ambisonic sound fields. Therefore, for the next chapter, we will survey some sound source localization and separation for ambisonic mixtures.

Chapter 3

Survey on sound source localization and sound source separation

In Chapter 2 Section 2.6, we brought to light three different categories of 6-DoF navigation in ambisonic sound fields. The third category is the one that we are interested in, with which the rendering is from a single ambisonic recording. Since the ambisonic format is a mixture, one must be able to decompose it into directional components and reconstruct it according to the user's movements.

Therefore, we can already say that in order to be able to navigate from a single ambisonic recording, we will need to know where the sound sources are coming from and estimate their signals. More information about navigation in ambisonic sound field are communicated in Chapter 4.

In this chapter, we study several approaches for both sound source localization and separation. Most of these approaches operate in the time-frequency domain, where sound fields generally have a sparse representation, which happens to be a decisive advantage to many sound source localization and separation methods. In the next section, we discuss one of the most popular time-frequency approaches. After the first section in the second one, we will discuss with more details the mixture models in both domains (ambisonic and microphone). In the third section, we will discuss several sound source localization approaches in both microphone and ambisonic domain and adapt the former ones in the microphone domain to the ambisonic domain. In the fourth section, we will discuss sound source separation approaches. In the last section, we will conclude this chapter.

3.1 Short Time Fourier Transform

Audio signals such as speech and music are non-stationary signals. Therefore spectral transforms such as the discrete Fourier transform (DFT) do not emphasize temporal information of real audio signals. An improved frequency representation must thereby have a specific time granularity as well. These kinds of time-frequency transformations help to have a sparsity representation of the audio signals. They were exploited in the literature for many audio applications such as source separation (especially underdetermined ones in which sparsity is crucial to be exploited), speech recognition, noise reduction, and echo cancellation.

The most popular time-frequency representation is the short-time Fourier transform (STFT). It consists in applying the DFT on successive time segments of the signal by applying a sliding window. In order to understand how a STFT decomposes a time signal, let us consider a signal s of a certain duration t. The signal is divided into successive time segments n by applying a time window with length Q as follows:

$$s_n(t) = w(t).s(t + nH) \quad 0 \leqslant t \leqslant Q - 1, \tag{3.1}$$

where Q is the window length, n is the frame index, H is the so-called "hop-size" (determining the amount of overlap between consecutive segments). The STFT consists in applying a simple DFT of F samples to each segment. Note that we change the notations only for this section for a more convenient presentation of the STFT, we usually note the time t as an index instead of putting it in between parentheses.

The value of the STFT for the time-frequency bin (f, n) is given by:

$$s_{f,n} = \sum_{t=0}^{Q-1} s_n(t) e^{-j\frac{2\pi ft}{F}} \tag{3.2}$$

where f is the frequency index. An illustration of the STFT analysis is given in Fig. 3.1.

We can reconstruct perfectly the original signal by using an ISTFT (STFT inverse), which is basically the inverse process of a STFT. First, an inverse DFT (IDFT) is applied to each local spectrum:

$$\hat{s}_n(t) = \frac{1}{F} \sum_{f=0}^{F-1} s_{f,n} e^{j\frac{2\pi ft}{F}} \quad 1 \leqslant n \leqslant N, \tag{3.3}$$

which are exactly the signals from Eq.(3.1). These signals can be written with respect to the original signal:

$$\hat{s}_n(t - nH) = w(t - nH)s(t) \quad nH \leqslant t \leqslant nH + Q - 1. \tag{3.4}$$

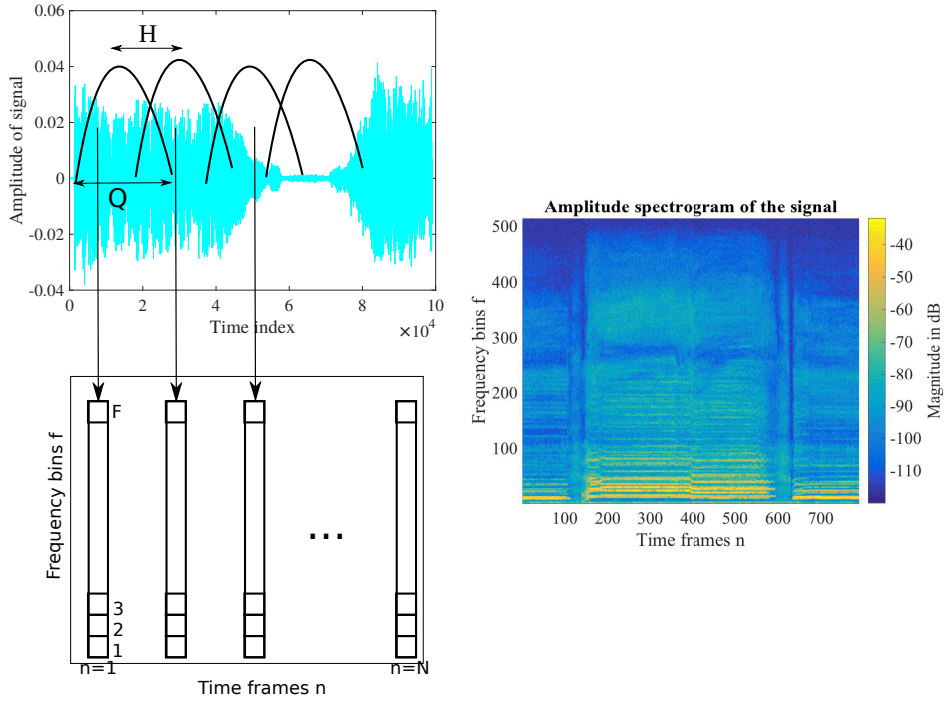


Fig. 3.1 The STFT process.

Second, the signals $\hat{s}_n$ are added to reconstruct the original signal $s(t)$. Looking at Eq. (3.4), in order to have a perfect reconstruction, the following constraint must be satisfied:

$$\sum_{n=1}^{N} w(t - nH) = 1. \tag{3.5}$$

A usual approach consists in using an overlap-add[1] process with a second window v.

$$\hat{s}(t) = \sum_{n=1}^{N} v(t - nH)w(t - nH)s(t). \tag{3.6}$$

Therefore, from Eq. (3.6), we can deduce that a perfect reconstruction can be

[1]Overlapping the windows helps to retrieve data when the window decrease to zero at boundaries. Usually in signal processing 50% overlap is used.

obtained if the following constraint is satisfied:

$$\sum_{n=1}^{N} v(t - nH)w(t - nH) = 1. \tag{3.7}$$

Knowing the time-frequency representation, we can now explain with more details the mixture model in the microphone domain and in the ambisonic domain. This is discussed in the next section.

3.2 Mixture models

3.2.1 Mixture model in the microphone domain

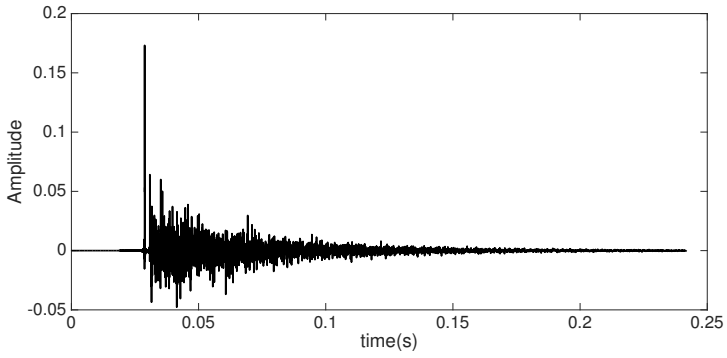


Fig. 3.2 An example of a room impulse response.

When a microphone captures a sound source in a reverberant environment, the captured signal is a filtered version of the sound source. The filter in question is referred to as the room impulse responses (RIR), which represents the interaction of the sound source with the environment from the microphone's point of view. In other words, it represents the propagation of the sound source signal in the room, between the sound source and the microphone. An example of RIR is represented in Fig. 3.2. An impulse response between a sound source and a microphone is known as a mixture filter in the sound source separation community. Considering a sound scene containing J sources, each one is emitting a signal s_j, and a recording system in the form of a microphone array of I microphones. The impulse response between the j^{th} source and the i^{th} microphone is denoted α_{ij}. The contribution of the j^{th} source in the i^{th} microphone is noted c_{ij}, and it is therefore at a time t given by:

$$c_{ij,t} = [\alpha_{ij} * s_j]_t = \sum_{\tau=0}^{N-1} \alpha_{ij,\tau} s_{j,t-\tau}, \tag{3.8}$$

where $*$ denotes the convolution product. In other words the contribution of a source in a microphone i is the sum of the direct path source signal, and its reflections. A graphic illustration of the contribution of a source in a microphone is given in Fig. 3.3.

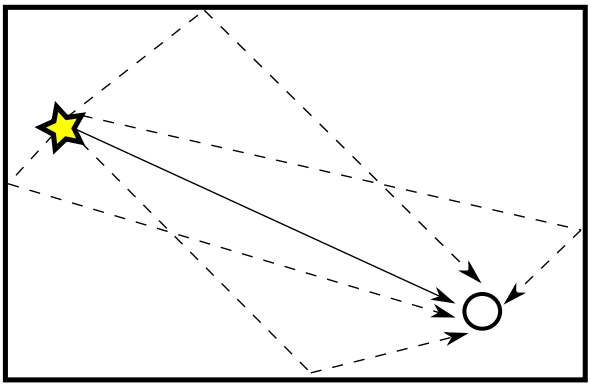


Fig. 3.3 Graphic illustration of the contribution of a source (yellow star) in a microphone (circle) in a reverberant environment

The i^{th} microphone signal x_i is the sum of all the contributions and therefore, it is given by:

$$x_{i,t} = \sum_{j=1}^{J} \sum_{\tau=0}^{N-1} \alpha_{ij,\tau} s_{j,t-\tau}. \tag{3.9}$$

The microphone array mixture is denoted $\mathbf{x}_t = [x_{i,t}]_{i=1\ldots I}$, it is a vector that contains the microphone signals. In the literature the mixture is written as a function of the sources contribution [22]:

$$\mathbf{x}_t = \sum_{j=1}^{J} \mathbf{c}_{j,t}, \tag{3.10}$$

where $\mathbf{c}_{j,t} = [c_{ij,t}]_{i=1\ldots I}$ is a vector that contains the contribution of the j^{th} source in each microphone.

Most sound source localization and separation techniques operate in the time-frequency (TF) domain (Chapter 3 Section 3.1). Under the narrow-band approximation,[2] and assuming the mixing filters are time invariant, the Short-Time Fourier Transform (STFT) of the microphone signals is given by:

$$\forall f \in [1, F], n \in [1, N], \quad \mathbf{x}_{f,n} = \sum_{j=1}^{J} \mathbf{c}_{j,f,n} = \mathbf{A}_f \mathbf{s}_{f,n}, \tag{3.11}$$

[2] Assuming that the mixing filters are short compared to the STFT window.

where f, and n denote the frequency bin and time-frame index, respectively. Thus, $\mathbf{A}_f \in \mathbb{C}^{I \times J}$ contains the frequency responses $a_{ij,f}$ of the filters $\alpha_{ij}(t)$, and embeds information on the sources DoA. Indeed, the frequency response $a_{ij,f}$ is a linear combination of the j^{th} source direct path steering vector (SV) and the ones corresponding to its reflections. Note that a SV represents the set of phase delays a plane wave experiences, evaluated at the antenna, which in other words contains an information about the time difference of arrival between the microphones and indirectly the DoA of a sound source. The SV of the p^{th} reflection of a source j recorded by an antenna of I microphones is given by:

$$\mathbf{a}_{j,p,f} = [k_{1,p,j} e^{-i2\pi f \tau_{1,p,j}} \dots k_{I,p,j} e^{-i2\pi f \tau_{I,p,j}}]^\top, \tag{3.12}$$

where on the one hand $k_{1,p,j}$ is coefficient that depends on the reflection coefficient and the distance that the wave travels to the i^{th} microphone. On the other hand $\tau_{I,p,j} = \frac{r_{i,p,j}}{c}$ is the delay of arrival to the i^{th} microphone with $r_{i,p,j}$ being the distance from the i^{th} microphone. The j^th column of the mixing matrix $\mathbf{A}_f$ is given by:

$$\mathbf{a}_{j,f} = \sum_{p=1}^{P} \mathbf{a}_{j,p,f}, \tag{3.13}$$

with P being the number of times the j^{th} sound source is reflected, and $p = 1$ corresponds to the direct path.

3.2.2 Mixture model in HOA domain

Let us consider a sound source j in a reverberant environment where it is reflected P times from P different directions (θ_{jp}, ϕ_{jp}) with $p \in [1, P]$. Note that $p = 1$ corresponds to the direct path. The ambisonic signals of a given reflection p are given by:

$$\mathbf{z}_{jp,t} = \mathbf{y}(\theta_{jp}, \phi_{jp}) s_{jp,t}, \tag{3.14}$$

where the signal $s_{jp,t}$ corresponds to the reflected signal $s_{j,1}$ from the direction (θ_{jp}, ϕ_{jp}). This signal is therefore delayed and attenuated up to a given time τ_{jp} and coefficient α_{jp}, respectively. This signal can be written in the time frequency domain as follows:

$$s_{jp,fn} = \alpha_{jp} e^{-i2\pi f \tau_{jp}} s_{j,fn}. \tag{3.15}$$

Given Eq. (3.15), we can write Eq. (3.14) in the time frequency domain as follows:

$$\mathbf{z}_{jp,fn} = \mathbf{y}(\theta_{jp}, \phi_{jp}) \alpha_{jp} e^{-i2\pi f \tau_{jp}} s_{j,fn}. \tag{3.16}$$

Considering all the reflections P of the sound source j, we can write the contribution the sound source j at each ambisonic channel $\mathbf{b}_{j,fn}$ as follows:

$$\mathbf{b}_{j,fn} = \sum_{p=1}^{P} \mathbf{y}(\theta_{jp}, \phi_{jp}) \alpha_{jp} e^{-i2\pi f \tau_{jp}} s_{j,fn} \tag{3.17}$$

$$= \mathbf{y}_{j,f} s_{j,fn}, \tag{3.18}$$

with $\mathbf{y}_{j,f} = \sum_{p=1}^{P} \mathbf{y}(\theta_{jp}, \phi_{jp}) \alpha_{jp} e^{-i2\pi f \tau_{jp}}$ is a frequency dependent composite vector.

Considering J sound sources the ambisonic signals are given by:

$$\forall f \in [1, F], n \in [1, N], \quad \mathbf{z}_{f,n} = \sum_{j=1}^{J} \mathbf{b}_{j,f,n} = \mathbf{Y}_f \mathbf{s}_{f,n}. \tag{3.19}$$

Similarly to $\mathbf{A}_f$ in the microphone domain $\mathbf{Y}_f \in \mathbb{C}^{M \times J}$ contains the frequency responses $y_{mj,f}$ of a time filter (explained later in the same section) that embeds information on the sources DoA. The frequency response $y_{mj,f}$ is indeed a linear combination of the j^{th} source direct path steering vector and the ones corresponding to its reflections.

There is a difference between both domains when it comes to the phase difference between channels or microphones when only one sound source is active. Indeed, On the one hand, as it is described before in the microphone domaine, the SV contains a phase difference between each microphone. On the other hand in the HOA domain there is no phase difference between the channels. We can see that with the expression of $\mathbf{y}_{j,f}$:

$$\mathbf{y}_{j,f} = \sum_{p=1}^{P} \mathbf{y}(\theta_{jp}, \phi_{jp}) \alpha_{jp} e^{-i2\pi f \tau_{jp}}, \tag{3.20}$$

Indeed, in the presence of only one sound source $\mathbf{y}(\theta_{jp}, \phi_{jp})$ is the spherical harmonic vector. it depends on the direction of arrival and not the time difference of arrival of. It's expression in the second order is given in Appendix A Eq. (A.15). Moreover, for a sound source j and a reflection p the expression $\alpha_{jp} e^{-i2\pi f \tau_{jp}}$ is similar at each channel.

3.3 Sound source localization

Several sound source localization techniques were developed and proposed in the sate of the art. Most of them were created first for narrowband signals, and they have been used for wideband cases using frequency domain analysis. We divide these sound source localization techniques into three different categories:

- Beamforming localization approaches.
- Subspace localization approaches.
- Time-frequency analysis techniques.

3.3.1 Beamforming localization approaches

In the beamforming localization approaches, we have techniques such as in such as Barlett and MVDR. These approaches are usually used as spatial filters in which the mixture content is used to minimize the power contributed by noise and undesired interference while maintaining a fixed gain in the look direction. The mixture content is exploited by using an estimation of the covariance matrix. An estimation of this matrix for a given frame corresponding to the sample $t \in [(k-1)T, \ldots, kT-1]$ under the hypothesis that the ambisonic signals have a zero mean, as follows:

$$\mathbf{\Gamma}_{in} = \mathbf{C}_k = \frac{1}{T} \sum_{(k-1)T}^{kT-1} \mathbf{z}_t \mathbf{z}_t^\top, \tag{3.21}$$

The DoAs can be estimated in the form of cartography by computing the power spectrum in each direction. For the direction (θ_0, ϕ_0) the power spectrum is given by:

$$P^{MVDR}_{[(k-1)T,\ldots,kT-1]}(\theta_0, \phi_0) = \frac{1}{\mathbf{y}(\theta_0, \phi_0)^\top . \mathbf{\Gamma}_{in}^{-1} . \mathbf{y}(\theta_0, \phi_0)}. \tag{3.22}$$

$$P^{Barlett}_{[(k-1)T,\ldots,kT-1]}(\theta_0, \phi_0) = \mathbf{y}(\theta_0, \phi_0)^\top . \mathbf{\Gamma}_{in} . \mathbf{y}(\theta_0, \phi_0). \tag{3.23}$$

The peaks occur whenever the SV is orthogonal to the noise subspace of the covariance matrix. For example for MVDR, the denominator in Eq. (3.22) represents a projection of the SV corresponding to the DoA (θ_0, ϕ_0) on the noise subspace. If ever (θ_0, ϕ_0) is a direction of arrival, the denominator will be a small value, which translates on the cartography as a peak and, therefore, as a DoA of a sound source.

Another beamforming approach was presented in [57]. It is based on the pseudointensity vectors. The key idea behind this approach is to compute the power corresponding to the output of a beamformer steered in different directions. The location with the highest power provides an estimate of the location of the sound source. Though this is a low-cost approach, it works for a single active sound source.

Although these approaches are simple, they are extremely vulnerable when the signal to noise ratio (SNR) varies at each frequency. Moreover, they offer a mediocre spatial resolution when the number of channels is small. Indeed, for instance, while using MVDR, the whole covariance matrix of the ambisonic signals is used with this

approach, the projection isn't only on the noise subspace. For these kinds of approaches, this problem can be solved with subspace localization approaches.

3.3.2 Subspace localization approaches

Unlike the approaches presented in Section 3.3.1, subspace localization approaches such as Esprit and MUSIC offer a higher spatial resolution. In the audio domain, MUSIC is heavily used. The key idea behind it is to project the potential SV only on the noise subspace generated by the eigenvectors corresponding to the smallest eigenvalues of the covariance matrix instead of the overall space. This is done by decomposing the covariance matrix using an eigendecomposition.

$$\mathbf{\Gamma}_{in} = \mathbf{U}\Lambda\mathbf{U}^{\top}, \tag{3.24}$$

where Λ is a diagonal matrix whose coefficients are the covariance's $\mathbf{\Gamma}_{in}$ eigenvalues, $\mathbf{U}$ is the matrix of the covariance eigenvectors.

The eigenvectors of the covariance matrix belong to either of two: the signal subspace or the noise subspace. The noise-subspace corresponds to the eigenvectors of the smallest eigenvalues. In the case where the noise is diffuse, the smallest eigenvalues are equal to each other, and they are identical to the noise variance.

Therefore to apply MUSIC, we first need to compute the eigenvalues and eigenvectors of the covariance matrix. Second, we observe the eigenvalues. If the number of sources is known J, the eigenvalues of the noise subspace can be identified. Indeed, if the eigenvalues are sorted in decreasing order, the eigenvectors corresponding to the J largest eigenvalues "generate" the signal subspace. The remaining $M - J$ "generate" the noise subspace, which are the identified eigenvalues. Otherwise, we need to make a decision based on the proximity of the smallest values. Third, we need to construct the noise subspace, which corresponds to the matrix $\mathbf{G}$ that contains the eigenvectors of the identified eigenvalues (Corresponding to the smallest eigenvalues). The spatial spectrum in the direction (θ_0, ϕ_0) is then computed as follows:

$$P^{MUSIC}_{[(k-1)T,\ldots,kT-1]}(\theta_0, \phi_0) = \frac{1}{\mathbf{y}^{\top}(\theta_0, \phi_0).\mathbf{G}_{[(k-1)T,\ldots,kT-1]}.\mathbf{G}^{\top}_{[(k-1)T,\ldots,kT-1]}.\mathbf{y}(\theta_0, \phi_0)}. \tag{3.25}$$

Using such approaches directly on a given time frame does not take into consideration several advantages about the type of the mixture and its properties that can be offered in the frequency domain.

3.3.3 Time-frequency analysis techniques

In the time-frequency domain, we can exploit the nonstationarity of speech and its sparsity properties. In general, we make the hypothesis that whether we have speech or other types of sound sources (that are statistically independent) that in some time-frequency bins, we can have only one active sound source with no significant contribution from room reflections. This hypothesis is known as the approximately W-Disjoint-Orthogonality (W-DO) hypothesis.

3.3.3.1 Time-frequency analysis techniques without weighing the importance of each time-frequency bin

In some sound source separation approaches such as Time-frequency masking the DoAs of the sound, sources are estimated under the made hypothesis. A popular method in the microphone domain is DUET which is a sound source separation approach, in which DoAs are estimated by analyzing each time-frequency bin. This provides potential directions by looking for the phase difference between two microphones. The possible directions are after clustered or used in a histogram to estimate the DoAs.

It seems that such approaches are not applicable to ambisonic mixture due to the fact that we can not use the phase difference between the ambisonic signals as an information when only one sound source is active. Indeed, for a given sound source, all the ambisonic signals have the same phase (see the last two paragraphs of Section 3.2.2). However, there is a way to adapt this approach to the ambisonic mixture to work similarly as in the microphone domain. In fact, we can use it to estimate the DoAs in terms of azimuth and elevation. We would need at least the first four channels. If we consider an anechoic environment, the mixing matrix $\mathbf{Y}_f$ in Eq. (3.19) can be simplified and considered frequency independent if we consider the phase shifting of the sound sources already in $\hat{\mathbf{s}}_{j,f,n}$ (this signals are attenuated and delayed by the propagation of $\mathbf{s}_{j,f,n}$), and therefore we can write the mixing matrix as follows:

$$\mathbf{Y} = \begin{pmatrix} 1 & \dots & 1 & \dots & 1 \\ cos(\theta_1)cos(\phi_1) & \dots & cos(\theta_j)cos(\phi_j) & \dots & cos(\theta_J)cos(\phi_J) \\ sin(\theta_1)cos(\phi_1) & \dots & sin(\theta_j)cos(\phi_j) & \dots & sin(\theta_J)cos(\phi_J) \\ sin(\phi_1) & \dots & sin(\phi_j) & \dots & sin(\phi_J) \end{pmatrix}, \tag{3.26}$$

where θ_j and ϕ_j denote the azimuth and the elevation of the sources. In the case of time-frequency bins in which we have one dominant sound source and with no significant contribution from room reflections, we can make the hypothesis that the composite vector in Eq. (3.20) is close to the direct path sound source and therefore, it looks like a column from the relative transfer mixing matrix $\mathbf{Y}$ in Eq. (3.26). Let the j^{th}

sound source be active at the time frequency bin (f, n), we can write the mixture $\mathbf{z}_{f,n}$ as follows:

$$\mathbf{z}_{f,n} = \hat{s}_{j,f,n} \begin{pmatrix} 1 \\ cos(\theta_j)cos(\phi_j) \\ sin(\theta_j)cos(\phi_j) \\ sin(\phi_j) \end{pmatrix}, \tag{3.27}$$

in such time-frequency bins, it is possible to recover the DoA of the sources. Alg. 1 can be used as an adaptation of the approach in [59] on ambisonic signals. There are some contribution that align with this work and with more constraint to have better estimation [78, 46, 48, 47]. These approaches are discussed in the Section 3.3.3.2.

Algorithm 1 Adaptation of The DoA as in DUET on ambisonic signals, under the **approximately W-DO Hypothesis**

Input $\mathbf{z}_t \quad \forall t \in \{1, 2, \ldots, T\} \quad J$ number of sound sources
Outputs $[\theta_j, \phi_j]_{j=1,\ldots,J}$

Perform a STFT on the ambisonic signals
$\theta = \varnothing$
for $n \leqslant N$ **do**
  for $f \leqslant F$ **do**
    $z_{frac,f,n} = ||\frac{1}{z_{1,f,n}} \begin{pmatrix} z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix}||,$
  end for
end for
if $z_{frac,f,n} = 1$ **then**
  $\theta_{fn} = \text{Arctg}(\frac{z_{3,f,n}}{z_{2,f,n}})$ and $\phi_{fn} = \text{Arcsin}(z_{4,f,n})$
  $\theta = \theta \cup [\theta_{f,n}, \phi_{f,n}]$
end if
- Design a histogram with all the potential DoA at each time-frequency bin. $H_{[\theta,\phi]} = \#\{[\theta, \phi]\}$ for $[\theta, \phi] \in \theta$
for $j \leqslant J$ **do**
  $[\theta_j, \phi_j] = \text{argmax}_{[\theta,\phi]}\{H_{[\theta,\phi]}\}$
  $H_\theta = H_{[\theta,\phi]} \backslash \max\{H_{[\theta,\phi]}\}$
end for

As you can see in Alg. 1, unlike in [59], the adaptation to the ambisonic sound field requires to look for the DoA as a pair of two angles (azimuth and elevation). This aspect might elevate the performance of the sound source localization in the ambisonic domain. Indeed, looking for a pair of angles may disregard false possibilities. An evaluation of this approach is going to be assessed in Chapter 4 Section 4.2 (The approach will be referred to as DUET).

Alg. 1 can be used in different ways. An interesting one deals with computing the vector of each time frequency bin:

$$\frac{1}{z_{1,f,n}} \begin{pmatrix} z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix}, \tag{3.28}$$

and cluster them into J clusters, with J being the number of sound source (beforehand known), and consider the centroids as DoA. This type of approach was subtly used in known ambisonic algorithms such as Directional Audio Coding (DirAC).

DirAC [99, 98, 119] is a method specially designed for ambisonic signals. It aims to improve the reproduction of 1^{st} order ambisonic signals. It is a technique that allows communicating a spatialized sound scene in each time-frequency bin with two parameters, which are the direction of the dominant sound source and the diffuseness. The overall method is called DirAC. However according to the articles [99, 98, 119], the part that determines the direction of arrivals is called "energetic analysis of the sound field". Similarly to the previously discussed approach (in the same section) this method operates in the time frequency domain as well. Let us consider the four first channels of the ambisonic signals in the time frequency domain:

$$\mathbf{z}_{f,n} = \begin{pmatrix} z_{1,f,n} \\ z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix}. \tag{3.29}$$

The analysis of the sound field is done by computing the instantaneous intensity vector $\mathbf{I}_{f,n}$:

$$\mathbf{I}_{f,n} = \Re(z_{1,f,n}[z_{2,f,n}\ z_{3,f,n}\ z_{4,f,n}]). \tag{3.30}$$

The DoAs are deduced from the instantaneous intensity vector $\mathbf{I}_{f,n}$. Consider the j^{th} sound source to be dominant in the time frequency bin (f, n), the estimation of the DoA is given by:

$$\mathbf{r}_{f,n} = \begin{pmatrix} cos(\theta_j)cos(\phi_j) \\ sin(\theta_j)cos(\phi_j) \\ sin(\phi_j) \end{pmatrix} = -\frac{\mathbf{I}_{f,n}}{||\mathbf{I}_{f,n}||}. \tag{3.31}$$

This method estimates along the DoAs, the diffuseness[3] at each time-frequency bin. This parameter computes the amount of diffuseness in a time-frequency bin. It is a value between zero (only one source is active) and one (the sound field is diffuse). It is given by:

$$\psi_{f,n} = \sqrt{1 - \frac{||\mathbf{I}_{f,n}||}{|z_{1,f,n}| + |||[z^2_{2_f,n}, z^2_{3,f,n} z^2_{4,f,n}]|||}}. \tag{3.32}$$

If we use only the DoA in Eq. (3.31), the approach will be similar to the one in Alg. 1.

There is another approach in the ambisonic domain that aims to sharpen first-order ambisonic sound scenes that can be used to estimates the DoA. This approach is known as High Angular Resolution Plane Wave Expansion (HARPEX) [12, 11]. The idea behind this approach is to decompose at each time-frequency bin the ambisonic signals into two plane waves and estimate the SV of each plane wave. This approach can be useful if we have two dominant sources in a time-frequency bin.

Consider the vector of first order ambisonic signals for time-frequency bin (f,n):

$$\mathbf{z}_{f,n} = \begin{pmatrix} z_{1,f,n} \\ z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix}. \tag{3.33}$$

The first-order ambisonic signals corresponding to a plane wave coming from the direction (θ, ϕ), and emitting a signal s is given by Eq. (3.19), with $\mathbf{y}_f = \mathbf{y}(\theta, \phi) = [1, \cos(\theta)\cos(\phi), \sin(\theta)\cos(\phi), \sin(\phi)]^\top$, and $\mathbf{s}_{f,n} = s_{f,n}$. Note that the vector $\mathbf{y}_f$ does not depend on the frequency because the delay and the attenuation propagations are already considered in the signal $s_{f,n}$. In other words $s_{f,n}$ contains the value of the signals at the position of the spherical microphone array and not at it starting position. The first-order ambisonic signals resulting from two plane waves incoming from directions (θ_1, ϕ_1), (θ_2, ϕ_2), and emitting respectively the signals s_1 and s_2 as follows:

$$\mathbf{z}_{f,n} = \begin{pmatrix} z_{1,f,n} \\ z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \cos(\theta_1)\cos(\phi_1) & \cos(\theta_2)\cos(\phi_2) \\ \sin(\theta_1)\cos(\phi_1) & \sin(\theta_2)\cos(\phi_2) \\ \sin(\phi_1) & \sin(\phi_2) \end{pmatrix} . \begin{pmatrix} s_{1,f,n} \\ s_{2,f,n} \end{pmatrix}. \tag{3.34}$$

In order to estimate the plane-wave signals s_1 and s_2, the following steps have to be executed :

[3] A value that quantify the diffuseness of an ambisonic sound field.

- Split the vector in Eq. (3.33) into 2 vectors; a real one and an imaginary one :

$$\mathbf{z}_{f,n} = \begin{pmatrix} z_{1,f,n} \\ z_{2,f,n} \\ z_{3,f,n} \\ z_{4,f,n} \end{pmatrix} = \begin{pmatrix} \mathrm{Re}(z_{1,f,n}) & \mathrm{Im}(z_{1,f,n}) \\ \mathrm{Re}(z_{2,f,n}) & \mathrm{Im}(z_{2,f,n}) \\ \mathrm{Re}(z_{3,f,n}) & \mathrm{Im}(z_{3,f,n}) \\ \mathrm{Re}(z_{4,f,n}) & \mathrm{Im}(z_{4,f,n}) \end{pmatrix} . \begin{pmatrix} 1 \\ i \end{pmatrix} . \tag{3.35}$$

- Apply a QR transform on the splitted matrix :

$$\begin{pmatrix} \mathrm{Re}(z_{1,f,n}) & \mathrm{Im}(z_{1,f,n}) \\ \mathrm{Re}(z_{2,f,n}) & \mathrm{Im}(z_{2,f,n}) \\ \mathrm{Re}(z_{3,f,n}) & \mathrm{Im}(z_{3,f,n}) \\ \mathrm{Re}(z_{4,f,n}) & \mathrm{Im}(z_{4,f,n}) \end{pmatrix} = \mathbf{QR}, \tag{3.36}$$

  where $\mathbf{Q}$ is a 4 by 2 matrix and $\mathbf{R}$ is a 2 by 2 matrix. Therefore, at each time-frequency bin the first-order ambisonic signals will be written as follows:

$$\mathbf{z}_{f,n} = \mathbf{QR} \begin{pmatrix} 1 \\ i \end{pmatrix} . \tag{3.37}$$

  Matrix $\mathbf{Q}$ is already a 4 by 2 matrix. Eq. (3.37) would be considered as the decomposition in Eq. (3.34) only if :

$$Q_{11} = Q_{12} = 1 \tag{3.38}$$
$$Q_{21}^2 + Q_{31}^2 + Q_{41}^2 = 1 \tag{3.39}$$
$$Q_{22}^2 + Q_{32}^2 + Q_{42}^2 = 1. \tag{3.40}$$

- Transform matrix $\mathbf{Q}$ into a matrix $\mathbf{D}$ that respect the above conditions :

$$b = \sqrt{2(Q_{12}^2 + Q_{12}^2)^2 - 1} \tag{3.41}$$
$$\mathbf{C} = Q_{11} \begin{pmatrix} 1 & 1 \\ -b & b \end{pmatrix} + Q_{12} \begin{pmatrix} b & -b \\ 1 & 1 \end{pmatrix} \tag{3.42}$$
$$\mathbf{D} = \mathbf{QC}. \tag{3.43}$$

Once the above conditions are met, the signals s_1 and s_2 and their SV are estimated

as follows:

$$\mathbf{z}_{f,n} = \mathbf{QR}\begin{pmatrix}1\\ i\end{pmatrix} = \mathbf{QCC}^{-1}\mathbf{R}\begin{pmatrix}1\\ i\end{pmatrix} = \mathbf{DC}^{-1}\mathbf{R}\begin{pmatrix}1\\ i\end{pmatrix}, \tag{3.44}$$

which makes $\mathbf{D} = \mathbf{QC}$ the matrix that contains the SV and the vector $\mathbf{C}^{-1}\mathbf{R}[1, i]^\top$ as the estimation of the signals s_1 and s_2.

Alg. 2 is used to exploit HARPEX for the estimation of the DoAs.

Algorithm 2 Exploit HARPEX to determine the DoA

Input $\mathbf{x}_t \quad \forall t \in \{1, 2, \ldots, T\}$, J number of sound sources
Outputs $[\theta_j, \phi_j]_{j=1,\ldots,J}$
Perform a STFT on the mixture
Initialize S =Null
for $n <= N$ **do**
  for $f <= F$ **do**
    QR decomposition of the bin(f,n)
    Compute **D**
)
  end for
end for
Apply k-means algorithm on the set S, the number of clusters must be greater than or equal the number of sources
The centroids are now potential directions

The problem of such approaches resides in the fact that some time-frequency bins can introduce wrong directions. To overcome this problem, we need to set a test and weight somehow the reliability of each time-frequency bin. As one can imagine also, these algorithms can be very sensitive to the length of the RIR. Having a longer RIR can effect the narrow band approximation hypothesis and therefore the approximately W-DO hypothesis.

3.3.3.2 Time-frequency analysis techniques with weighing the importance of each time-frequency bin

Under the approximately W-DO hypothesis, as long as there are bins (f, n) where only one source is dominant, there are other bins where it is not the case. Therefore, the obstacle dwells on knowing which bins are reliable. Another problem when it comes to source localization consists in identifying the number of sources in the sound field, which is difficult given only the mixture.

The Direction Estimation of the Mixing matriX (DEMIX) algorithm was proposed in [7] in order to solve the above-listed problems. The main idea of the algorithm is

to estimate two values for each time-frequency bin by taking neighboring bins into consideration. Assuming that each time-frequency bin has a time-frequency region $\Omega_{f,n}$, we can estimate two parameters:

- $\hat{\mathbf{a}}(\Omega_{f,n})$ the SV of the most dominant source in the treated time-frequency bin.
- The local confidence measure $\mathcal{T}(\Omega_{f,n})$, a value that can discriminate bins where more than one source is dominant.

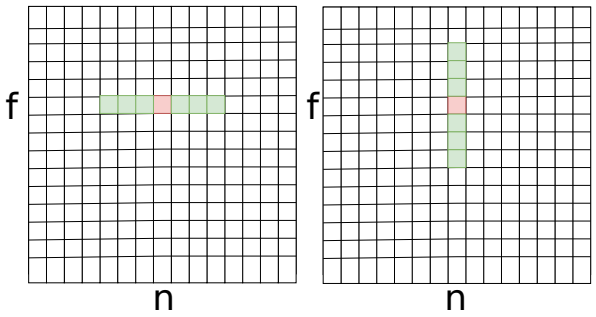


Fig. 3.4 Example of neighbors of a time frequency bin (f,n). left is $\Omega^N_{f,n}$ and right is $\Omega^F_{f,n}$, with $K = 3$.

The neighbors of a time frequency bin (f, n) are defined as follows:

$$\Omega^N_{f,n} = \{f, n + k \mid |k| \leqslant K\} \tag{3.45}$$

$$\Omega^F_{f,n} = \{f + k, n \mid |k| \leqslant K\}, \tag{3.46}$$

where $K \in \mathbb{N}$ is a chosen number (from more information refer to [7]). An illustration of the neighbors is given in Fig. 3.4. Each region Ω provides a complex-valued local scatter plot $\mathbf{X}(\Omega)$. It is a $(I \times (2K + 1))$ matrix that columns are $X(\tau, \omega) \in \mathbb{C}^I$ with $(\tau, \omega) \in \Omega$. I represents the number of sensors. Next, a Principal Component Analysis (PCA) decomposition should be applied to the matrix $\mathbf{X}(\Omega)$ for each time-frequency bin:

- The estimate of $\hat{\mathbf{a}}(\Omega_{f,n})$ the SV is the Principal Component (PC) of $\mathbf{X}(\Omega)$. This is done using an singular value decomposition on the local covaraince matrix $\mathbf{X}(\Omega)$.
- The local confidence measure $\mathcal{T}(\Omega_{f,n})$ is presented as follows :

$$\mathcal{T}(\Omega_{f,n}) = \hat{\lambda_1}(\Omega) / \frac{1}{I-1} \sum_{i=2}^{I} \hat{\lambda_i}(\Omega), \tag{3.47}$$

where $\hat{\lambda}_i$ denotes the i-th eigenvalue of the performed PCA, the eigenvalues being sorted in decreasing order $\hat{\lambda_1} \geqslant \hat{\lambda_2} \geqslant ... \geqslant \hat{\lambda_I}$. The local confidence measure of a time-frequency bin is a value that represents the proportion of the variance. A higher value is a sign that the treated time-frequency bin does not contain several decorrelated sound sources. In [7], the authors proposed to weight the channels of the local SV by the local confidence and plot the first channel as a function of the second one.

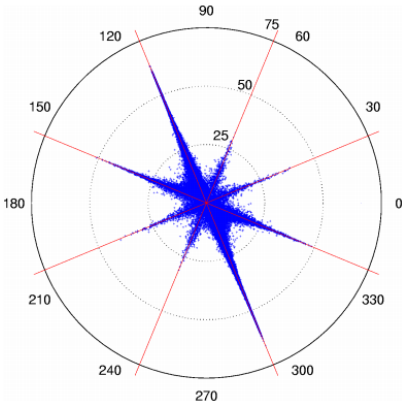


Fig. 3.5 Example of a scatter plot of points $\hat{\mathbf{a}}(\Omega_{f,n})$ weighted by their confidence measure [7].

Thus the points that represent the DoA are visually far from the other ones and along the DoAs, as it is shown in Fig. 3.5.

DEMIX approach comes in the form of two algorithms :

- Cluster creation
- Direction estimation

The number of clusters is an estimation of the number of sources.

In order to adapt DEMIX to ambisonic mixtures, we should incorporate at least the first four channels. The adapted algorithms are similar to the classic case, except the scatter plot of the points are going to be in three dimensions. Indeed, the DoA in the HOA is given by the azimuth and the elevation, unlike in the microphone domain. Similarly to the usual algorithm, the principal component will be considered as the SV of the treated time-frequency bin. The singular value decomposition of a time-frequency bin neighborhood is given as follows:

$$\mathbf{Z}(\Omega_{f,n}) = \mathbf{USV}^{\top}, \tag{3.48}$$

where $\mathbf{Z}(\Omega_{n,f}) \in \mathbb{C}^{4,2K+1}$ is the matrix that contains the neighbors of $\mathbf{z}_{f,n}$. The matrix

Algorithm 3 DEMIX, Cluster creation

Input $\mathbf{x}_t \quad \forall t \in \{1, 2, \ldots, T\}$

Outputs C_k the number of clusters K is determined in the algorithm and depends on a given threshold

Perform a STFT on the mixture
for $n \leqslant N$ **do**
  for $f \leqslant F$ **do**
    Compute $\hat{\mathbf{a}}(\Omega_{f,n})$ for each time frequency bin. It embedded information about DoA if only one sound source is active.
    Compute $\mathcal{T}(\Omega_{f,n})$. This value provides relative information about the activity of the sound sources in the treated time-frequency bin.
  end for
end for
$\forall f, \forall n \quad P = \{\hat{\mathbf{a}}(\Omega_{f,n})\}$
initialize $k = 0, P_k = P_0 = P$
while $P_k \neq 0$ **do**
  $\Omega_k = \text{argmax}_{\Omega \in P_k} \mathcal{T}(\Omega_{f,n})$
  Create a cluster C_k with all region $\Omega \in P$ with which $\hat{\mathbf{a}}(\Omega)$ is close to $\hat{\mathbf{a}}(\Omega_k)$. The proximity is computed by the distance $\text{d}(|\hat{\mathbf{a}}(\Omega)|, |\hat{\mathbf{a}}(\Omega_k)|)$ and judged by a chosen threshold ζ
  Update $P_{k+1} = P_k \backslash C_k$
  $k = k + 1$
  $K = k$
end while

Algorithm 4 DEMIX, Direction estimation

Input $C_k \quad \forall k \in \{1, 2, \ldots, K\}$

Outputs $\hat{\mathbf{a}}_j$

determine the number of Clusters N
for $k \leqslant N$ **do**
  Determine a confidence threshold : $\eta_k = \max_{\Omega \in C_k \cap [\cup_{j \neq k} C_j]} \mathcal{T}(\Omega)$
  Keep in the cluster only the regions with highest empirical confidence values : $C'_k = \{\Omega \in C_k | \mathcal{T}(\Omega) \geqslant \eta_k\}$
  Estimate the centroid $\hat{\mathbf{a}}(C'_k)$
end for

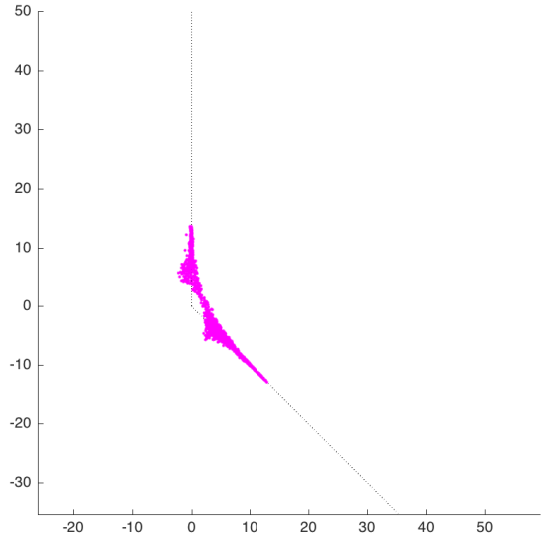


Fig. 3.6 Example of a scatter plot of points $\hat{\mathbf{y}}(\Omega_{t,f})$ weighted by their confidence measure for an ambisonic mixture. The representation is in 3D but represented as the upper view XoY.

S contains the singular values:

$$\mathbf{S} = \begin{pmatrix} \hat{\lambda_1} & 0 & 0 & 0 & \dots & 0 \\ 0 & \hat{\lambda_2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \hat{\lambda_3} & 0 & \dots & 0 \\ 0 & 0 & 0 & \hat{\lambda_4} & \dots & 0 \end{pmatrix}. \tag{3.49}$$

The singular values are sorted in decreasing order $\hat{\lambda_1} \geqslant ... \geqslant \hat{\lambda_4}$. The estimate of the local confidence measure is given by Eq. (3.47).

The estimate of the SV is the singular vector corresponding to the largest singular value :

$$\hat{\mathbf{y}}(\Omega_{n,f}) = \mathbf{u}_1. \tag{3.50}$$

Since $\hat{\mathbf{y}}(\Omega_{n,f}) = \mathbf{u}_1 = [u_{1m}]_{m=1...4}$ is considered as the estimate of the SV, in order to have a visual representation as in [7], we plot in 3D u_{12}.sign(u_{11}) according to u_{13}.sign(u_{11}) and u_{14}.sign(u_{11}), where sign denotes the sign function. An example of visual representation is given in Fig. 3.6.

An evaluation of this approach is going to be assessed in Chapter 4 Section 4.2. When this approach was tested in Chapter 4 Section 4.2, we considered the number of sound sources known and theirfore, we didn't use the same method of clustering as in

Alg. 3. We used k-means clustering instead.

A similar approach was proposed for HOA mixtures in [78]. The authors suggest in the same manner to construct new local time-frequency bins covariance matrices using time neighbors and conserve only reliable bins using what they refer to as direct path dominance test. Similarly to DEMIX, this test identifies time-frequency bins in which only one sound source is dominant and with no significant contribution from room reflections. The selection of reliable time-frequency bins is given by identifying rank-1 covariance matrices. One way to do that is computing the eigenvectors and using the local confidence Eq. (3.47). Instead of clustering the possible directions, they propose to present the results in cartography using the MUSIC spatial spectrum. To do that, they recommend "fusing" the selected time-frequency bins information to construct the overall spatial spectrum. One way to "fuse" these time-frequency bins is to sum the spatial spectrums of all the chosen time-frequency bins. For more information, please check [78].

Some other approaches have emerged since late 2017. These contributions align with this approach and complement it further. The idea behind these contributions is to find a better way to cluster the potential directions. Sadly, we were not aware of them when we assessed the localization approaches through numerical experiments, so we didn't experiment with them. The first approach is known as MSEC weighting [48], in which an adaptative k-means clustering is used. To each time-frequency bin, two parameters are computed being:

- The cluster weight, which represents a time-frequency bin the normalized measure of concentration in its associated cluster.

- The member weight, which represents for a time-frequency bin, is the normalized measure of closeness to its associated centroid.

These parameters are used to compute each time-frequency bin's MSEC weight, and only the Time-frequency bins with the strongest weights are preserved. DoAs are deduced using a histogram. For more information, please check [48].

The second approach is known as DBSCAN clustering [46, 47], which can be used as an extension for a better clustering. This approach is based on what they call local density metric, which is defined as the number of points within the neighborhood of small specified distance, and a threshold density. The points with a higher density than the threshold are labeled score points and are grouped using density connectivity. For more information about this approach, please check [46, 47].

3.3.4 Conclusion on sound source localization

We can say that we have several algorithms for sound source localization. Most of these algorithms require to know the number of sound sources in the mixture. Note that the main goal of my Ph.D. work is to use my research in order to be able to create content where a user can move visually and audibly with 6-DoF in a recorded environment. Therefore, cameras must be used to capture the environment visually. In this case, the cameras can be used to recover the number of sound sources in the environment and even have an idea about their DoAs.

One of these algorithms can be used to locate the sound sources for two reasons:

- Help the sound source separation by giving the location of the sound sources as additional information.
- Estimate the phase difference and the magnitude to be applied to the sound source signals in order to estimate the values of these signals at the newest user position.

These algorithms are going to be assessed through some numerical experiments in order to compare them and recommend one or several of them to use. This will be studied in Chapter 4 Section 4.2.

Note that we will not conduct any more research on this subject, judging that we have enough algorithms to use for our main goal.

In Section 3.4, we explain the sound separation problem, and survey some approaches that can be used in the ambisonic domain.

3.4 Sound source separation

The source separation problem consists in recovering the spatial images of the sources $\mathbf{c}_j$ from the microphone mixture $\mathbf{x}$ or $\mathbf{b}_j$ from the ambisonic mixture $\mathbf{z}$. The sound sources s_j are recovered by dereverberating the spatial images of the sources $\mathbf{c}_j$ in the microphone domain or $\mathbf{b}_j$ in the ambisonic domain, which involves a deconvolution with the filter corresponding to the room impulse responses. This problem will not be considered in this thesis. A beamforming can estimate the single-channel source signals s_j.

In the time frequency domain we wrote the microphone or ambisonic mixing model Eq. ((3.11)(3.19)), respectively, in the form of linear equation system:

$$\forall f \in [1, F], n \in [1, N], \quad \mathbf{x}_{f,n} = \mathbf{A}_f \mathbf{s}_{f,n} \tag{3.51}$$

$$\mathbf{z}_{f,n} = \mathbf{Y}_f \mathbf{s}_{f,n}. \tag{3.52}$$

It seems that we can recover the single-channel source signals s_j at each time-frequency bin (f, n) if the mixing matrix is known $\mathbf{A}_f$ in the microphone domain, and $\mathbf{Y}_f$ in the ambisonic domain, by simple inversion. However, several difficulties occur while trying to solve the problem:

- First, blind sound source separation consists in solving the problem without having the mixing matrix.
- Second, even when the mixing matrix is known, the problem has a solution if the mixing matrix is square and is full rank (the number of microphone/channels must be equal to the number of sources $I = J$ (microphone) or $M = J$ (ambisonics). When the problem is underdetermined or over determined ($I < J, M < J$ or $I > J, M > J$) the problem is more challenging and "ill-posed", and can be solved by supposing more hypothesis, adding some constraints, or rely on singular value decomposition of the demixing matrix in the case of overdetermined problems.
- Third, the type of the filter can add more difficulties. There exist different types of mixing filters; instantaneous mixture (no delay), anechoic mixture (delay), real-life mixture (corresponding to a convolution with a finite impulse response filter).
- Fourth, the sound source may be static, which corresponds to a time-invariant mixing matrix, or dynamic (moving sources), which corresponds to a time-variant mixing matrix.

For simplicity, we consider only static sources. In the case of dynamic sound sources, we suggest applying what we propose on small frames of the mixture.

The performance of a sound source separation algorithm can be judged by listening to the outcome of the separation and check if the sounds were well separated with no interference, less distortions and less artifacts. However, you can imagine how much time such a task would take on many examples. Therefore an objective measure is needed. In the next subsection we discuss some objective measures that are considered as a reference in the sound source separation community. Then, we will describe state-of-the art sound source separation algos.

3.4.1 Objective evaluation of sound source separation

Vincent et *al.* [124] proposed an evaluation method that helps to measure different type of errors that affect the separated sources. The separation community highly adopts this method, and this is the reason why this method was selected to evaluate the performance of the different approaches discussed throughout this thesis.

The method consists in decomposing each separated source $\hat{s}_j$ as the following sum:

$$\hat{s}_j = s_{target} + e_{noise} + e_{interf} + e_{artif}, \tag{3.53}$$

where s_{target} is an allowed distortion of the target source s_j, e_{noise} , e_{interf}, and e_{artif} represent respectively the noise, interference and artifacts error terms. These errors are computed by least-square projection of the estimated source $\hat{s}_j$ onto the corresponding subspaces [124].

An orthogonal projector onto the subspace spanned by the vectors $y_1, ..., y_k$ is denoted $\prod\{y_1, ..., y_k\}$, the projector is a $T \times T$ matrix, with T being the length of the vectors $y_1, ..., y_k$. The considered orthogonal projectors in order to estimate the terms in Eq. (3.53) are given by:

$$\begin{aligned} P_{s_j} &= \prod\{s_j\}, & (3.54)\\ P_{\mathbf{s}} &= \prod\{(s_{j'})_{1\leqslant j'\leqslant J}\}, & (3.55)\\ P_{\mathbf{s},\mathbf{n}} &= \prod\{(s_{j'})_{1\leqslant j'\leqslant J}, (n_i)_{1\leqslant i\leqslant I}\}, & (3.56) \end{aligned}$$

where J and I denote respectively the number of sources, and the number of microphones/channels. n_i represents the noise in the microphone/channel i. The terms in Eq. (3.53) are then computed as follows[124]:

$$\begin{aligned} s_{target} &= P_{s_j}\hat{s}_j, & (3.57)\\ e_{interf} &= P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j, & (3.58)\\ e_{noise} &= P_{\mathbf{s},\mathbf{n}}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j, & (3.59)\\ e_{artif} &= \hat{s}_j - P_{\mathbf{s},\mathbf{n}}\hat{s}_j. & (3.60) \end{aligned}$$

After decomposing the estimated source s_j the following objective measures can be computed as energy ratio criteria in decibels (dB):

- Source to Distortion Ratio (SDR):

$$SDR = 10\log_{10}\frac{||s_{target}||^2}{||e_{noise} + e_{interf} + e_{artif}||^2}. \tag{3.61}$$

- Source to Interference Ratio (SIR):

$$SIR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2}. \tag{3.62}$$

- Source to Artifacts Ratio (SAR):

$$SAR = 10\log_{10}\frac{||s_{target} + e_{noise} + e_{interf}||^2}{||e_{artif}||^2}. \tag{3.63}$$

- Source to Noise Ratio (SNR):

$$SNR = 10\log_{10}\frac{||s_{target} + e_{interf}||^2}{||e_{noise}||^2}. \tag{3.64}$$

Note that the larger these measures are, the better the performance of the source separation is. Indeed the larger these measures are the smaller the denominators are, which means the lower the errors are. This can be interpreted for SDR, SIR, SAR, and SNR, respectively, as less distortion, interference, artifacts, and noise in the estimated sound source signal $\hat{s}_j$.

Having the objective measures, we can now tackle the sound source separation problem. In the rest of this chapter we survey some methods about this subject.

3.4.2 Time-frequency masking

Time-frequency masking is widely used for source separation in the case of under-determined mixtures. It was introduced first as a sound source separation solution for single microphone mixtures. As the name of the technique implies, it operates in the time-frequency domain. Indeed the technique profits from the assumption of the disjoint-orthogonality in the TF domain. Time-frequency masking was heavily recommended for the separation of speech from speech-in-noise mixtures such as in [131]. TF algorithms are based on computational auditory scene analysis (CASA) [19].

The time frequency source image s_{ij} (the contribution of the source j at the microphone i) is estimated from the mixture at the microphone i by:

$$s_{i,j,f,n} = m_{i,j,f,n}x_{i,f,n}, \tag{3.65}$$

with $m_{i,j,f,n}$ being a coefficient of the mask matrix $\mathbf{M}_{i,j}$ dedicated to the i^{th} microphone and the j^{th} sound source.

Oracle masks are TF masks that are deduced from the true contribution of the sound sources. There are two types of oracle masks. They are going to be used later in Chapter 4 Section 4.3.1.1 as benchmarks.

Oracle Soft Mask (OSM): also known as Ideal ratio Mask (IM) or the single-channel Wiener filter, with $0 \leqslant m_{i,j,f,n} \leqslant 1$. It is given by:

$$m_{i,j,f,n} = \frac{|s_{i,j,f,n}|^2}{\sum_{j'=1}^{J} |s_{i,j',f,n}|^2}. \tag{3.66}$$

Oracle Binary Mask (OBM): $m_{i,j,f,n}$ is whether equal to 0 or 1. They are deduced as follows:

$$m_{i,j,f,n} = \begin{cases} 1 & \text{if} \quad 20\log_{10} \frac{|s_{i,j,f,n}|^2}{\sum_{j'=1,j'\neq j}^{J} |s_{i,j',f,n}|^2}) \geqslant \eta \\ 0 & \text{otherwise} \end{cases}, \tag{3.67}$$

where $\eta \in]0, 1[$.

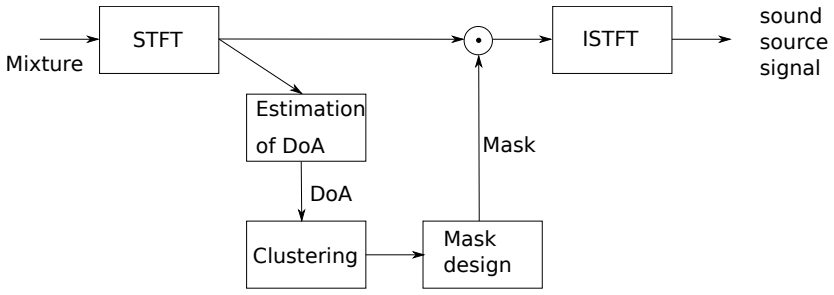


Fig. 3.7 Example of a time frequency masking algorithm [132].

Several algorithms [131, 132, 102, 65, 6, 5, 66] were proposed in the literature to estimate the OBMs. DUET and MENUET [132, 6] are examples. We introduced these algorithms as sound source localization techniques, but in reality, they are sound source separation techniques based on binary masking, they estimate the sound sources DoA in their process.

In the case of DUET and MENUET, the clustering process consists in identifying the time-frequency bin in which only one source is active. The mask design is based on a binary mask approach. Each time-frequency bin corresponding to the dominant sound source are set to 1, the rest to 0.

Such an algorithm can be adapted to the ambisonics domain. The only difference resides in estimating the DoA.

3.4.3 Local Gaussian approach in the microphone domain

This approach consists also in estimating the contribution $\mathbf{c}_{j,t} \in \mathbb{R}^I$ of each source $j = 1, ..., J$ in each microphone $i = 1, ..., I$ and at each time instant $t = 1, ..., T$ while relying on spatial and spectral cues [125, 35], unlike beamforming techniques

and TF approaches. This approach was first used on instantaneous mixtures in [37, 123]. Note that it is possible to perform a multicahnnel sound source separation and dereverberation using a Wiener filter and have the same type of output as beamforming.

The sources contributions $\mathbf{c}_j$ as defined in Eq.(3.10) can be addressed using the multichannel Wiener filtering framework (MWF). This framework requires the selection of a distribution model for the variables to estimate. In [123], the authors recommend using the local Gaussian model which is described as follows:

$$\forall f \in [1, F]\,, n \in [1, N]\,, \quad \mathbf{c}_{j,f,n} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}})\,, \tag{3.68}$$

where $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = \mathbb{E}\left[\mathbf{c}_{j,f,n}\,\mathbf{c}_{j,f,n}^{\mathsf{H}}\right]$ is the covariance matrix of the contribution of the j source to every microphone at frequency f and time frame n. In line with the literature, this matrix can be further decomposed as the product of a scalar spectral part, $v_{j,f,n}$, with a time-invariant spatial matrix, $\mathbf{R}_{\mathbf{c}_{j,f}}$ [35], as follows: $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = v_{j,f,n}\,\mathbf{R}_{\mathbf{c}_{j,f}}$. Notably, the so-called spatial covariance matrix $\mathbf{R}_{\mathbf{c}_{j,f}}$ respects the relation $\mathbf{R}_{\mathbf{c}_{j,f}} = \mathbf{A}_{j,f}\,\mathbf{A}_{j,f}^{\mathsf{H}}$ when the assumptions of Eq. (3.11) hold.

The multi-channel source separation problem can be solved by looking for the filter that minimizes the expected squared error for every source j and every time frequency bin (f, n):

$$\begin{aligned} &\forall j \in [1, J], f \in [1, F] \text{ and } n \in [1, N], \\ &\mathbf{W}_{j,f,n} = \underset{\mathbf{W}}{\operatorname{argmin}}\, \mathbb{E}\left[\|\mathbf{c}_{j,f,n} - \mathbf{W}\,\mathbf{x}_{f,n}\|_2^2\right]\,. \end{aligned} \tag{3.69}$$

The filter $\mathbf{W}_{j,f,n}$ is known as the multichannel Wiener filter (MWF) and is given by:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{(\mathbf{c}_{j,f,n}, \mathbf{x}_{f,n})}\,\mathbf{\Sigma}^{-1}_{(\mathbf{x}_{f,n}, \mathbf{x}_{f,n})}\,, \tag{3.70}$$

where the matrices $\mathbf{\Sigma}_{(\mathbf{x}_{f,n}, \mathbf{x}_{f,n})}$ and $\mathbf{\Sigma}_{(\mathbf{c}_{j,f,n}, \mathbf{x}_{f,n})}$, represent the covariance of the mixture $\mathbf{x}_{f,n}$ and the cross-correlation between the vectors $\mathbf{c}_{j,f,n}$ and $\mathbf{x}_{f,n}$, respectively.

From Eq. (3.68), and assuming the sources are statistically independent, the Wiener filter can be simplified as:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} \left(\sum_{j'=1}^{J} \mathbf{\Sigma}_{\mathbf{c}_{j',f,n}}\right)^{-1}\,. \tag{3.71}$$

Thus, the source separation problem reduces to the problem of estimating the covariance matrices $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}$. Each source contribution is obtained by applying element-wise its corresponding Wiener filter to the mixture: $\hat{\mathbf{c}}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{x}_{fn}$, and finally using

inverse STFT with overlap-add to reconstruct the time-domain signal.

There is an off the shelf toolbox that is based on this approach called the flexible audio source separation toolbox (FASST) [103, 90]. It is a software toolbox which estimates the Wiener filter parameters as well as applies it to estimates the contribution of the sound sources in each microphone. In FASST the parameters are estimated by maximizing the log-likelihood of the observations with an Expectation-Maximization (EM) algorithm [32, 23, 8, 55], and a multichannel non negative matrix factorization (NMF) model can be enforced on the source covariances $\mathbf{\Sigma}_{c_{j,f,n}}$.

This approach is studied in more details for ambisonics in Chapter 5.

3.4.4 Plane wave decomposition (beamforming)

In the context of sound source separation, there are also plane wave decomposition approaches, with which we can estimate a signal coming from a specific direction using beamforming [69, 91, 92]. In contrary to the above approach (MWF), Beamforming approaches are purely based on spatial cues, which requires to know only the direction of the signal we want to extract. Unlike MWF, where the sound source separation is based on both spatial and spectral cues. There are two different categories:

- Approaches based on applying the beam directly.
- Approaches based on exploiting the mixture content.

3.4.4.1 Approaches based on applying the beam directly

There are several plane wave decomposition approaches of this kind:

- First, we consider **basic projection or better known as the matched filter** [101, 56],[4] in which the ambisonic signals are projected on the spherical harmonic vector corresponding to the DoA of the desired sound, it is given by:

$$\hat{s}_{j,t} = \frac{\mathbf{y}(\theta_j, \phi_j)^\top}{||\mathbf{y}(\theta_j, \phi_j)||^2}\mathbf{z}_t. \tag{3.72}$$

  Eq. (3.76) is applied for each identified sound object. Although it gives an estimation of the desired signal, it does not set a constraint on interfering sound sources. In the case where the sound sources are close to each other, we may get interference in the estimated signal. For our application, we need to avoid interfering sound objects while extracting each of them.

[4] This beamformer will be referred to as basic projection or PWD in this manuscript.

To explain and showcase how beamforming works in the HOA domain. Let us consider an example where we apply this first beamformer in both time and TF domains. Let us consider an ambisonic sound field with J sound source. Each source is reflected P-1 times. In the time domain the representation of the mixture is given by Eq. (2.8). This representation can be misleading where each sound source and each reflection is considered separetly in the vector $\mathbf{s_t}$, and the matrix $\mathbf{Y}$ contains the spehrical harmonic vector of each sound source and each reflection. To be perfectly clear the sound source $\mathbf{s}_t$is given as follows:

$$\mathbf{s}_t = \left[s_1, s_{1,1}, \ldots, s_{1,P-1}, s_2, s_{2,1}, \ldots, s_{2,P-1}, \ldots, s_{J,P-1}\right]_t^\top, \tag{3.73}$$

with $s_{j,p}$ is the magnitude of the p^{th} reflection of the j^{th} sound source **at the position of the microphone array**. The matrix $\mathbf{Y}$ is given as follows:

$$\begin{aligned}\mathbf{Y} =&[\mathbf{y}(\theta_1, \phi_1), \mathbf{y}(\theta_{1,1}, \phi_{1,1}), \ldots, \mathbf{y}(\theta_{1,P-1}, \phi_{1,P-1}),\\ &\mathbf{y}(\theta_2, \phi_2), \mathbf{y}(\theta_{2,1}, \phi_{2,1}), \ldots, \mathbf{y}(\theta_{2,P-1}, \phi_{2,P-1}), \ldots, \mathbf{y}(\theta_{J,P-1}, \phi_{J,P-1})].\end{aligned} \tag{3.74}$$

An application of a matched filter beamforming towards the direct path of the first sound source for instance in the time domain estimates s_1 with Eq. (3.76) as follows:

$$\begin{aligned}\hat{s}_{1,t} = s_{1,t} + \frac{\mathbf{y}(\theta_1, \phi_{j1})^\top}{||\mathbf{y}(\theta_1, \phi_1)||^2}\Big(&[\mathbf{y}(\theta_{1,1}, \phi_{1,1}), \ldots, \mathbf{y}(\theta_{1,P-1}, \phi_{1,P-1}), \mathbf{y}(\theta_2, \phi_2),\\ &\mathbf{y}(\theta_{2,1}, \phi_{2,1}), \ldots, \mathbf{y}(\theta_{2,P-1}, \phi_{2,P-1}), \ldots, \mathbf{y}(\theta_{J,P-1}, \phi_{J,P-1})].\\ &\left[s_{1,1}, \ldots, s_{1,P-1}, s_2, s_{2,1}, \ldots, s_{2,P-1}, \ldots, s_{J,P-1}\right]_t^\top\Big)\end{aligned} \tag{3.75}$$

Note that the multiplication of $\frac{\mathbf{y}(\theta_1,\phi_1)^\top}{||\mathbf{y}(\theta_j,\phi_j)||^2}$ by any other spherical harmonic vector of a different direction than (θ_1, ϕ_1) is inferior to 1.

The application of beamforming is possible in the time frequency domain as well. In the case of matched filter beamformer, we can apply it as follows:

$$\hat{s}_{j,f,n} = \frac{\mathbf{y}(\theta_j, \phi_j)^\top}{||\mathbf{y}(\theta_j, \phi_j)||^2}\mathbf{z}_{f,n}. \tag{3.76}$$

Unlike in the microphone domain, the beamformer does not depend on the frequency. Despite $\mathbf{Y}_f$ being frequency-dependent in Eq. (3.20). We can apply the beamformer and **estimate the sound source at the position of the micro-**

phone array. Indeed the only thing that makes $\mathbf{Y}_f$ frequency-dependent is the delays of propagating the sounds from their starting position to the recording position. Therefore, estimating a sound source with beamforming in the ambisonic domain consider already the delay of arrival and the attenuation of the signal while traveling from the starting position to the ambisonic microphone array.

- Second, we consider **the pseudo-inverse beamformer** [100, 101].[5] It consists in multiplying the ambisonic signals with the pseudo-inverse of the matrix containing the spherical harmonic vectors of the sound objects DoAs, it is given by:

$$\hat{\mathbf{s}}_t = \mathbf{Y}^{\dagger}\mathbf{z}_t, \tag{3.77}$$

where $\hat{\mathbf{s}}_t \in \mathbb{R}^{J\times 1}$, and the matrix $\mathbf{Y} \in \mathbb{R}^{M\times J}$ contains the spherical harmonic vectors corresponding to the DoA of the identified sound objects. Note that this beamformer is a particular case of the Linearly Constrained Minimum-Variance (LCMV) beamformer (explained later in Section 3.4.4.2) [27]. While extracting a sound object with this approach, the beam avoids interfering DoA by setting the beam pattern gain to 0dB. Note that the interference are avoided when the number of sources J to extract is lower than the number of channels $J \leqslant M$. If ever the $J \geqslant M$ the number of equations (being number of channels) will not be sufficient to find the unknowns (being the number of sound sources we want to estimate). With fourth-order ambisonics we can have 25 channels, which can be considered enough channels for a realistic amount of sound sources. Although this approach avoids identified interference sources, it introduces secondary beam patterns with gain that can be way larger than 0dB, which means, while it avoids the identified interfering sound sources it amplify none-identified ones coming from other direction such as echoes.

- Third, In order to avoid the problem of extremely amplified secondary beams, in the case where the number of sound objects J is lower than the number of channels, we propose to take advantage of the rest of channels to set a constraint on the secondary beam patterns using **regularized pseudo-inverse**.[6] This was one of the treated aspect in our first article [49]. We propose to regularize the expression in Eq. (3.77), inspired from Tikhonov regularization [43]. This is done by defining a number J' of directions of interest in the plane-wave basis in which sources are expected (or known) to be located. For each direction of interest we

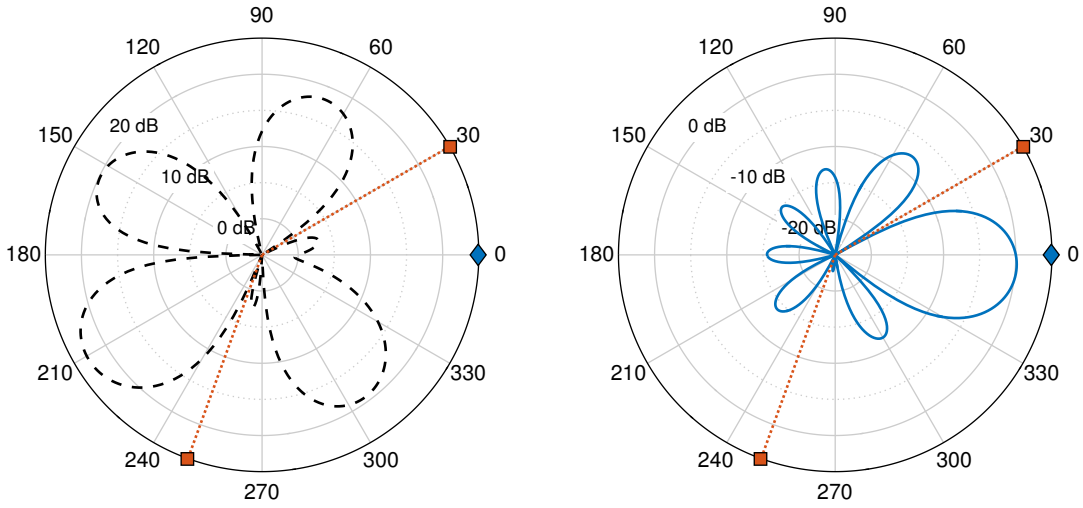
[5] This beamformer will be referred to as PIV in this manuscript.
[6] This beamformer will be referred to as regularized PIV.

estimate the corresponding source signal $\hat{s}_{j,t}$, with the following formula:

$$\hat{s}_{j,t} = \frac{\mathbf{y}^\top(\theta_j, \phi_j)(\mathbf{Y}_{samp}\mathbf{\Omega}\mathbf{Y}_{samp}^\top)^{-1}}{\mathbf{y}^\top(\theta_j, \phi_j)(\mathbf{Y}_{samp}\mathbf{\Omega}\mathbf{Y}_{samp}^\top)^{-1}\mathbf{y}(\theta_j, \phi_j)}\mathbf{z}_t \tag{3.78}$$

where $\mathbf{Y}_{samp} \in \mathbb{R}^{M\times Q}$ is the matrix of the spherical harmonic coefficients for a basis of $Q \geqslant M$ plane-wave directions regularly distributed around the sphere including the direction of interest (θ_j, ϕ_j), $\mathbf{y} \in \mathbb{R}^{M\times 1}$, $\mathbf{z} \in \mathbb{R}^{M\times 1}$,and $\mathbf{\Omega} \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix of weights assigned to the plane-wave directions:

$$\begin{aligned} \mathbf{\Omega} &= \text{diag}(\mathbf{w}) \\ \mathbf{w} &= [w_1, w_2, \ldots, w_Q] \end{aligned} \tag{3.79}$$



(a) Plain pseudo-inverse.

(b) Regularized pseudo-inverse.

Fig. 3.8 Beam pattern resulting from the application of Eq. (3.77) or Eq. (3.78), and Eq. (3.80) for the direction (0,0) in the case of a 2D; 4^{th} order ambisonic sound field. Note that the scale is not the same. The main lobe for the sound of interest have a gain of $0dB$ in both figures.

One can think of different schemes for choosing the weights w_q. Since our point is to separate sound sources of interest from each other, we propose the following choice of weights:

$$w_q = \begin{cases} 1 & \text{if } (\theta_q, \phi_q) \text{ is a direction of interest,} \\ \frac{1}{Q-J} & \text{otherwise.} \end{cases} \tag{3.80}$$

This choice of weights results in beam patterns that separate the sources of interest from each other while maintaining "reasonable" side lobes. To illustrate our point, let us consider a 2D (for a visual purpose);[7] 4^{th} order ambisonic sound field (number of channels $M = 2L + 1 = 9$), that contains 3 direction of interest $\theta = [0°, 30°, 250°]$. We plotted in Fig. 3.8 the beam pattern corresponding to the extraction of the source signal at 0°. With a plain pseudo inverse that directions of interference are avoided. However, the side lobes are huge in other directions with a gain that can exceed 20dB. With the proposed regularization, we managed to avoid the directions of interference, as well as side lobes that do not exceed −5dB. For the rest of this manuscript this beamformer will be referred to as regularized PIV.

3.4.4.2 Approaches based on exploiting the mixture content

Unlike in Section 3.4.4.1, the methods presented in this section exploit in real-time the content of the ambisonic signals. The gains of the beam patterns are set automatically according to the power present in each direction. This information is contained in the covariance matrix of the ambisonic signals. We can estimate the covariance matrix for a given frame corresponding to the sample $t \in [(k-1)T, \ldots, kT-1]$ under the hypothesis that the ambisonic signals have a zero mean, as follows:

$$\mathbf{C}_k = \frac{1}{T} \sum_{(k-1)T}^{kT-1} \mathbf{z}_t \mathbf{z}_t^\top, \tag{3.81}$$

where the matrix $\mathbf{C}_k \in \mathbb{R}^{M \times M}$. In practice in order to avoid abrupt changes between two consecutive frames, it is recommended to apply a temporal smoothing by taking into consideration the previous frame up to a certain factor $0 \leqslant \alpha < 1$. This can be applied as follows:

$$\mathbf{C}_k = (1 - \alpha)\mathbf{C}_{k-1} + \alpha\mathbf{C}, \tag{3.82}$$

with $\mathbf{C}$ being the current covariance matrix.

- First, we consider the **Minimum-Variance Distortionless Response (MVDR)**[8] beamformer [15, 118, 101]. With this approach the beam pattern set a gain of 0dB toward the direction of interest while minimizing the total energy. The

[7] By 2D ambisonic sound field, we consider the representation format in the plane XoY, for that the elevation is considered to be equal to zero and the elevation channels are discarded since they are equal to zero. Therefore, the number of channels is M=2L+1.

[8] This beamformer will be referred to as MVDR in this manuscript.

extraction of the j^{th} source is given by:

$$\hat{s}_{j,t} = \frac{\mathbf{y}(\theta_j, \phi_j)^\top . \mathbf{C}_k^{-1}}{\mathbf{y}(\theta_j, \phi_j)^\top . \mathbf{C}_k^{-1} . \mathbf{y}(\theta_j, \phi_j)} \mathbf{z}_t \tag{3.83}$$

with $\hat{s}_{j,t} \in \mathbb{R}$ is the estimate signal magnitude at the time t. In the case of extracting J sound sources, the matrix of extraction is given by:

$$\begin{aligned} \mathbf{D}_{MVDR} &= [\mathbf{d}_1^{MVDR}, \mathbf{d}_2^{MVDR}, \dots, \mathbf{d}_J^{MVDR}]^\top \\ \mathbf{d}_j^{MVDR} &= \frac{\mathbf{y}(\theta_j, \phi_j)^\top . \mathbf{C}_k^{-1}}{\mathbf{y}(\theta_j, \phi_j)^\top . \mathbf{C}_k^{-1} . \mathbf{y}(\theta_j, \phi_j)}, \end{aligned} \tag{3.84}$$

where $\mathbf{D}_{MVDR} \in \mathbb{R}^{J \times M}$, and $\mathbf{d}_j^{MVDR} \in \mathbb{R}^{1 \times M}$.

With this approach the extraction lobes have a unitary gain (distortionless response), while having a minimal total energy (minimum variance). Although this approach takes into consideration the contents of the ambisonic signals and tries to minimize the energy of the side lobes, it does not set a strict constraint on the interfering directions.

- Second, we consider the **Linearly-Constrained Minimum Variance (LCMV)** [118, 101].[9] With this approach we can solve the problem of the MVDR beamformer. It works similarly as the MVDR while adding a strict constraint on the sources of interest, which is done by solving the following problem:

$$\mathbf{d}_j^{LCMV} = \{\arg\min_{\mathbf{w}} (\mathbf{w}^\top \mathbf{C}_k \mathbf{w}) | \mathbf{Y}^\top \mathbf{w} = \mathbf{u}_j\}, \tag{3.85}$$

where $\mathbf{u}_j$ is a discrete Dirac impulse given by:

$$\begin{aligned} \mathbf{u}_j &= [u_{1,j}, u_{2,j}, ..., u_{J,j}]^\top, \\ u_{m,j} &= \begin{cases} 0, & m \neq j \\ 1, & m = j \end{cases}. \end{aligned} \tag{3.86}$$

In other words, the extracted signal of the j^{th} source is given by the beam that insures the lowest total energy for the side lobes, while making sure that the gain is equal to 1, and 0, for the j^{th} sound source DoA and interference sound sources, respectively. The extraction matrix is given by:

$$\mathbf{D}_{LMCV} = [\mathbf{Y}^\top \mathbf{C}_k^{-1} \mathbf{Y}]^{-1} \mathbf{Y}^\top . \mathbf{C}_k^{-1}. \tag{3.87}$$

[9] This beamformer will be referred to as LCMV in this manuscript.

with $\mathbf{D}_{LMCV} \in \mathbb{R}^{J,M}$.

3.4.4.3 Mixed plane wave decomposition

We implemented a beamformer in which we mixed both strategies as in Section 3.4.4.1 and Section 3.4.4.2. This approach is going to be referred to as "mixed beamformer". The extraction matrix for this approach is given by:

$$\begin{aligned} \mathbf{D}_{Mixed} &= [\mathbf{d}_1^{Mixed}, \mathbf{d}_2^{Mixed}, \dots, \mathbf{d}_J^{Mixed}]^\top \\ \mathbf{d}_j^{Mixed} &= \frac{\hat{\mathbf{C}}_k^{-1}\mathbf{y}(\theta_j, \phi_j)}{\mathbf{y}(\theta_j, \phi_j)^\top \hat{\mathbf{C}}_k^{-1}\mathbf{y}(\theta_j, \phi_j)}. \end{aligned} \tag{3.88}$$

where $\mathbf{d}_j^{Mixed} \in \mathbb{R}^M$ and $\mathbf{D}_{Mixed} \in \mathbb{R}^{J \times M}$. The covariance matrix $\hat{\mathbf{C}}_k$ is a mix between the current time frame covariance matrix $\mathbf{C}_k$ and the one corresponding to a regularized pseudo-inverse PIV approach. It is given by:

$$\hat{\mathbf{C}}_k = \frac{\mathbf{C}_k}{\text{tr}(\mathbf{C}_k)} + \frac{\mathbf{Y}\mathbf{Y}^\top + J'\mathbf{I}}{\text{tr}(\mathbf{Y}\mathbf{Y}^\top + J'\mathbf{I})}, \tag{3.89}$$

where tr() denotes the trace operator, the matrix $\mathbf{I} \in \mathbb{R}^{M \times M}$ is the identity matrix, and the matrix $\mathbf{Y} \in \mathbb{R}^{M \times J}$ contains the spherical harmonic vectors of the sound source of interest.

3.5 Conclusion

In this chapter, we first surveyed some sound source localization approaches. We considered some known methods in the microphone domain, and we proposed their adaptation in the ambisonic domain. We discussed other known approaches that were intended for ambisonics. We can say that we have several methods when it comes to sound source localization to chose from if we ever want to know the sound sources direction of arrivals. We just need to compare them to each other with some numerical experiments. This is going to be presented in the next Chapter 4. Moreover, note that the motivation for this work is the production of immersive audio-visual experiences. In this context, several cameras are likely to be employed to provide different viewpoints on the scene. The visual information recorded by these cameras could thus be used to track the number of sound sources, such as actors, over time. Therefore, in the following, we assume that we have enough information to use the algorithms provided in Chapter 3 Section 3.3.

Secondly, we surveyed some sound source separation approaches. In this case, we have three main methods: binary masking and multichannel Wiener filter based on the local Gaussian model, and plane wave decomposition. In the case of plane wave decomposition we proposed two beamforming approaches that we can use on ambisonic mixtures. The first one referred to as regularized PIV beamformer, which was inspired from the Tikhonov regularization. This allows us to take advantage of the number of channels in ambisonic mixtures. The second one referred to as mixed beamformer, in which we mixed a regularized PIV while exploiting the mixture content such as with MVDR. . These approaches are going to be assessed through some numerical experiments in Chapter 4. The second approach seems to be very interesting because it is known that a Wiener filter is a smooth filter that introduces minimum artifacts if it is well estimated. With this approach, the estimation of the Wiener filter parameters is based on a strong hypothesis on the contribution of the sound sources on each microphone. We need to check the validity of the local Gaussian hypothesis in the ambisonic domain, and find if there is a way to apply it on ambisonic mixtures. This is going to be the main subject of Chapter 5. Note that the surveyed approaches are quite old and classic. And the current trend is deep learning-based sound source separation and localization approaches.

Chapter 4

Pre-validation of the global approach

In this chapter, we explain in the first section, the approach that we adopted to respond to the navigation problem. In this first section, we present two different variations of our global strategy.

The need for the estimation of the sound sources DoA will be expressed in the first section. As a consequence, we evaluate in the second section the surveyed methods in Chapter 3 Section 3.3. We evaluate for the first time approaches that have never been used in the ambisonic domain, and compare them to used ones.

In the third section, we briefly express the need to investigate more when it comes to sound source separation. In the fourth section, we investigate and validate our navigation strategy using an objective metric. Finally in the last section we conclude this chapter and prepare the reader for the next part.

4.1 Global approach

We use ambisonics for its flexibility of recording sound fields and the ability to recover sound objects as in the object-oriented audio. By that, we mean recovering from ambisonic mixtures, the sound source signals, and their positions. This operation will be referred to as decomposing ambisonic sound fields.

Note that this decomposition will not provide an accurate presentation of the sound field with object-oriented audio. Indeed, in my P.hD. work, we are not trying to recover the room impulse responses corresponding to each sound objects, which can be on its own another subject. We will recover the sound sources signals and collect enough information about their positions (their DoA, and distance from the microphone array). With these information, we propose to process each sound object according to the user's

movement from a point of space to an other one, and add to each other to simulate a mixture in the user's current position.

In Fig. 4.1, we illustrate our main idea in a diagram. First, the ambisonic sound scene is captured using a spherical microphone array. The microphone signals are encoded to produce the ambisonic signals. These signals are processed to estimate the signal of each sound source as well as their DoA. Given these DoA, their distances from the main antenna and the information about the user's movement from it's original position to a new one, the current DoA are deduced.

Taking into consideration these new DoA, the previous ones (corresponding to the user's starting position), the signal of the sound sources, and an estimation of the sound sources distance from the microphone array, we can estimate the ambisonic signals corresponding to the starting position and current position of each source signal. With these signals, we can estimate the ambisonic signals corresponding to the current and starting listening position. We can consider stopping the processing and considering the ambisonic mixture corresponding to the movement from the starting position to the current one as the estimated ambisonic signals from the extracted sound sources. However, in reality there are some other sources that we are not able to locate and extract because they create a diffuse field.[1] In order to overcome this problem we propose to compute a residual ambisonic sound field and add the ambisonic signals that were computed from adapting the extracted sound source to the user's final position. In the case the user turns his or her head, a rotation is applied to the entire ambisonic mixture. The only way to make the main idea clear is through an example with some equations.

Considering a sound field that contains j sound sources. This sound field was captured with a spherical microphone array in point A. Each sound source is located from the direction $(\theta_{j_A}, \phi_{j_A})$ and at a distance r_{j_A} from the recording position. These sound sources are emitting a signal that is perceived at a time t and point A as $s_{j_A,t}$. We want to move the listening point from point A to point B. Note that the sound sources are at a distance r_{j_B}, and at a direction $(\theta_{j_B}, \phi_{j_B})$ from the point B. Using the scheme proposed in Fig.4.1, the process block will produce the ambisonic signals of each specific sound source corresponding to point A and to the current point B, which are given by:

$$\mathbf{z}_{j_A,t} = s_{j_A,t}\mathbf{y}(\theta_{j_A}, \phi_{j_A}) \tag{4.1}$$

$$\mathbf{z}_{j_B,t} = \hat{s}_{j_B,t}\mathbf{y}(\theta_{j_B}, \phi_{j_B}), \tag{4.2}$$

[1] A diffuse sound source correspond to the same sound source signal coming from different directions with very small time differences.

where $\mathbf{y}(\theta, \phi)$ are the spherical harmonics corresponding to the direction (θ, ϕ) and $\hat{s}_{j_B,t}$ are an estimation of the sound source signals at time t and position B. Note that a time signal is delayed and attenuated while traveling from point A to point B in the air. The estimation of these signals is given by:

$$\hat{s}_{j_B,f} = \frac{r_{j_A}}{r_{j_B}} s_{j_A,f} e^{\frac{i2\pi f(r_{j_B} - r_{j_A})}{c}}. \tag{4.3}$$

First, let us assume for simplicity that all the sounds DoA were successfully estimated, and their signals were successfully separated after the decomposition of the ambisonic mixture. In this case, the ambisonic signals corresponding to point B are given simply by adding the ambisonic signals of each sound source at point B, which is provided by:

$$\mathbf{z}_{B,t} = \sum_j \mathbf{z}_{j_B,t} \tag{4.4}$$

However, it is impossible to recover each sound because, in most sound field recordings, we usually have direct path sound sources and their reflections. The direct path sound sources can be identified, and their signals can be recovered (more details about how we chose to handle echoes are given in Section 4.1.2) . We can try to identify some echoes, but it is impossible to identify each one of them. We usually assume that the late reverberation is diffuse. Under this assumption, we propose translating the ambisonic sound field by keeping the diffuse part the same. After identifying the sounds that must be adapted, the translation block in Fig.4.1 is given by:

$$\mathbf{z}_{B,t} = \mathbf{z}_{A,t} - \sum_j \mathbf{z}_{j_A,t} + \sum_j \mathbf{z}_{j_B,t}, \tag{4.5}$$

where $\mathbf{z}_{A,t} - \sum_j \mathbf{z}_{j_A,t}$ is the residual sound field.

The head tracking block is a head tracker that is usually assembled with the user's headphones. It provides the user's head direction, which allows us to adapt to the ambisonic sound field by simple matrix multiplication. For more information about rotation matrices, please refer to Chapter 2 Section 2.4.1.

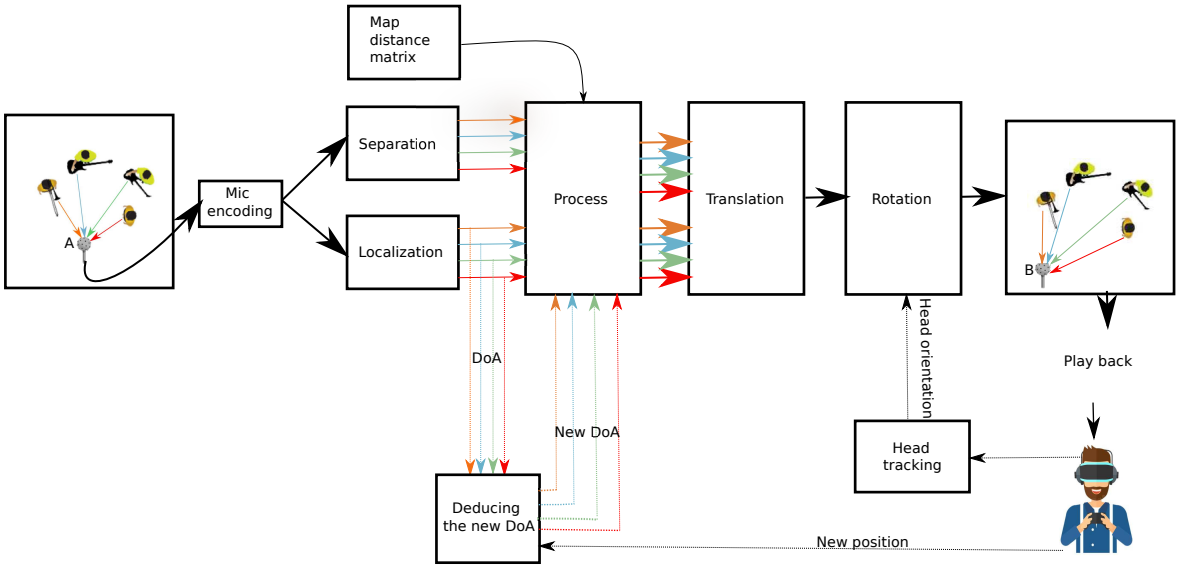


Fig. 4.1 Our main idea for navigation in ambisonic sound scenes.

As part of the scheme proposed in Fig. 4.1, we suggest two different approaches to handle the sound source separation:

- A first variant based on simple beamforming for the decomposition of the ambisonic sound field into plane waves.
- A second variant based on multichannel sound source separation followed by beamforming for the decomposition of the ambisonic sound field into plane waves.

Note that we assume that we can easily have information about the distance of each sound source as well as any boundary or object in the sound field from the recording position. Indeed, given the direction in which we are interested, we can use a time-of-flight type camera that can automatically give us the distance from each object. This also can help to give us the sound sources DoA, which is one of the reasons we didn't do further research on the sound source localization problem.

4.1.1 Navigation based on a simple plane wave decomposition

Our first approach is showcased in Fig. 4.2. Similar to the explanation given for our main idea, our first approach contains two steps. In the first step, the ambisonic sound field is decomposed into plane waves. To each sound source, we associate information about locations such as angle and distance. In the second step, the sound scene is reconstructed by panning the extracted sound sources with taking into consideration the user's current position.

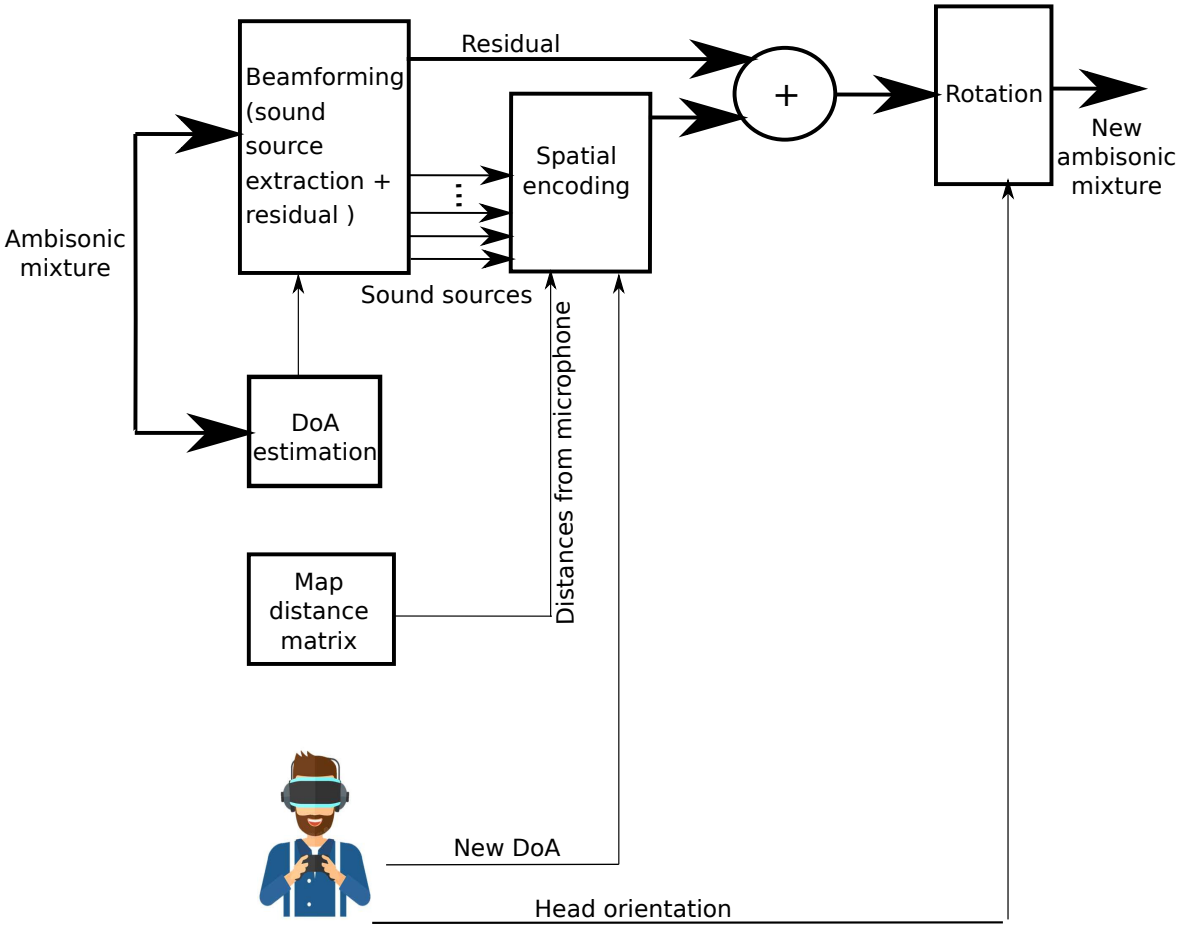


Fig. 4.2 Navigation based on a plane wave decomposition of the ambisonic sound fields using full band beamformers.

The separation, in this case, is called spatial, because it is only based on the position of the desired objects to extract. In this step, we can define two types of extractions. The first type is the extraction of the direct path source signals, which is done by applying a beamforming that is guided by the sound sources DoA. The beam shape can be guided by dynamically taking into consideration the sound field (different types of beamforming were discussed in the last chapter and they are going to be assessed through some numerical experiment in this chapter).

The DoAs can be known or estimated using a sound source localization algorithm (a survey on different types of sound source localization for ambisonics was studied and discussed in Chapter 3 Section 3.3). Extracting and adapting the direct path sound sources is sometimes not sufficient. As explained in Section 4.1, the sound field contains secondary components, which correspond mainly to the reflections of the sound sources on the boundaries such as walls, ceiling, and floor. Other types of components can be present such as diffuse noise.

These secondary components can be left intact if considered diffuse. However, some times, some reflections may affect navigation. In reality, the reverberation consists of primary echoes and diffuse reverberation. The first echoes are far from being diffuse. We propose to extract the first echoes by a directional sampling of the sound field. In other words, the sound field is decomposed into a grid, and the directions that contain more power are extracted from the sound field. To this second type of extraction, we propose to apply a similar process as a direct path sound signal by panning them in regards to the user's new position. A similar approach was considered in [2]. Indeed, the authors propose to extract dominant sources using a matching pursuit algorithm, followed by labeling the extracted components. Labeling the components helps to identify the direct path components and their primary reflections. This is done by computing the correlation of each primary source with the rest of the components. A component is labeled to be a reflection to a given primary source if they are the more correlated.

Our approach and the one presented in [2] depend hugely on the order of the decomposed ambisonic sound scene. The decomposition can be very accurate for large orders. Note that the larger the number of channels, the more accurate the estimation of the sound source signals. With a commercially available spherical microphone array, we can have at best fourth-order ambisonic sound scenes (25 channels). Compared to a typical microphone recording, 25 channels can be considered a large number of equations for a plane wave decomposition. However, complex sound scenes (a longer reverberation time, large amount of sources, and close sound sources) make it very hard to have an accurate spatial separation, because of the correctness of the assumptions for longer reverberation time, for instance the narrow band approximation. For the proposed approach, we studied different strategies for the plane wave decomposition, and we recommend to use a specific type of beamforming to take advantage of the entire 25 channels. The used algorithm to handle reflection is given in Alg. 5. It will be described with more details later in Section 4.1.2.

We can summarize our approach of navigation in a transformation matrix $\mathbf{T}$ as in Chapter 2 Section 2.4, if ever the phase shifting of the signals in Eq. (4.3) is discarded. This choice can be considered reasonable in indoor environments with small dimensions, such as a conference room. We can assume that the phase-shifting accumulation to the farthest point possible can still not be noticeable by human ears. This choice is no longer available in large environments such as gymnasium, for instance. In this case, the phase-shifting should be incorporated, and the application of the matrix T should be applied in the frequency domain.

For simplicity we consider the indoor environments with small dimensions. Thus we can summarize our approach of navigation in a translation matrix $\mathbf{T}$ that can be applied as a transformation in the time domain. This matrix will be given in Eq. (4.14).

This hypothesis is sufficient even for high frequencies. Indeed, it is very difficult for a human being to notice the delay when the translation is small.

First, the sound sources are identified, and their signals are extracted using a particular type of beamforming, which is done by projecting the ambisonic signals on a matrix $\mathbf{D} \in \mathbb{R}^{J' \times M}$, which is given by:

$$\hat{\mathbf{s}}_{A,t} = \mathbf{D}\mathbf{z}_{A,t}, \tag{4.6}$$

where $\hat{\mathbf{s}}_{A,t} \in \mathbb{R}^{J'}$, J' can be equal to the number of sources J or larger,[2] $M = (L+1)^2$ being the number of channels. This matrix depends on the spherical harmonic vectors corresponding to the direction of the sound signals that we would like to extract (more details are given later in this section depending on the used plane wave decomposition strategy and method).

The residual sound field $\mathbf{z}_{res,t}$ is computed as follows:

$$\mathbf{z}_{res,t} = \mathbf{z}_{A,t} - \mathbf{C}_A \hat{\mathbf{s}}_{A,t}, \tag{4.7}$$

where the matrix $\mathbf{C}_A \in \mathbb{R}^{M \times J'}$ contains the spherical harmonic vectors of the sources DoA to the point A.

Since we discard[3] the phase-shifting while moving from point A to point B, we can write the ambisonic signals corresponding to the contribution of the extracted sound sources $\mathbf{z}_{B_{obj},t}$[4] in the time domain as follows:

$$\mathbf{z}_{B_{obj},t} = \mathbf{C}_B \mathbf{G} \hat{\mathbf{s}}_{A,t} \tag{4.8}$$

where the matrix $\mathbf{C}_B \in \mathbb{R}^{M \times J'}$ contains the spherical harmonic vectors of the sources DoA to point B. The matrix $\mathbf{G} \in \mathbb{R}^{J' \times J'}$ is diagonal, and it contains the gains corresponding to the amplitude changes from Eq. (4.3). Its coefficients are given by:

$$g_j = \frac{r_{j_A}}{r_{j_B}}. \tag{4.9}$$

Under the hypothesis that the residual sound field expressed in Eq. (4.7) is diffuse,

[2] by $J' > J$ we mean that we have J sources and we identified $J' - J$ echoes that are worthy of being extracted and panned in function of the current position.

[3] Under the hypothesis that the navgation is done in indoor environment with small dimension, the phase shifting can be unnoticeable by humans.

[4] Here the index B refers to the position B, and the index "obj" indicate that the ambisonic signals are computed from the extracted sound objects.

we can estimate the ambisonic signals in point B as follows:

$$\begin{aligned}\mathbf{z}_{B,t} &= \mathbf{z}_{B_{obj},t} + \mathbf{z}_{res,t} \\ &= \mathbf{C}_B\mathbf{G}\hat{\mathbf{s}}_{A,t} + \mathbf{z}_{A,t} - \mathbf{C}_A\hat{\mathbf{s}}_{A,t} \\ &= \mathbf{C}_B\mathbf{G}\mathbf{D}\mathbf{z}_{A,t} + \mathbf{z}_{A,t} - \mathbf{C}_A\mathbf{D}\mathbf{z}_{A,t} \\ &= (\mathbf{C}_B\mathbf{G}\mathbf{D} + \mathbf{I} - \mathbf{C}_A\mathbf{D})\mathbf{z}_{A,t} \\ &= \mathbf{T}\mathbf{z}_{A,t},\end{aligned} \tag{4.10}$$

where the matrix $\mathbf{T} \in \mathbb{R}^{M\times M}$ is the translation matrix from point A to point B, and it is given by:

$$\mathbf{T} = \mathbf{C}_B\mathbf{G}\mathbf{D} + \mathbf{I} - \mathbf{C}_A\mathbf{D}. \tag{4.11}$$

In the case the user turns his/her head we can apply the rotation matrix $\mathbf{R} \in \mathbb{R}^{M,M}$ deduced from a head tracking device, which is applied as follows:

$$\mathbf{z}_{B,t} = \mathbf{R}\mathbf{T}\mathbf{z}_{A,t}. \tag{4.12}$$

In the case, the phase-shifting must be incorporated, the matrix $\mathbf{G}$ is frequency-dependent, and its coefficients are given by:

$$g_{j,f} = \frac{r_{j_A}}{r_{j_B}} e^{\frac{i2\pi f(r_{j_B} - r_{j_A})}{c}}. \tag{4.13}$$

This makes the matrix $\mathbf{T}$ frequency dependent as well. Its expression is given by:

$$\mathbf{T}_f = \mathbf{C}_B\mathbf{G}_f\mathbf{D} + \mathbf{I} - \mathbf{C}_A\mathbf{D}. \tag{4.14}$$

The application of the matrix should be in the frequency domain, which is given by:

$$\mathbf{z}_{B,f} = \mathbf{T}_f\mathbf{z}_{A,f}. \tag{4.15}$$

4.1.2 Navigation based on a multichannel sound source separation

The decomposition in the first approach of our strategy is based on plane wave decomposition in order to decompose the sound field (Section. 4.1.1). This decomposition depend only on spatial cues, which makes it depends on the number of channels the sound field comes with. When the sound scene is complex, the approach is very limited by the decomposition. As you can imagine if the sound sources are close to each other, the performance of the plane wave decomposition will present a lot of interferences. To

this aim, instead of decomposing the ambisonic sound field directly into plane waves, we propose to apply first a multichannel sound source separation to the ambisonic sound field.

Specifically, we would like to search for the contribution of each source in each ambisonic channel. This allows us to decompose an ambisonic sound field that contains J sound sources into J ambisonic sound fields. Each resulting ambisonic sound field is the contribution of its dedicated sound source in each channel. In other words, the j^{th} ambisonic sound field is the ambisonic mixture of the j^{th} sound source direct path and its reflections from the boundaries. This sound source separation can be seen as if the recording was done for each sound source separately (Fig. 4.3). This allows us to have control over each sound source separately.

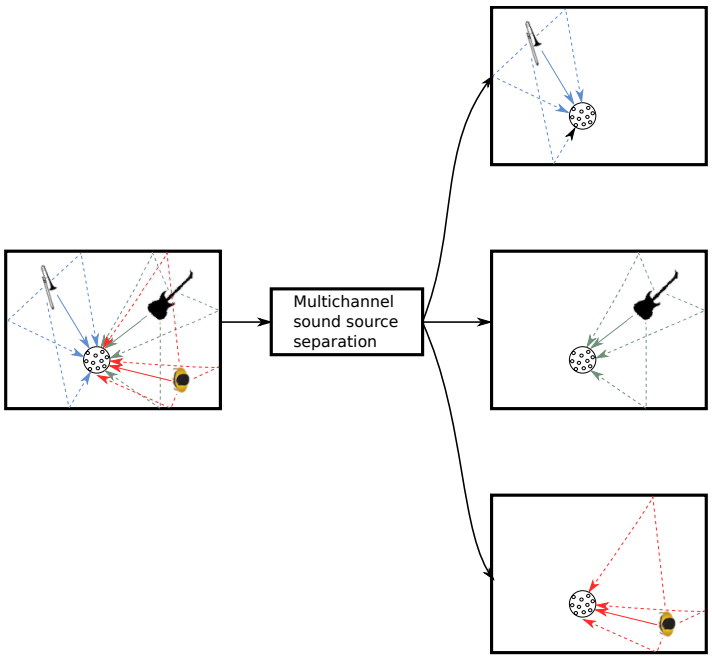


Fig. 4.3 Example of a multichannel sound source separation. The number of sound sources may vary.

The multi-channel source separation problem can be solved by looking for the filter that minimizes the expected squared error for every sound source j and every time frequency bin (f, n):

$$\begin{aligned} &\forall j \in [1, J], f \in [1, F] \text{ and } n \in [1, N], \\ &\mathbf{W}_{j,f,n} = \underset{\mathbf{W}}{\operatorname{argmin}}\, \mathbb{E}\left[\|\mathbf{b}_{j,f,n} - \mathbf{W}\,\mathbf{z}_{f,n}\|_2^2\right] . \end{aligned} \tag{4.16}$$

As discussed earlier in the state of the art (Chapter 3 Section 3.4.3) [125], this problem

is solved by the multichannel Wiener filter which is given in the ambisonic domain by:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{(\mathbf{b}_{j,f,n}, \mathbf{z}_{f,n})} \, \mathbf{\Sigma}^{-1}_{(\mathbf{z}_{f,n}, \mathbf{z}_{f,n})} \, , \tag{4.17}$$

where the matrices $\mathbf{\Sigma}_{(\mathbf{z}_{f,n}, \mathbf{z}_{f,n})}$ and $\mathbf{\Sigma}_{(\mathbf{b}_{j,f,n}, \mathbf{z}_{f,n})}$, represent the covariance of the ambisonic mixture $\mathbf{z}_{f,n}$ and the cross-correlation between the vectors $\mathbf{b}_{j,f,n}$ and $\mathbf{z}_{f,n}$, respectively. A proposition to reduce the problem and a couple of approaches are discussed in Chapter 5, Chapter 6, and Chapter 7 of this manuscript.

Assuming we can estimate successfully the contribution of each sound source in each ambisonic channel, to navigate, we suggest a second decomposition step in which a plane wave decomposition is applied. In other words, we propose to apply a multichannel sound source separation to retrieve the contribution of each source in each channel (Fig. 4.3) followed by the navigation approach proposed in Section 4.1.1 to each separated ambisonic contribution (Fig. 4.4). The second decomposition aims to recover the direct path sound source and the primary echoes.

On the one hand, with the 1st approach (navigation based on a simple plane wave decomposition), we proposed to use several variants of plane wave decomposition (basic projection, PIV, regularized PIV, Mixed, MVDR, LCMV), some of them take into account interfering sound sources. On the other hand, in the 2nd approach, in the second step of the decomposition, we already eliminate interference sources (if the multichannel sound source separation works perfectly). We can, however, use these methods (plane wave decomposition such as PIV, regularized PIV, Mixed, etc.), and consider early echoes as interference sources.

The DoA of the echoes can be estimated using the same approach as the one proposed in [2], which is based on a matching pursuit algorithm. First the sphere of the same center as the recording position is sampled densely and regularly. Second, in an iterative algorithm, the ambisonic signals are projected into the spherical harmonic vector corresponding to the direction of each sample. Third, the direction that contains the most power is selected and stored. Fourth, the residual is computed by subtracting from the ambisonic signals the estimated ones of the sound signals with the most power. Fifth, the second step is repeated while considering the ambisonic mixture as the residual computed in the fourth step until a chosen threshold. The threshold can be up to a certain number of iterations or an amount of power in the residual sound field. The described algorithm is given in Alg. 5. Knowing from which direction each sound has been extracted, and having at our disposal the map distance matrix that was acquired from a time-of-flight type camera, we can change the point of view in terms of the extracted sounds using Eq. (4.3). Instead of using this algorithm directly to the mixture, unlike in [2], we propose to use it on each ambisonic signals provided from the multichannel sound source separation. We suggest to fix the number of iterations

to 4. The 1^{st} given direction would be the direct path, the 2^{nd}, 3^{th}, and 4^{th} directions would be considered as the DoA of the 1st, 2nd, and 3rd echoes.

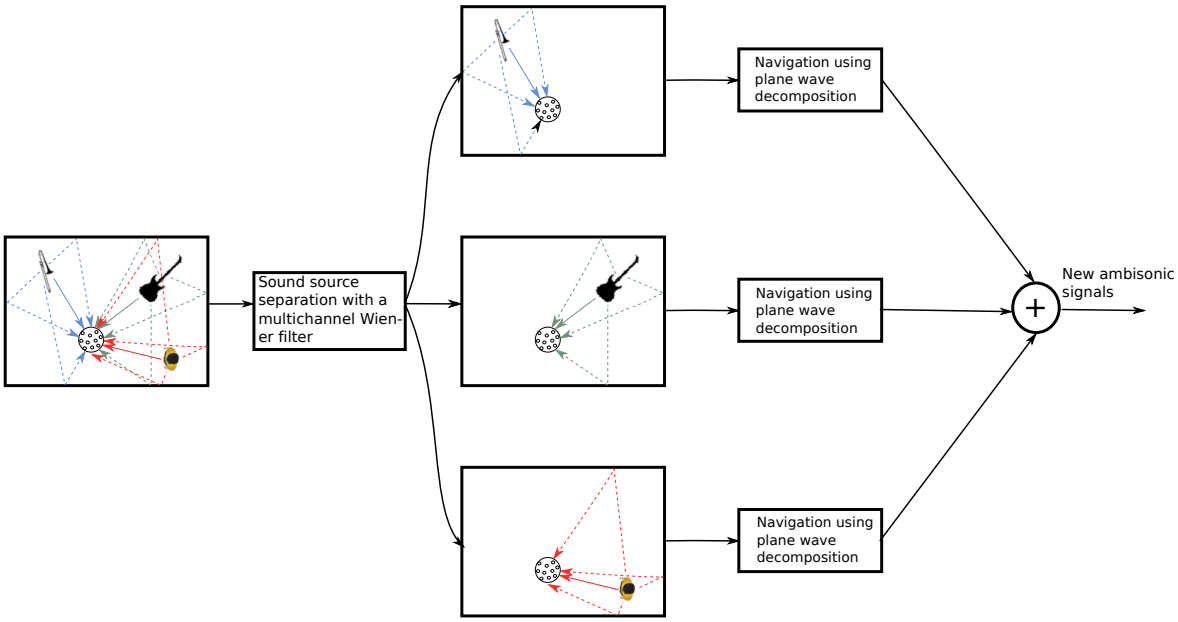


Fig. 4.4 Navigation with our second approach. Note that the number of sound objects could be more or less than 3. The number of sound sources may vary.

In the following we will experiment with the surveyed sound source localization and separation approaches.

4.2 Validation of the localization bricks

Our navigation strategy is heavily based on locating the sound sources in the sound field. Therefore, in this section, we check the performance of the sound source localization approaches that have been surveyed in Chapter 3 Section 3.3. First, we begin by presenting our simulation setup in which we discuss the simulation data and the used objective measure with which we evaluate the performances. Second, we discuss the parameters that we took into consideration for each approach and how the azimuth and the elevation were extracted. And finally, we present and discuss the results, and compare the order of magnitude with a new sophisticated baseline approach for ambisonics that is based on neural networks.

Algorithm 5 Directional decomposition

Input The mixture $\mathbf{z}_t$, and the DoA of the primary sound sources $\{\theta_j, \phi_j\}_{j \in J}$
Output The DoA of the echoes $DoA_{Echo} = \{\theta_q, \phi_q\}_{q \in Q}$

Set the maximum of iteration $MaxIter = Q$
Compute the spherical harmonics vector $\mathbf{y}_j$ corresponding to (θ_j, ϕ_j)
Sample the sphere into directions $\mathbf{Y} = [\mathbf{y}(\theta_i, \phi_i)]_{i \in I}$
Compute the direction dictionary spherical harmonic vectors $\{\mathbf{y}_i\}_{i \in I}$
$\mathbf{z}_{residual,t} = \mathbf{z}_t$
j = 0
q = 0
$DoA_{Echo} = \varnothing$
for $j \leqslant J$ **do**
    $\mathbf{z}_{residual,t} = \mathbf{z}_{residual,t} - \frac{\mathbf{y}(\theta_j,\phi_j)^\top}{||\mathbf{y}(\theta_j,\phi_j)||^2}(\mathbf{y}(\theta_j, \phi_j)^\top.\mathbf{z}_t)$
    $j = j + 1$
end for
for $q \leqslant MaxIter$ **do**
    $e = \arg\max_{i \in I} \left|\left| \frac{\mathbf{y}(\theta_i,\phi_i)^\top}{||\mathbf{y}(\theta_i,\phi_i)||^2} \mathbf{z}_{residual,t} \right|\right|^2$
    $s_{e,t} = \frac{\mathbf{y}(\theta_e,\phi_e)^\top}{||\mathbf{y}(\theta_e,\phi_e)||} \mathbf{z}_{residual,t}$
    $\mathbf{z}_{residual,t} = \mathbf{z}_{residual,t} - s_{e,t}\mathbf{y}(\theta_e, \phi_e)$
    $DoA_{Echo} = DoA_{Echo} \cup \theta_e, \phi_e$
    $q = q + 1$
end for

4.2.1 Simulation set-up

4.2.1.1 Dataset

We took into consideration 3×4 scenarios. Each one represents a particular configuration. Each configuration regroups a given amount of sound sources and a given reverberation time. The considered number of sound sources were either one, two, or three sound sources in the sound field. The considered reverberation times were $RT_{60}(s) = [0, 0.2, 0.4, 0.7]$. For each scenario, with the help of the adopted simulation framework MCRoomSim [130], we generated 20 Eigenmike RIR. The sound source positions, were randomly chosen while ensuring that at least 10° is between two close sources and at least 2.5 m apart from the recording device. The sound sources were not considered to be omnidirectional and each source direction direction was oriented toward the Eigenmike. All of this results in $3 \times 4 \times 20 = 240$ RIR.

For each RIR, 10 seconds of one, two, or three (depending on the RIR) speech sounds were randomly selected from the SiSEC campaign data set [122] and convolved with RIR. If the number of sound sources is larger than one, the resulted signals were added to each other to create an Eigenmike mixture. These mixtures were encoded to get ambisonic mixtures. In total, we had 240 fourth-order (M=25) mixtures. For more information about how we generate out simulation mixture please refer to Appendix B.

4.2.1.2 Evaluation measure

One of the ways to judge and compare the performances of sound source localization approaches is to compute the angle differences between the true DoAs and the estimated ones. Note that averaging the angle differences over all the examples is misleading because the estimation might be spot on for some cases and not great for others. There is a way to overcome this problem using the parameters defined in the MBSS locate toolbox.[5] The idea is to set an angle tolerance and report the accuracy and the error for the examples that achieved in having a lower angle difference than the angle tolerance. The accuracy will present the percentage of cases that have succeeded, and the error will be computed only for these examples.

However, there isn't a consensus on the angle tolerance in the literature community. In some articles, the angle tolerance is set only for the azimuth, other for the elevation, and some believe it is necessary to establish the angle tolerance on both the azimuth and the elevation. In our case, the third option is the one that we will adopt.

In order to compute the evaluation measures, we used a file called "$MBSS\_eval.m$" in the Multichannel BSS locate toolbox. With this file it is possible to compute the

[5] `http://bass-db.gforge.inria.fr/bss_locate/`

error and the accuracy of the localization for a chosen angular tolerance constraint. We based our evaluation on these measures.

4.2.1.3 Algorithm parameters

- DUET[6]: After constructing a matrix that represents the values of the histogram over the azimuth and the elevation, we took into consideration only the selected directions that were larger than a given threshold. Our threshold was set to be the mean over the maximums of occurrences for each azimuth. These selected directions are now potential DoA. Knowing the number of sound sources, we used a k-means clustering algorithm. The centroids of the clusters are the estimated DoAs.

- DEMiX: We set the number of neighbors K to 20. We chose a frequency neighbor type $\Omega^F_{f,n} = \{f + k, n \mid |k| \leqslant K\}$. The potential directions are the one that have a larger confidence measure. Therefore, we set a threshold for the confidence measure, and it was equal to the mean over the max of each frequency bin. In this approach, there are many occurrences of potential directions as well. So we set another threshold for the number of occurrences, and it was similar to the used one for DUET. The directions that achieved all the required conditions are now potential directions. Knowing the number of sound sources, we used a k-means clustering algorithm. The centroids of the clusters are the estimated DoAs.

- HARPEX: For each time-frequency bin, we have two potential directions. We computed all of them and set an occurrence condition that is similar to DUET. The ones that succeeded in the set condition are taken into consideration as potential directions. Similarly to the other approaches with the knowledge of the number of sound sources, we used a k-means clustering algorithm. The centroids of the clusters are the estimated DoAs.

- DIRAC: This approach is similar to the adaptation of DUET. The diffuseness coefficient could not be used as a condition since it measures the amount of diffuseness in a time-frequency bin and not if only one sound source is active. If we used it with the occurrence condition, it would be precisely a DUET.

- MVDR and MUSIC: There were no parameters to set with these approaches. However, we found some problems to pick the DoA from the cartography. The cartography is visually appealing to the eyes. The sound source DoA can be

[6]We chose this name to refer to the sound source separation approach in DUET, which is a time-frequency analysis approach without any test on the reliability of the treated time-frequency bins.

roughly estimated with human eyes. But finding a suitable way or algorithm to extract them automatically was difficult. Therefore, we report the measures for the cases where only one sound source is active in which the extraction of the DoA was smooth, and it corresponds to the maximum of the spectra.

4.2.2 Results and discussion

First, we will present a visual representation of the DoAs estimation to help the reader visualize how the DoA are extracted. We considered a complex scenario randomly chosen from the data set. The chosen situation was $RT_{60} = 0.7s$, and the number of sound sources is equal to three. The visual representation is given in Fig 4.5.

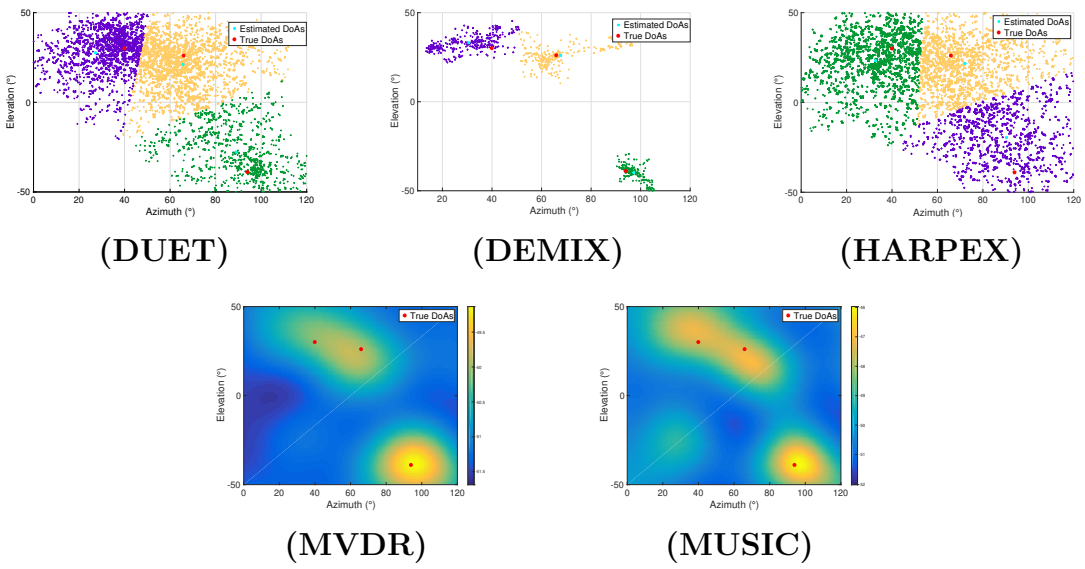


Fig. 4.5 Visual representation of the surveyed DoA approaches. For this example, the time reverberation was 0.7s, and the number of sound sources was 3.

Second, the results in terms of the accuracy and angular error are given in Table 4.1 for one sound source, in Table 4.2 for two sound sources and Table 4.3 for three sound sources. Note that the accuracy and the mean error over the examples are given for the angle tolerance of 5°, 10° and 15° for each scenario.

Third, we present the angular error in terms of azimuth and elevation with no angle tolerance in Fig. 4.6 for the complex scenario ($RT_{60} = 0.7s$, and three sound sources are present in the sound scene).

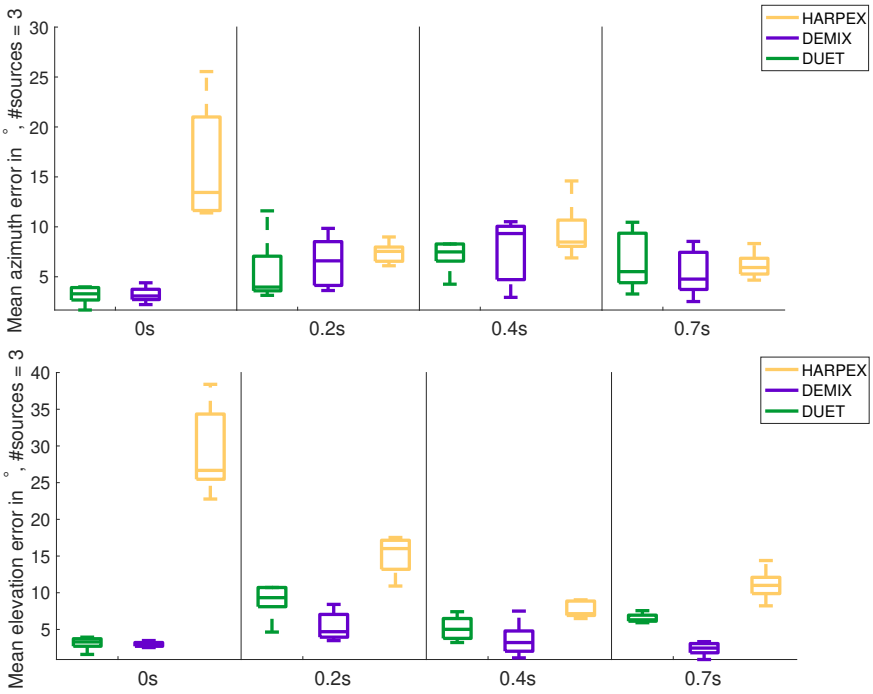


Fig. 4.6 Angular error with no angular tolerance constraint for the cases where three sound sources are present in the sound scene. The outliers are not represented.

In Fig. 4.5, in the top row, we propose a visual representation of the process of estimating the DoAs with the methods DUET, DEMIX, and HARPEX. We chose to represent all the potential directions on cartography that succeeded to fulfill the following conditions:

- DUET: the number of occurrences must be larger than the proposed threshold in Section. 4.2.1.3.
- DEMIX: the number of occurrences and the confidence measure must both of them be larger than the proposed thresholds in Section. 4.2.1.3.
- HARPEX: the number of occurrences must be larger than the proposed threshold in Section. 4.2.1.3. This option was not proposed in the original HARPEX approach, but we propose to add this constraint to use HARPEX as a localization algorithm.

As we can see in the top row of Fig. 4.5, the above conditions help to discard wrong directions that are given by time-frequency bins where the approximately (W-DO) hypothesis is not true. We can see that with DEMIX, more false directions are discarded. This must be because we have one more condition with which we can judge the reliability of a time-frequency bin. HARPEX seems to have more potential directions that

succeeded to fulfill the set condition on the number of occurrences. This is clearly due to the fact that each time-frequency bin gives two potential directions.

In the bottom row of Fig. 4.5, we showcase the cartography that is given by the MVDR and the MUSIC approaches. Visually we can see that these approaches help to locate the sound sources. In this particular case, we have two sound sources that are close to each other. With the MVDR approach, we can see that the peaks are located where the true location of the sound sources are. However, for the close sound sources, we only have one broad peak between the location of the two sound sources. We can see that MUSIC is an improvement of the MVDR approach because, with this approach, that broad peak got refined into two peaks. Note that the bottom row was generated using the whole 25 ambisonic channels, unlike the top row, where only four channels were used. Indeed, the MVDR and the MUSIC cartography resolution get refined if more channels are used. Using only the first four channels results in a bad accuracy of the peaks (very large). We do not compare these approaches to the other, one and therefore the fairness of the comparison will not be an issue. With a fair comparison, these approaches (MVDR and MUSIC) are not as performant as the other one.

In terms of accuracy, we can see as showcased in Table 4.1 Table 4.2 and Table 4.3, DEMIX is the best approach with which we can have at least 73% of accuracy for an angle tolerance of 10° whatever is the scenario. DUET seems to have decent results as well, with an accuracy of at least 40%. When it comes to HARPEX, this approach seems to struggle more compared to the other ones, which is surprising since this approach is highly regarded in the ambisonic domain. In terms of angular error on the whole examples we can say that DEMIX comes in the first position followed by DUET and finally HARPEX in the final position. A sample of this summary can be seen in Fig. 4.6.

# src	Approach	RT_{60} (s)	Accuracy (%)			Mean error (°)			
			5°	10°	15°		5°	10°	15°
1	DUET	0	100	100	100	θ	1.7	1.7	1.7
						ϕ	4.3	4.3	4.3
		0.2	40	100	100	θ	2	3.4	3.4
						ϕ	2.3	2.3	2.3
		0.4	40	100	100	θ	1.7	1.9	1.9
						ϕ	3.9	7.4	7.4
		0.7	40	100	100	θ	4.2	4.2	4.2
						ϕ	3.8	5.8	5.8
	DEMIX	0	100	100	100	θ	3.1	3.1	3.1
						ϕ	4.8	4.8	4.8
		0.2	80	100	100	θ	0.8	2.9	2.9
						ϕ	1.6	1.6	1.6
		0.4	40	100	100	θ	1.3	1.3	1.3
						ϕ	3.7	5.1	5.1
		0.7	40	100	100	θ	1.7	9.1	9.1
						ϕ	3.9	4.9	4.9
	HARPEX	0	50	100	100	θ	2.2	6.1	6.1
						ϕ	4.1	4.3	4.3
		0.2	40	100	100	θ	3	7	7
						ϕ	3.9	4.5	4.5
		0.4	30	100	100	θ	3.6	7.4	7.4
						ϕ	3.5	4.9	4.9
		0.7	20	100	100	θ	4.2	8.3	8.3
						ϕ	2.7	4.5	4.5
	MVDR	0	100	100	100	θ	1	1	1
						ϕ	1.2	1.2	1.2
		0.2	100	100	100	θ	1.9	1.9	1.9
						ϕ	0.6	0.6	0.6
		0.4	100	100	100	θ	0.4	0.4	0.4
						ϕ	0.2	0.2	0.2
		0.7	100	100	100	θ	3	3	3
						ϕ	0.3	0.3	0.3
	MUSIC	0	100	100	100	θ	1	1	1
						ϕ	1.2	1.2	1.2
		0.2	100	100	100	θ	1.9	1.9	1.9
						ϕ	0.6	0.6	0.6
		0.4	100	100	100	θ	0.4	0.4	0.4
						ϕ	0.2	0.2	0.2
		0.7	100	100	100	θ	3	3	3
						ϕ	0.3	0.3	0.3

Table 4.1 Performance of the sound source localization of the approaches surveyed in Chapter 3 Section 3.3 when only one sound source is present in the sound field.

# src	Approach	RT_{60} (s)	Accuracy (%)			Mean error (°)			
			5°	10°	15°		5°	10°	15°
2	DUET	0	100	100	100	θ	2.4	2.4	2.4
						ϕ	2	2	2
		0.2	35	75	90	θ	3.7	5	5.1
						ϕ	2.6	2.6	2.6
		0.4	35	85	100	θ	3.2	4.5	4.5
						ϕ	3.2	3.9	4
		0.7	15	40	85	θ	2.6	4.6	4.6
						ϕ	4.4	6.6	6.8
	DEMIX	0	55	100	100	θ	3.1	3.8	3.8
						ϕ	2.6	6.7	6.7
		0.2	90	90	90	θ	2.1	2.1	2.1
						ϕ	1	1	1
		0.4	50	85	90	θ	3.1	3.1	5.2
						ϕ	2.7	3.1	3.7
		0.7	5	95	100	θ	2.2	2.2	2.3
						ϕ	1.9	6.6	6.6
	HARPEX	0	15	25	40	θ	3.1	3.2	3.2
						ϕ	3.9	4	9.3
		0.2	65	90	90	θ	3	4	4
						ϕ	2	2	2
		0.4	30	75	100	θ	3.5	4.2	4.2
						ϕ	4.1	5.2	5.3
		0.7	30	50	95	θ	2.7	2.9	3
						ϕ	4.2	7.1	7.7

Table 4.2 Performance of the sound source localization of the approaches surveyed in Chapter 3 Section 3.3 when two sound sources are present in the sound field.

# src	Approach	RT_{60} (s)	Accuracy (%)			Mean error (°)			
			5°	10°	15°		5°	10°	15°
3	DUET	0	80	93	96	θ	3.1	3.5	3.5
						ϕ	3	3	4
		0.2	16	60	80	θ	3.6	4.8	5.5
						ϕ	4.6	7.8	9.2
		0.4	23	70	83	θ	4.2	6.7	6.7
						ϕ	3.7	4.9	4.9
		0.7	16	53	83	θ	3.9	5.6	6.5
						ϕ	4.1	6.2	6.7
	DEMIX	0	23	76	96	θ	3.1	3.1	3.1
						ϕ	2.9	5.4	5.9
		0.2	53	73	76	θ	4	6.4	6.4
						ϕ	4.1	5.8	6
		0.4	53	73	76	θ	3.8	4.1	7.6
						ϕ	2.9	5.1	5.4
		0.7	60	76	90	θ	3.8	4.2	5.1
						ϕ	2.2	5.2	5.2
	HARPEX	0	10	13	13	θ	4	5.6	12.3
						ϕ	4.2	7.1	9.2
		0.2	6	40	60	θ	4.6	4.6	7.1
						ϕ	4	7	12.5
		0.4	13	46	53	θ	3.8	5.2	9
						ϕ	3.9	7.5	9
		0.7	10	30	63	θ	4.6	6	6.4
						ϕ	3.9	9.3	11

Table 4.3 Performance of the sound source localization of the approaches surveyed in Chapter 3 Section 3.3 when two sound sources are present in the sound field.

4.2.3 Comparison to the state of the art

In [93, 95][7], the authors proposed a sophisticated approach recently for sound source localization for ambisonic mixtures. It is based on using neural networks. The approach was tested on quite a similar data set with a reverberation time between $0.2s$ and $0.8s$. They obtained an accuracy of 51.6%, 91.1%, and 95.2% for an angular error tolerance of 5°, 10° and 15°, respectively. We can say that we are in the same magnitude order if we average the accuracy over all the reverberation times. For instance, with three sound sources, the average accuracy for DEMIX is 47.25%, 74.5%, and 84.5% for an angular error tolerance of 5°, 10° and 15°, respectively. We do not have enough information on their simulated RIRs. We do not know if the ambisonic mixtures were created from

[7]Note that we didn't implement this approach, nor we had its code. However, we simulate our data set to be close to the ones presented in the baseline approach dedicated articles. The reported accuracies comes from these articles.

perfect ambisonic RIRs or microphone ones, as our case. The only way to compare the approaches in an honest way would be to run tests on the exact same test data set. Note that with the old-fashioned approaches there is no need to generate a huge amount of RIRs to train them. They can be applied straightforward to any ambisonic mixture. In [93], their approach was compared to the one proposed in [10], which gave an accuracy of 27.5%, 56.5%, and 71.2% for an angular error tolerance of 5°, 10° and 15°. We can say that the approaches that we surveyed and adapt to ambisonics outperform the one in [10]. Details of the method weren't available and prevented a full reprodictability at the time when this work was done, the method was under process of being a patented.

Note that the algorithms that we surveyed are not very time consuming. In Table 4.4 we report the execution time on our computer (MacBook pro with 2,2 GHz intel Core i7 processor and 16Go of Ram) for 10 seconds of the ambisonic signals.

Method	Execution time
DUET	8.75s
DEMIX	19.03s
HARPEX	19.98s
MVDR	1.01s
MUSIC	1.01s

Table 4.4 Execution time for 10 seconds of ambisonic signals.

In Section 4.3, we will evaluate the surveyed sound source separation approaches.

4.3 Evaluation of the sound source separation bricks

In this section, we experiment a little bit with the surveyed approaches of sound source separation. First, we will compare all the plane wave decomposition approaches discussed before. Second, we will evaluate the performance of the Oracle time-frequency masks to have a reference for the range of SDR, SIR and SAR values, as well as have an idea of the performance of such type of sound source separation.

4.3.1 Plane wave decomposition

In this section, we compare the performance of the different types of plane wave decomposition. To this aim, let us consider a spherical microphone array (the one modeled in Appendix B), and two sound sources, both of them at the same elevation as the spherical microphone array. The sound sources were positioned at the circle with a similar center as the spherical microphone array and with a radius 2.5m. The room dimension and reverberation time were fixed. The reverberation time $RT_{60} = 0.35s$. The position of the first sound source was fixed, the second sound source was at a different position but still at the same distance from the microphone array. 36 RIR were generated corresponding to the angle between the sound sources. The first RIR corresponds

to an angle difference of $\theta = 5^\circ$. For each new RIR, the angle difference gets larger by 5°, which makes the last RIR corresponds to the angle difference of $\theta = 180^\circ$. 14 tracks were randomly chosen from DSD100[8] data set and randomly grouped in pairs. In order to get the spherical microphone array mixtures, each pair was convolved with the generated spherical microphone array RIR. This corresponds to $36 \times 14/2 = 252$ different examples. The computed spherical microphone array mixtures were encoded to get the ambisonic signals (see Appendix B from more information on the generation of our simulations). For each ambisonic signals, we applied the listed beamformers in order to extract both sound sources.

The performance of the plane wave decomposition was judged by the introduced energy ratios in Chapter 3 Section 3.4.1. The results of the plane wave decomposition were compared to the ground truth sound sources signal at the position of the microphone array. The performance of the different plane wave decompositions are displayed in regards to angle difference between the sound sources. In Fig. (4.7a, 4.7b, 4.7c), the energy ratios are average over both the sound and the examples.

In terms of SDR, when the sound sources are close to each other:

- The MIXED beamformer and the regularized PIV beamformer seems to give the best performances when the sound sources are close to each other.
- The regularized PIV beamformer seems to give better scores than regular PIV beamformer.
- The MVDR beamformer seems to give better results than both static approaches (basic projection and PIV).
- The LCMV seems to give lower scores than the MVDR.
- The basic projection seems to give over all the worst scores.

In terms of SDR, when the sound sources are far from each other:

- The regularized PIV beamformer and the regular PIV beamformer seems to have similar scores and the best ones.
- The MIXED beamformer keeps giving great scores compared to other beamformer.
- The MVDR beamformer and the LCMV beamformer seems to gives similar scores, which are lower than the static approaches.
- The basic projection performances seems to be improved.

[8]The DSD100 is a dataset of 100 full lengths music tracks of different styles along with their isolated drums, bass, vocals and others stems. `https://sigsep.github.io/datasets/dsd100.html`.

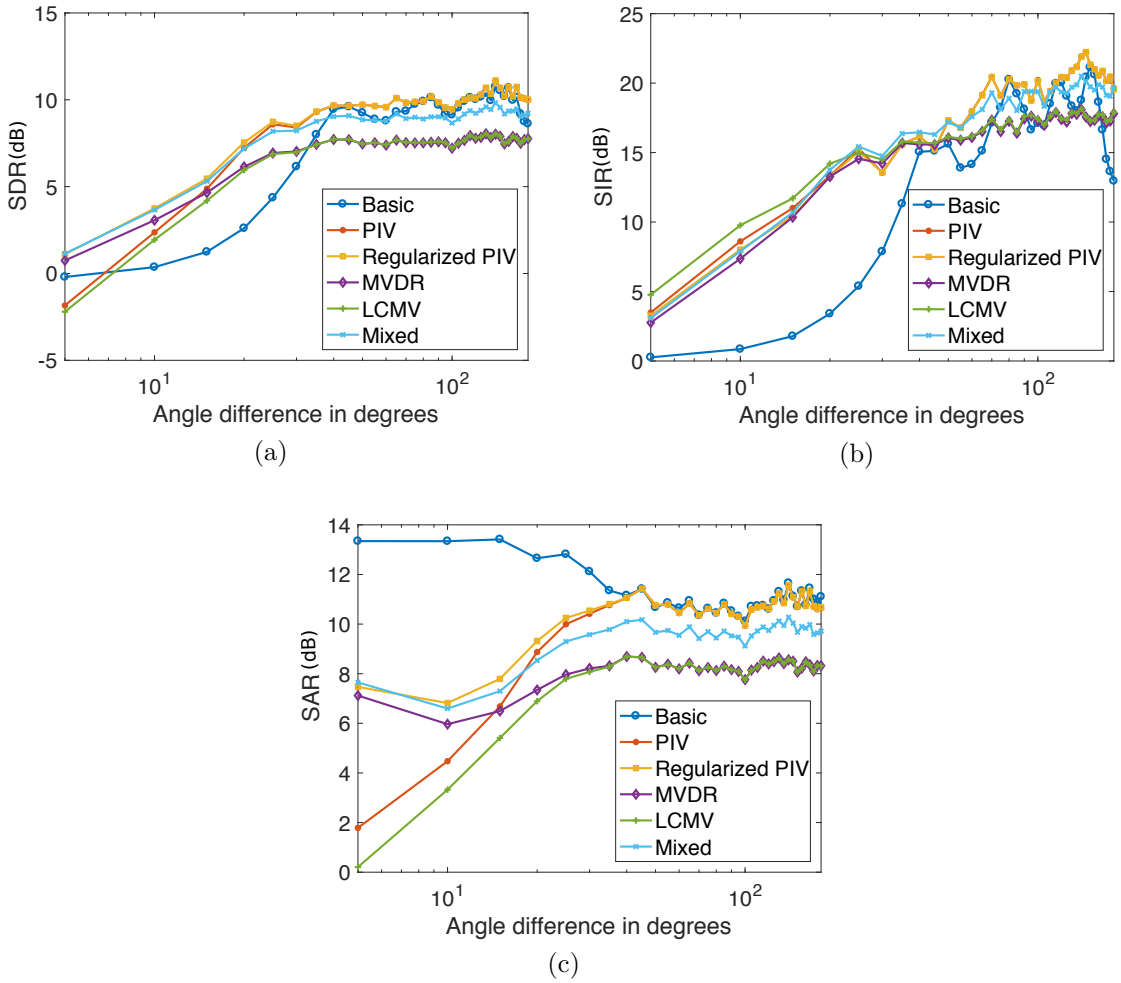


Fig. 4.7 Performance of the beamformers in terms of SDR, SIR, and SAR in dB. Scores are averaged over all the examples. Note that "Regularized PIV" refers to regularized pseudo-inverse, and "PIV" to a plain pseudo-inverse.

It isn't surprising that the MIXED beamformer and the regularized PIV beamformer gave similar scores when the sources were close to each other, because by definition the covariance matrix of the PIV beamformer is part of the MIXED beamformer covariance matrix. On the one hand, it is reasuring that over all the regularized PIV beamformer gave better scores than PIV beamformer because the first one is supposed to be an improvement of the second one by setting more constraints. On the other hand LCMV is supposed to be an improvement of the MVDR by setting strict constraint toward the interference sound sources, but it does not seem to be the case in terms of SDR. However, in terms of SIR we can see the improvement because over all the LMCV beamformer approach gives way betters cores than the MVDR beamformer.

In terms of SAR, the LCMV introduce lot of artifacts followed by the pseudo-inverse beamformer for close sound sources, and the MVDR.

Overall the approach that we proposed in Chapter 3 Section 3.4.4.1 (regularized PIV) is the best compromise for the plane wave decomposition. Therefore, it is the

recommended beamformer for the beamforming block in Fig. 4.2.

4.3.1.1 Time-frequency masking

In this section we present some numerical simulations that illustrate how oracle time-frequency masks behave in the context of HOA recordings.

Let us consider the signals of a recorded ambisonic sound field. The sound field contains four sound sources.

The room dimension and the boundaries coefficient of reflection, which influences the reverberation time, were set in the simulation software in order to have a random reverberation time between [0.2s 0.8s]. We added a diffuse noise with a random SNR between $[0dB, 25dB]$. With this simulation software, we are able to have at our disposal the ambisonic signals corresponding to the separate contribution of each source. With this, we can compute both masks as in Eq. (3.67) and Eq. (3.66), and apply them on the ambisonic mixture in the time-frequency domain as follows:

$$\hat{s}_{m,j,f,n} = M_{m,j,f,n} z_{m,f,n}, \tag{4.18}$$

with m being the channel of the ambisonic signals. For the OBM we took into consideration several values of the threshold value $\eta = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$.

In Fig. 4.9 (top), we present the amplitude spectrogram of the 1st channel of the mixture and the first source, respectively. In Fig. 4.9 (middle), we show the OSM of the first source and the estimation of the first source in the first channel by applying the OSM to the mixture. Similarly, in Fig. 4.9 (bottom) we show the OBM and the corresponding estimated source signal for $\eta = 0.5$.

We listened to the source signals estimated using the two methods. The separation works in both cases, and the interference sound sources were suppressed. However, the rendering of the OBM is different from the rendering of the OSM. The source signal estimated using the OBM presents more hearable artifacts than that estimated using the soft mask. However, the OBM seems to suppress interference more efficiently.

Table 4.5 presents the value of the SDR, SAR and SIR obtained using the soft mask and the OBM for the different threshold values.

The results presented in Table 4.5 confirm our informal listening observations. Indeed with an OSM, fewer distortions and artifacts are present in the separated sources as confirmed by the SDR and SAR scores. Not that the performances can be higher if the mixture did not contain diffuse noise with high SNR.

For the OBM, the threshold η seems to have an influence on the scores. As we can see in Table 4.5, and Fig. 4.8, the SDR increases with $\eta \in]0, 0.5]$, and then it decreases for values between $\eta \in]0.5, 1[$. The SAR seems to decrease when η getting larger. However, the larger η, the larger the SIR.

Each approach has its own usage. We can say that the OBM should be used in cases where interference suppression is the priority, such as in telecommunication applications. However, in the case of multimedia applications, such as 6-DoF navigation, using a soft masking method such as the Wiener filter may be more appropriate.

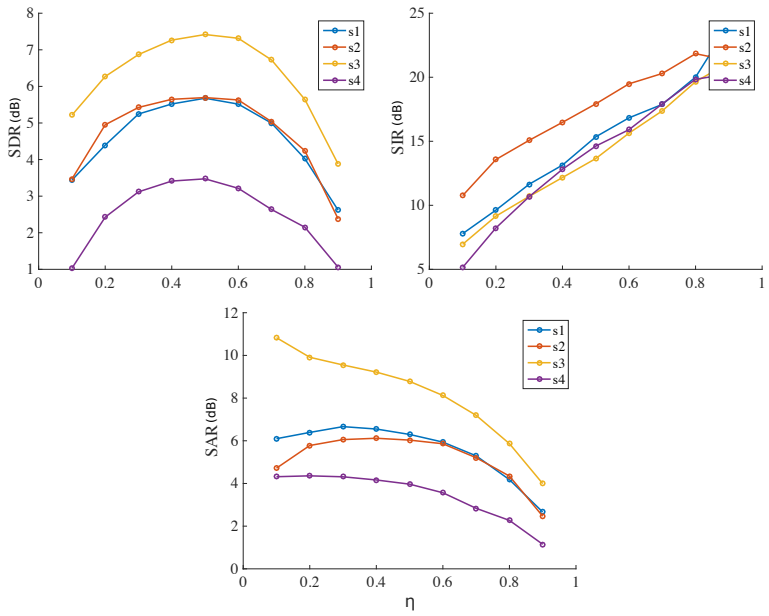


Fig. 4.8 The resulted energy ratios after applying a OBM in regards of the threshold parameter η. The scores are in dB.

4.3.1.2 Conclusion and discussion

We discussed previously in Chapter 3 Section 3.4.4 that in the ambisonic domain, sound sources could be separated using simple beamforming techniques, which rely only on spatial information. In the presence of complex sound scenes, however, the quality of the separation could be improved using multichannel source separation approaches. However, as discussed in Section 4.1.2 we recommend to decompose the contributions of each source in each channel. We recommend specifically the approach we proposed in Chapter 3 Section 3.4.4.1 (regularized PIV) whenever the number of channels is larger than the number of sound sources.

By multichannel sound source separation, we mean looking for the contribution of each source in each channel. OBMing approaches could be considered as a multichannel sound source separation. It is possible to apply such algorithms in the ambisonic domain. However, we showed in Section 4.3.1.1 that such approaches end up giving a poor SAR scores, which means they introduce so many artifacts. The scores were computed from the exact contributions, which means these scores are the best that we could get from a OBMing approach.

The Wiener filter is known in the literature as a smoothing filter that introduces fewer artifacts. We got in Section 4.3.1.1 a sample of its performances by computing the scores given by the OSMs and compare them to scores given by the OBMs. Note that with the OSM, a single-channel Wiener filter was used, it was computed from the spectra of the true sound source signals. We wonder how the performances will improve if the spatial aspect was modeled as well and used along the spectral aspect to create a Wiener filter to perform the sound source separation.

	Sources	SDR (dB)	SIR (dB)	SAR (dB)
OSM	s_1	**6.7724**	12.7773	**8.2503**
	s_2	**6.8331**	15.6374	**7.5633**
	s_3	**8.4121**	12.3417	**10.9103**
	s_4	**4.7277**	10.5677	**6.4042**
OBM $\eta = 0.1$	s_1	3.4384	7.7844	6.0978
	s_2	3.4689	10.7558	4.7172
	s_3	5.2143	6.9523	10.8300
	s_4	1.0296	5.1478	4.3163
OBM $\eta = 0.2$	s_1	4.3942	9.6249	6.3918
	s_2	4.9514	13.5756	5.7792
	s_3	6.2708	9.1523	9.9121
	s_4	2.4263	8.2325	4.3575
OBM $\eta = 0.3$	s_1	5.2457	11.6463	6.6630
	s_2	4.9514	15.0759	6.0560
	s_3	6.8701	10.6931	9.5515
	s_4	3.1179	10.6763	4.3124
OBM $\eta = 0.4$	s_1	5.5196	13.1393	6.5504
	s_2	5.4250	16.4778	6.1182
	s_3	7.2654	12.1682	9.2179
	s_4	3.4126	12.8131	4.1637
OBM $\eta = 0.5$	s_1	5.6762	15.3461	6.2971
	s_2	5.6472	17.9208	6.0318
	s_3	7.4176	13.6484	8.7827
	s_4	3.4733	14.6227	3.9674
OBM $\eta = 0.6$	s_1	5.5157	16.8314	5.9381
	s_2	5.6295	19.4880	5.8605
	s_3	7.3156	15.6372	8.1240
	s_4	3.2146	15.9071	3.5648
OBM $\eta = 0.7$	s_1	4.9948	17.8803	5.2944
	s_2	5.0239	20.2947	5.2060
	s_3	6.7223	17.3779	7.1915
	s_4	2.6351	17.9237	2.8350
OBM $\eta = 0.8$	s_1	4.0242	20.0084	4.1783
	s_2	4.2412	21.8389	4.3457
	s_3	5.6432	19.6518	5.8661
	s_4	2.6351	19.8407	2.2703
OBM $\eta = 0.9$	s_1	2.6153	**24.1358**	2.6628
	s_2	2.3780	**21.3504**	2.4660
	s_3	3.8773	**21.1450**	3.9927
	s_4	1.0564	**20.2628**	1.1495

Table 4.5 Comparing the OSM approach to OBM approach. The best scores are in bold.

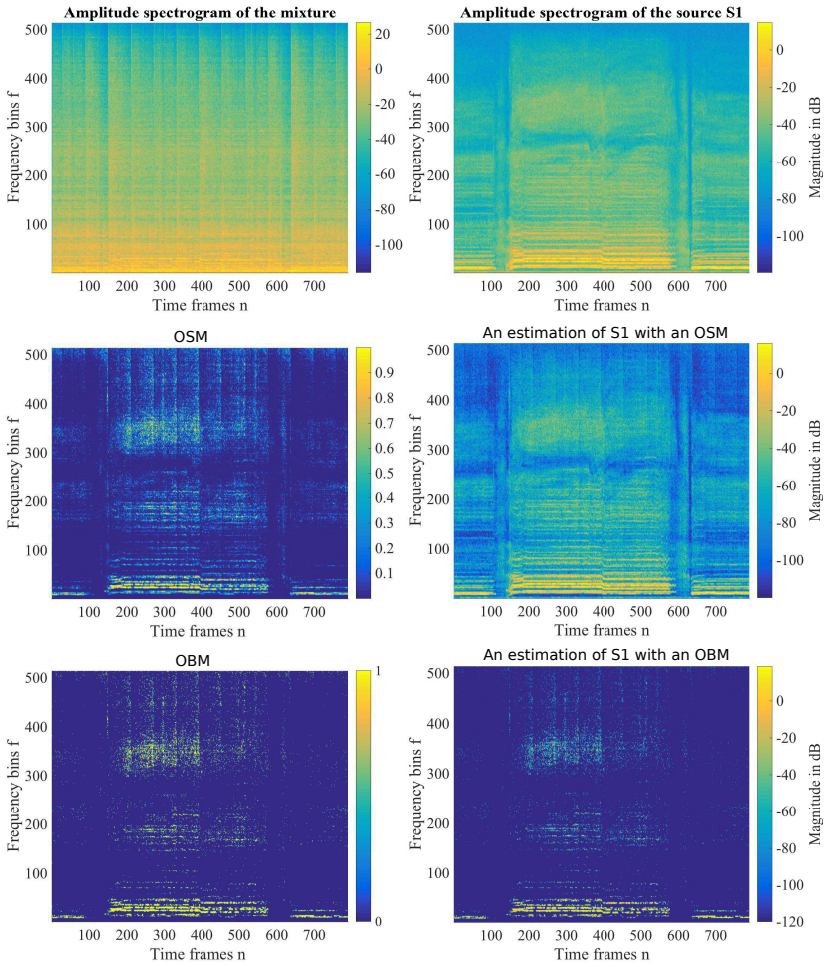


Fig. 4.9 Top: Spectrogram of the mixture's first channel and the source s_1.
Middle: The estimated spectrogram of the source $s1$ (right) by applying the OSM (left) to the first channel of the mixture, the OSM was computed using Eq. (3.66).
Bottom: The estimated spectrogram of the source $s1$ (right) by applying the OBM (left) to the first channel of the mixture, the OBM was computed using Eq. (3.67), here $\eta = 0.5$.

In the literature, there is a need of contributions in terms of multichannel sound source separation in the ambisonic domain. We surveyed in Chapter 3 Section 3.4.3 a microphone approach that helps to simplify the expression of the Wiener filter and estimate its parameters. This approach is based on a strong assumption on the contributions. The contributions in the ambisonic domain are not the same as in the microphone domain. Therefore, in Chapter 5, we will derive the equation of the model and check if such a model can be adapted to the ambisonic domain.

Since last year three contributions have emerged so far in terms of multichannel sound source separation in the ambisonic domain: the one presented in [94] for speech enhancement, the one presented in [85] based on a multichannel non-negative matrix factorization and our contribution [50], which is discussed with more detail in Chapter 5.

In the following, we experiment with our navigation strategy to evaluate its performance compared to state of the art.

4.4 Validation of the navigation approach

In this section, we evaluate the performance of our strategies for navigation. To this aim, the same procedure as in [2] (see Section 2.6) was adopted.[9] We generate two ambisonics room impulse responses for each simulation:

- The first RIR corresponds to the mixture to which the navigation is going to be applied.
- The second RIR as the ground truth mixture.

In other words, the second RIR corresponds to the mixture after using perfect navigation on the mixture corresponding to the first RIR. As an objective quality metric, we use the same one as in [2].

4.4.1 The objective quality metric

The Virtual Speech Quality Objective Listener (ViSQOL) is a metric that models human speech quality using a spectro-temporal measure of similarity between a reference and a test speech signal [51]. The comparison is made in terms of one channel. Given the fact that we have $(l+1)^2$ channels per mixture, it is essential to combine as much information in one channel before the comparison. To this aim, we binauralize our ambisonic mixtures before the comparison. The ViSQOL gives a score called the ViSQOL MOS_LQO score. It is between one and five; the higher the score, the closest the test

[9] The approach was discussed before in Chapter 2.

speech signal to the reference one. Since the ViSQOL is a metric that models human speech quality to measure the similarity of the test signal to the reference one, in this evaluation, we only consider speech signals.

Note that we are aware that this metric doesn't take into consideration the spatial quality of the navigation. For that, the only way to judge the performance of the navigation correctly is through some subjective tests.

4.4.2 Simulation setup

We generated two SRIR for each source in each simulation, the first one as an input for the approaches to be tested, and the second one as a reference output. We already explained our simulation protocol in Appendix B. We adopted the same protocol, and therefore, we validate the translation of a fourth-order ambisonics sound field for the proposed strategies.

We considered four rooms; two small (2.7 m width x 3 m depth x 2.4 m height), medium (4 m width x 5 m depth x 3m height), and large (5m width x 6m depth x 3.5 m height). The rooms' reflection coefficient of the boundaries are frequency independent and were fixed to 0 for the first small room, 0.7 for the second small room, 0.8 for the medium room, and finally 0.9 for the large room. We computed the reverberation time RT_{60} of each room; they are frequency-dependent, we presented them in Fig. 4.10.

We considered four sound sources, placed randomly in the rooms. We randomly placed the spherical microphone array for the first RIR, and strategically for the second RIR; we considered several cases such as the spherical microphone array is close to a sound source, far from all the sources, close to a wall, or in the center of the room. For each room, we considered 450 examples, corresponding to 15 different speech conversations randomly chosen from the TSP McGill speech database [2], and 30 different sound scene configurations. Therefore, we obtained 1800 examples. After the generation of room impulse responses, we identified the complex cases, for instance, sound sources are close to each other, or the spherical microphone array is close to a specific sound source. We considered a smaller dataset with these examples; we spotted around 20 sound scene configurations, which gave us a sub-dataset with 1200 samples.

First, we generated the contribution of each source in each microphone by convolving their speech signal, with its RIR of the spherical microphone array. Second, we computed the contribution of each source in each channel by encoding the results of the first step into ambisonics signals. The outcomes of the second step (the contribution of each source in each channel) are going to be considered as the inputs of our second approach. Third, we generate the fourth-order ambisonics mixture by summing all the contributions computed in the second step.

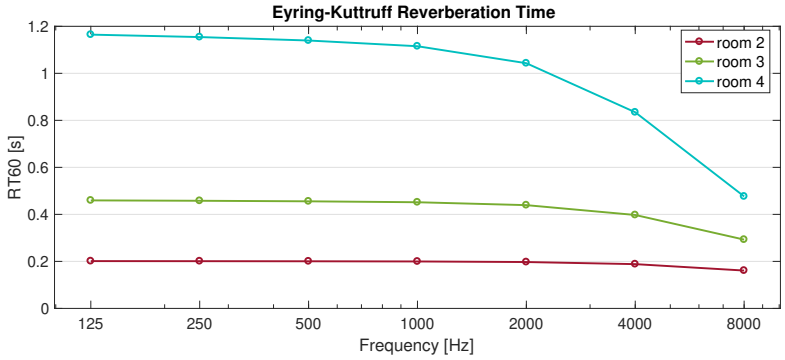


Fig. 4.10 The reverberation time RT_{60} of each room.

4.4.3 Considered methods

For simplicity, we consider here that the room size, the position of the spherical microphone array, and the position of the sound sources are known. This allows us to have a clear idea of the distances.

We considered three methods: Our first approach with a mixed plane wave decomposition, the method proposed in [2] (referred to it as c), and our second approach corresponding to the multichannel sound source separation (referred to it as d).

For our first approach, we considered two variants: First, we took into consideration the first echoes by adapting them to the current user position (referred to it as a). We searched for the echoes DoA using the matching pursuit algorithm described in Alg. 5. Second, without taking into consideration any echo (referred to it as b).

For our second approach, we considered a perfect multichannel sound source separation by taking as inputs the contribution of each source in each channel. To each one of them, we apply the matching pursuit algorithm described in Alg. 5 to decompose them into plane waves.

This study aims to confirm the fact that the multichannel sound source separation helps better to decompose the ambisonic sound field into plane waves.

4.4.4 Results and discussion

Box-plot results for the experiments on the entire dataset are shown in Fig. 4.11 (top), and on the sub-dataset in Fig. 4.11 (bottom). Based on the MOS_LQO score, we can see that our second approach outperforms all the other methods. We can see that when the sound source configuration is involved, we obtained a much larger score compared to methods that are only based on a spatial decomposition.

Our first approach seems to have similar behavior as the one proposed in [2]. We computed the ΔMOS_LQO which is a subtraction between the MOS_LQO of one of our strategies with the method proposed in [2]. The results are given in Table 4.6.

It looks like the reverberation time influences the navigation with all the considered approaches. When the RT_{60} gets larger, the MOS_LQO score decreases. This influence is less significant for our second approach.

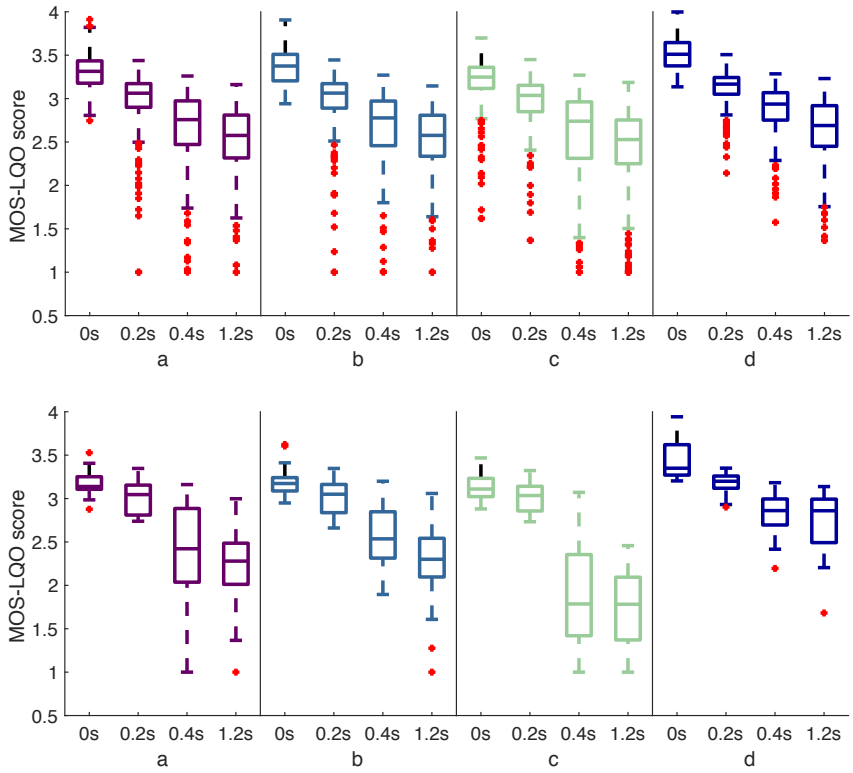


Fig. 4.11 Binaural MOS-LQO scores. The MOS_LQO score was computed on the whole datasat for the top, and on the sub-data set for the bottom. Recall that $MOS\_LQO \in [1, 5]$, the higher the better

	$RT_{60}(s)$	0	0.2	0.4	1.2
$MOS\_LQO_a - MOS\_LQO_c$	max	1.19	0.86	1.33	1.23
	min	-0.28	-0.80	-1.04	-0.54
$MOS\_LQO_b - MOS\_LQO_c$	max	1.46	0.59	1.25	1.56
	min	-0.14	-0.97	-1.09	-0.50
$MOS\_LQO_d - MOS\_LQO_c$	max	1.72	1.80	1.90	2
	min	-0.038	-0.20	-0.51	-0.71

Table 4.6 Comparing our strategies to the one presented in by computing the ΔMOS_LQO.

4.4.5 Conclusion

Using a multichannel sound source separation before the plane wave decomposition gives better results than the best approach of state of the art in terms of reconstruction of the sound field, especially when the sound scene is complex with sound sources close to each other or the microphone. One can consider our comparison to be biased because we had the sound sources' true contributions. However, note that we want to showcase that applying a plane wave decomposition on the contributions instead of the mixture directly decomposes the mixture better (we avoid having interference sources while decomposing the mixture into plane waves.) and therefore has better navigation, as shown by the numerical experiment. To our knowledge, there isn't a navigation contribution that uses the same approach as our second strategy that relies on decomposing the ambisonic sound field into the sound source contributions. The used objective metric does not take into consideration the spatialization aspect of the reconstruction. A more reliable way to judge the performance of these approaches is through some subjective tests. Although in [2], they performed extensive informal listening, which resulted in being consistent with the objective results using the same objective metric (MOS_LQO).

For the next of this manuscript, we will concentrate on approaching these contributions with the multichannel sound source separation in the ambisonic domain and trying to be close to the oracle contributions.

4.5 Conclusion

In this chapter, we explained our strategy to respond to the main goal of my Ph.D. subject, which is allowing the user to navigate with 6-DoF in 3D sound fields that were recorded. We based our main approach on decomposing and reconstructing the sound field. We proposed two strategies:

- A plane wave decomposition
- A multichannel sound source separation followed by a plane wave decomposition (We demonstrate the navigation with this strategy in **`https://hafsatimohammed.github.io/HTML_Files/Example_Navigation.html`** with an example. For this experience, we simulated three sound sources in a reverberant environment that were recorded live with an ambisonic antenna and three spot microphones (the reason is revealed in Chapter 6). Each one was close to a given sound source. Using Chapter 6' approach, we applied a multichannel sound source separation to decompose the ambisonic sound scene into ambisonic sound source contributions.

To each output, we applied the matching pursuit algorithm in Alg. 5 to decompose each output (the output contributions) into plane waves. All of this was done before launching the demo. During the demo in real-time, we reconstructed the ambisonic sound field regarding the user's position, which was manipulated using the keyboard.

For both strategies, a sound source localization is required. To this aim, we validated the localization bricks that we adapted in Chapter 3 Section 3.3 through some numerical experiments. It turned out that the adaptation of old approaches in the microphone domain to the ambisonic domain gives sufficiently good results. These results seem to be in the same order of magnitude as the current algorithms of DoA estimation in the ambisonic domain.

We experimented with the sound source separation approaches surveyed in Chapter 3 Section 3.4. We concluded on the lack of performance in terms of SAR for time-frequency masking approaches. This method introduces artifacts, which is unsuited to our application.

We experimented with all the plane wave decomposition approaches discussed before. We concluded on the outperformance of the proposed approach (regularized pseudo-inverse) compared to the rest when the number of channels is larger than the number of sound sources, which is most of the time the case. These plane wave decomposition approaches can be used for both navigation strategies that we proposed. We recommend the regularized pseudo-inverse for both strategies.

When it comes to multichannel sound source separation in the ambisonic domain, there is a lack of existing contributions, which is going to be the main axis of research for the next part of my manuscript. This decision was encouraged by the results we got in the last section. We finished this chapter by checking the principle of our navigation approach. This was done by using an objective metric. We compared both of our strategies to one of the best methods in state of the art. It turned out that in all cases, that the strategy that we proposed (navigation with multichannel sound source separation followed by a plane wave decomposition) outperformed state of the art.

This second strategy is based on a multichannel sound source separation, which is a field of research that is lacking in the ambisonic domain. To this aim, the next part of this manuscript will face and discuss this problem.

Chapter 5

Multichannel decomposition of HOA sound fields using the local Gaussian model

In this chapter we summarize the work published in [50] . We investigate how the local Gaussian model (LGM) can be applied to separate sound sources in the higher-order ambisonics (HOA) domain. First, we show that in the HOA domain, the mathematical formalism of the local Gaussian model remains the same as in the microphone domain. Second, using an off-the shelf source separation toolbox (FASST) based on the local Gaussian model, we validate the efficiency of the approach in the HOA domain by comparing the performance of toolbox in the HOA domain with its performance in the microphone domain (considering an informed case where the sound sources DoA is known). To do this we discuss and run some simulations to ensure a fair comparison. Third, we check the efficiency of the local Gaussian model compared to other available source separation techniques in the HOA domain. Simulation results show that separating sources in the HOA domain results in a 1 to 12 dB increase in signal-to-distortion ratio, compared to the microphone domain.

5.1 Mixture model

5.1.1 The mixture model in the microphone domain

In this section, we recall briefly the mixture model in the microphone domain. For more information about the mixture model see Chapter 3 Section 3.2.1.

By term identification between the left and the right of the equal sign in Eq. (3.11), we can write the contribution of each source in each microphone in the time frequency

domain under the narrow band approximation as follows:

$$\mathbf{c}_{j,f,n} = \mathbf{A}_{j,f}\mathbf{s}_{j,fn}\,. \tag{5.1}$$

The estimation of the $\mathbf{c}_{j,t}$ can be addressed using the multichannel Wiener filtering framework, which will be presented with more details in Section 5.2. This framework requires to select a distribution model for the variables to estimate. For simplicity we use the local Gaussian model presented in [123]:

$$\forall f \in [1, F]\,, n \in [1, N]\,, \quad \mathbf{c}_{j,f,n} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}})\,, \tag{5.2}$$

where $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = \mathbb{E}\left[\mathbf{c}_{j,f,n}\,\mathbf{c}_{j,f,n}^{\mathsf{H}}\right]$ is the covariance matrix of the contribution of the j^{th} source to every microphone at frequency f and time frame n. In line with the literature, this matrix can be further decomposed as the product of a scalar spectral part, $v_{j,f,n} = |\mathbf{s}_{j,fn}|^2$, with a time-invariant spatial matrix, $\mathbf{R}_{\mathbf{c}_{j,f}}$, as follows: $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} = v_{j,f,n}\,\mathbf{R}_{\mathbf{c}_{j,f}}$. Notably, the so-called spatial covariance matrix $\mathbf{R}_{\mathbf{c}_{j,f}}$ respects the relation $\mathbf{R}_{\mathbf{c}_{j,f}} = \mathbf{A}_{j,f}\,\mathbf{A}_{j,f}^{\mathsf{H}}$ when the assumptions of Eq. (3.11) hold. This can be found by using Eq. (5.1) and assuming that $\mathbf{s}_{j,f,n}$ is a random variable and $\mathbf{A}_{j,f}$ is deterministic, which is given follows:

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} &= \mathbb{E}\left[\mathbf{c}_{j,f,n}\,\mathbf{c}_{j,f,n}^{\mathsf{H}}\right] && (5.3)\\
&= \mathbb{E}\left[\ (\mathbf{A}_{j,f}\mathbf{s}_{j,fn})(\mathbf{A}_{j,f}\mathbf{s}_{j,fn})^{\mathsf{H}}\right] && (5.4)\\
&= \mathbb{E}\left[\ \mathbf{A}_{j,f}\mathbf{s}_{j,fn}\mathbf{s}_{j,fn}^{\mathsf{H}}\mathbf{A}_{j,f}^{\mathsf{H}}\right] && (5.5)\\
&= |\mathbf{s}_{j,fn}|^2\mathbf{A}_{j,f}\,\mathbf{A}_{j,f}^{\mathsf{H}}. && (5.6)
\end{aligned}$$

5.1.2 The mixture model in the HOA domain

In the Higher-Order Ambisonic (HOA) framework, the sound field is decomposed over a basis of spherical harmonic functions. As explained in Chapter 2 Section 2.3, the HOA signals, $\mathbf{z}_t$ are typically obtained by applying a set of finite impulse response filters, known as encoding filters, to the signals recorded by a spherical microphone array [83]. Thus, assuming the encoding filters are short enough, the vector of the HOA signal STFTs $\mathbf{z}_{f,n} \in \mathbb{C}^M$ is given by:

$$\mathbf{z}_{f,n} = \mathbf{E}_f\,\mathbf{x}_{f,n}, \tag{5.7}$$

where $\mathbf{E}_f$ is the matrix of the encoding filter frequency responses. Using Eq. (3.11),

we can now model the HOA mixture as follows:

$$\mathbf{z}_{f,n} = \sum_{j=1}^{J} \mathbf{E}_f \, \mathbf{c}_{j,f,n}, \tag{5.8}$$

and identify the contribution of the j^{th} source to the different HOA channels as:

$$\mathbf{b}_{j,f,n} = \mathbf{E}_f \, \mathbf{c}_{j,f,n}. \tag{5.9}$$

As is the case in the microphone domain, in the ambisonic domain source separation consists in estimating the contribution of every source to every channel $\mathbf{b}_{j,f,n}$, which can be solved using a Wiener filtering approach. To this aim we assume the following local Gaussian model:

$$\mathbf{b}_{j,f,n} \sim \mathcal{N}_c(0, \mathbf{\Sigma}_{\mathbf{b}_{j,f,n}}). \tag{5.10}$$

Similar to the microphone domain, the covariance $\mathbf{\Sigma}_{\mathbf{b}_{j,f,n}} = v_{j,f,n} \, \mathbf{R}_{\mathbf{b}_{j,f}}$ can be further decomposed into a spectral part, $v_{j,f,n}$, and a spatial covariance matrix given by:

$$\mathbf{R}_{\mathbf{b}_{j,f}} = \mathbf{E}_f \, \mathbf{R}_{\mathbf{c}_{j,f}} \, \mathbf{E}_f^{\mathsf{H}} \, . \tag{5.11}$$

5.2 Source separation with Wiener filtering

For more information about the Wiener filter refer to Chapter 3 Section 3.4.3. We recall the expression of the Wiener filter in the microphone domain, it is given by:

$$\mathbf{W}_{j,f,n} = \mathbf{\Sigma}_{\mathbf{c}_{j,f,n}} \left(\sum_{j'=1}^{J} \mathbf{\Sigma}_{\mathbf{c}_{j',f,n}} \right)^{-1} . \tag{5.12}$$

Thus, the source separation problem reduces to the problem of estimating the covariance matrices $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}$ or, equivalently in the HOA domain, $\mathbf{\Sigma}_{\mathbf{b}_{j,f,n}}$. Each source contribution is obtained by applying element-wise its corresponding Wiener filter to the mixture: $\hat{\mathbf{c}}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{x}_{f,n}$, $\hat{\mathbf{b}}_{j,f,n} = \mathbf{W}'_{j,f,n}\mathbf{z}_{j,f,n}$ in the HOA domain, respectively. and finally using inverse STFT with overlap-add to reconstruct the time-domain signal.

The estimation of the Wiener filter parameters is done by maximizing the log-likelihood for every time-frequency (f, n). The function to maximize happens to be nonconvex. A possible approach to find the maximum a posteriori is to use an EM algorithm [32, 23, 8, 55], which is an iterative algorithm where first, the desired parameters are initialized, and second, it alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using

the current estimate for the parameters or the initialized one during the first iteration, and a maximization (M) step, which computes the desired parameters maximizing the expected log-likelihood found on the E step.

5.3 The nonnegative matrix factorization constraint

Nonnegative matrix factorization (NMF) is an unsupervised decomposition that favors the sound sources' statistical independence in the sound source separation problems. It was first used for mono-channel sound source separation problems [128, 108, 68, 127]. It has been adapted in the multichannel case by Emanuel VINCENT during his Ph.D. thesis [121]. The property of such modeling of the sound source spectra in the multichannel case reduces the risk of overfitting [120]. The key idea behind the NMF decomposition for sound source separation is to model the sound source spectra as the multiplication of two nonnegative matrices:

$$\mathbf{V}_j = \mathbf{W}_j \mathbf{H}_j \tag{5.13}$$

where the matrix $W \in \mathbb{R}_+^{F \times K_j}$ contains spectral patterns characteristic of the spectrum, the matrix $H \in \mathbb{R}_+^{K_j \times N}$ represents the activation coefficients that approximate the spectrum samples onto the dictionary, and $K_j \in \mathbb{N}$ is the NMF rank, which is smaller than F and N. In each time frequency bin the spectrum of a sound source j is given by:

$$v_{j,f,n} = \sum_{k=1}^{K_j} w_{j,f,k} h_{j,k,n}. \tag{5.14}$$

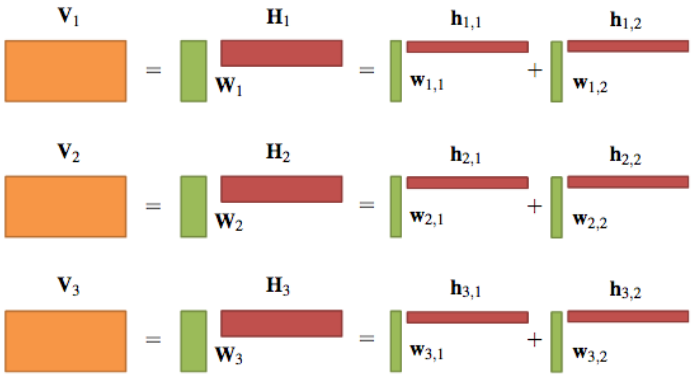


Fig. 5.1 The NMF decomposition of the sources spectra in FASST. Here the number of sources is 3 and the NMF rank $K_j = 2$. Figure from [89].

A visual representation of this decomposition is given in Fig. 5.1. This figure was used in [89]. As you can see, the spectrum of each sound source is modeled as the multiplication of two matrices, which is a linear combination of column vectors with line vectors as it is described in Eq. (5.14).

The NMF constraint is integrated into FASST as an option and can be activated or deactivated. For our case in this chapter, this option was activated. Indeed, we noticed its ability to overcome overfitting.

5.4 Experimental protocol

In this work we use the flexible audio source separation toolbox (FASST) [103, 90], a software toolbox which allows to estimate Wiener filter parameters and apply it. In FASST the parameters are estimated by maximizing the log-likelihood of the observations with an Expectation-Maximization (EM) algorithm, and a multi-channel non negative matrix factorization (NMF) model can be enforced on the source covariances $\Sigma_{c_{j,f,n}}$ [88]. The direct path of the sound sources is going to be considered known for the rest of this chapter.

5.4.1 Dataset

In order to evaluate the source separation performance, we built a dataset as follows. First, fifty songs were randomly chosen from the Mixing Secret Dataset (MSD100).[1] In the MSD100 database, each song consists of four sound sources (voice, bass, drums and "others") provided as separate tracks.

In this work, microphone array recordings were then simulated using MCRoomSim [130], a room acoustics simulation software. More details about scene generation are given in Appendix. B .

Room dimension	$(10m \times 8m \times 3m)$															
Considered $RT_{60}(s)$	0				0.2				0.4				0.7			
Room configuration	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D

Table 5.1 Information about the used dataset. The room dimension is fixed and similar for all the considered cases. For each considered time reverberation, four room configuration are considered (A,B,C,D), see Fig. 6.2.

[1] https://siSectioninria.fr/sisec-2015/2015-professionally-produced-music-recordings/.

As it is described in Tab. 5.1, a total of 16 simulations were run, corresponding to four rooms and four source configurations. The four rooms had the same dimensions, 10 m × 8 m × 3 m, but different wall absorption coefficients, which resulted in the following reverberation times: 0 s, 0.2 s, 0.4 s and 0.7 s. The four source configurations are illustrated in Fig. 6.2. We chose different reverberation time and different sound source position (from close to each other to far from each other) to study their influences on the performance of the sound source separation.

In every simulation the microphone array was modeled to match the characteristics of the Eigenmike[2] and was located at the same position in the room. In order to calculate the microphone mixtures, for each song and each of the 32 conditions the separate source tracks were then convolved with the simulated impulse responses and summed with each other.

We then built two different inputs for source separation: a microphone mixture $\mathbf{x}$ obtained by the whole 32 microphones of the Eigenmike, and a fourth order ambisonic mixture $\mathbf{b}$ (25 channels) obtained by an encoding of the microphone mixture $\mathbf{x}$.

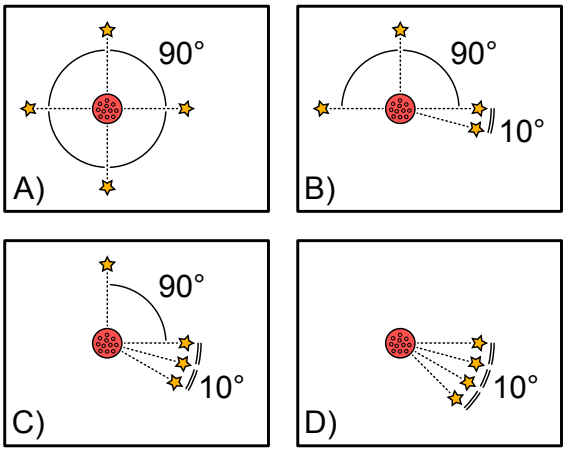


Fig. 5.2 The four sound source configurations considered in our simulations. Note: stars represent sound source locations.

5.4.2 Evaluation criteria

In order to validate the adaptability of FASST in the HOA domain, we propose to compare its performance to the one given by applying FASST in the microphone domain in **highly-informed sound source separation** (the DoAs are consider to be known). A fair comparison requires to compute the chosen performance measures in the same domain. However, given the used (multi-capsules) microphone for our simulations,

[2] `https://mhacoustics.com/products`.

switching back to the microphone domain after separation in the HOA domain is not possible if the used ambisonic mixture is of the fourth order or less (corresponding to at most 25 channels). Indeed, after encoding Eigenmike signals, we introduce two problems in the resulted ambisonic signals. The first one is spatial aliasing, and the second one is related to the loss of information in lower frequencies (more information about this subject are given in Chapter 2 at the end of Section 2.3). To alleviate this issue, instead of computing the evaluation measures in terms of the contribution of each source in each channel/microphone (FASST's outputs), we propose to compute them in terms of sound objects.

In practice we obtain the *sound objects* by applying beamformers on the sources contribution in each channel/microphone, which are the estimated signals by FASST. In this study we use the same type of beamforming for both domains. The used beamforming is known as the matched filter or what we refer to as a basic projection in Chapter 3 Section. 3.4.4.1, which allows to decompose the sound field into plane waves (PWD). Considering the fact that we would like to estimate the sound object j, we form this beam toward the known direction of the direct path source (θ_j, ϕ_j). This is done by projecting each estimated source contribution on the estimation of the steering vector corresponding to its direct path source direction (θ_j, ϕ_j). In other words, the estimated source object j in the microphone and HOA domains are calculated as follows:

$$\hat{\hat{s}}^{\text{Mic}}_{j,f,n} = \frac{\mathbf{a}^H_{j,f}}{||\mathbf{a}_{j,f}||^2}\hat{\mathbf{c}}_{j,f,n} \tag{5.15}$$

$$\hat{\hat{s}}^{\text{HOA}}_{j,f,n} = \frac{\mathbf{y}^T_j}{||\mathbf{y}_j||^2}\hat{\mathbf{b}}_{j,f,n} \tag{5.16}$$

The reference signals are also given by applying the same beamforming but on the true (*i.e* oracle) contribution of each source in each channel/microphone. In other words for the source j the reference signal in the microphone and HOA domains are given by:

$$\hat{s}^{\text{Mic}}_{j,f,n} = \frac{\mathbf{a}^H_{j,f}}{||\mathbf{a}_{j,f}||^2}\mathbf{c}_{j,f,n} \tag{5.17}$$

$$\hat{s}^{\text{HOA}}_{j,f,n} = \frac{\mathbf{y}^T_j}{||\mathbf{y}_j||^2}\mathbf{b}_{j,f,n} \tag{5.18}$$

Source separation performance is assessed by comparing the signals given by Eq. (5.15), and Eq. (5.16) to the ones given by Eq. (5.17) and Eq. (5.18), respectively, with the performance measures proposed in [124], and explained in Chapter 3 Section 3.4.1: Signal to Distortion Ratio (SDR), Signal to Artifact Ratio (SAR), and Signal to In-

terference Ratio (SIR). These measures are then calculated with the BSS-eval toolbox [38].[3]

5.4.3 Evaluated methods

The first evaluated method is FASST applied on the HOA mixtures (see Section. 5.2.) It is compared to its application on the corresponding regular microphone mixtures. The evaluation is done on the same mixtures in both domains, and with the same number of channels.

The first beamformer has already been introduced in Section. 5.4.2. It is the matched filter beamformer(PWD), but this time applied directly to the HOA mixture, which is given by:

$$\bar{\bar{s}}^{\mathrm{HOA}}_{j,f,n} = \frac{\mathbf{y}_j^T}{||\mathbf{y}_{j,f}||^2}\mathbf{z}_{f,n}. \tag{5.19}$$

The second beamformer, which refer to as the pseudo-inverse (PIV) beamformer (see Chapter 3 Section. 3.4.4.1), consists in a plane-wave decomposition with nulls steered toward the directions of interfering sources. It is given by:

$$\bar{\bar{s}}^{\mathrm{HOA}}_{j,f,n} = \mathbf{Y}^{\dagger}\mathbf{z}_{f,n}, \tag{5.20}$$

where the matrix $\mathbf{Y}$ contains the SVs of the sources directions of arrivals.

5.4.4 FASST parametrization and initialization

The FASST toolbox requires choosing configuration parameters, as well as providing initial values for the covariance matrices $\mathbf{\Sigma}_{\mathbf{c}_{j,f,n}}$ in Eq. (6.2). Tab. 5.2 summarizes the parameters used for all experiments. Further, in order to match the scene configuration, the number of sources was fixed to 4 in the anechoic condition and 5 in reverberant conditions, where we observed that it was beneficial to add a source accounting for diffuse noise or late reverberation.

The used algorithm is given in Alg.6. As described in it, the covariances are decomposed into a spectral part and a spatial part, and the spectral part is further modeled by NMF, as proposed by [121, 90].

With FASST, it is possible to model the parameter of each sound source differently. For each of the first four sources, the spatial covariance was initialized as in [87] to the **rank-1** matrices $\mathbf{R}_f^{\mathrm{HOA}} = \mathbf{y}_j\mathbf{y}_j^{\mathsf{H}}$ and $\mathbf{R}_f^{\mathrm{Mic}} = \mathbf{a}_{j,f}\mathbf{a}_{j,f}^{\mathsf{H}}$ for the HOA and microphone

[3]BSS-eval version 3.0 for Matlab, `http://bass-db.gforge.inria.fr/bss_eval/`.

Transform type	STFT
Sampling frequency	44100 Hz
Window length	69 ms (3072 samples)
NMF rank	16
Stopping criterion	150 iterations

Table 5.2 FASST parameters

Algorithm 6 The used algorithm

The number of EM iterations L
for $j = 1$ to $J + 1$ **do**
  Initialize : $\mathbf{W}_{j,fn}$ and $\mathbf{H}_{j,fn}$ ▷ as random matrices
  $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j$
  if $j \leqslant J$ **then**
    Initialize : $\mathbf{R}_{j,f}$ ▷ as $\mathbf{y}_j\mathbf{y}_j^{\mathsf{H}}$
  else
    Initialize : $\mathbf{R}_{j,f}$ ▷ as $I \times I$ identity matrix
  end if
end for
for $l = 1$ to L **do**
  $\mathbf{R}_{b,fn} = \sum_{j=1}^{J+1} v_{j,fn}\mathbf{R}_{j,fn}$
  for $j = 1$ to J **do**
    $\mathbf{W}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\mathbf{R}_{b,fn}^{-1}$
    $\hat{\mathbf{b}}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{b}_{fn}$
    $\hat{\mathbf{R}}_{bj,fn} = \hat{\mathbf{b}}_{j,fn}\hat{\mathbf{b}}_{j,fn}^{H} + (\mathbf{I} - \mathbf{W}_{j,fn})v_{j,fn}\mathbf{R}_{j,fn}$
    $\mathbf{R}_{j,fn} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_{j,fn}}\hat{\mathbf{R}}_{bj,fn}$
    $v_{j,fn} = \frac{1}{I}(\text{trace}(\mathbf{R}_{j,fn}^{-1}\hat{\mathbf{R}}_{bj,fn})$
    $[\mathbf{W}_j, \mathbf{H}_j] = \text{NMF}(\mathbf{V}_j)$
    $\mathbf{V}_j = \mathbf{W}_j\mathbf{H}_j$
  end for
end for

domain, respectively. In the microphone domain, the steering vectors $\mathbf{a}_{j,f}$ were estimated from the microphone array characteristics and **the direct path** sound source positions using Eq. (3.12). In the HOA domain, the steering vectors $\mathbf{y}_j$ were derived as the vector of the first nine spherical harmonic functions evaluated in **the direct path** sound source directions using Eq. (A.15). In the reverberant case, the fifth source was assumed to have a **full-rank** spatial covariance, and therefore, we initialized it with the identity matrix in both domains (ambisonics and microphone). This decision was taken after several tests about the rank of the covariance matrices while monitoring

the log-likelihood. We had the best increase in the likelihood during a certain amount of iterations (around 150) with this decision. Moreover, another motivation behind the initialization of the sound sources spatial covariance matrices to rank-1 is that the comparison is made in terms of the direct path sources. Lastly, regarding the spectral part of the covariance, NMF factors were initialized as random numbers.

5.5 Validation of the approach

As explained before the main goal is to validate experimentally the local Gaussian model assumption for source separation in the HOA domain. To this aim among other simulations we are comparing the performance of FASST in HOA domain and microphone domain.

5.5.1 Selection of the number of microphones/channels

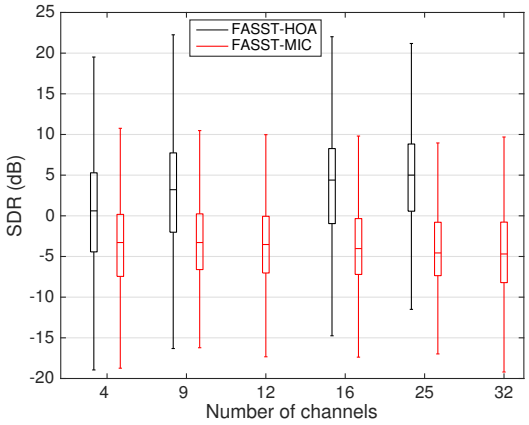
The computational cost of FASST depends primarily on the square of the number of channels of the mixture, and considering the size of our dataset and the number of channels $M = 25$ and microphones $I = 32$, it is important to spare time and resources in the main experiment that will soon be described. A naive approach would be to adopt a lower HOA order $L < 4$, and consider on the one hand HOA mixtures with $M = (L+1)^2$ channels, and on the other one, the same mixtures given by a sub-antenna of the Eigenmike, where the number of the chosen capsules is $I = M = (L+1)^2$.

However, one could argue that while HOA mixtures are obtained by **considering the whole 32 capsules of the Eigenmike**,[4] the microphone mixtures are given by only $M = (L+1)^2$ selected microphones, and therefore, the comparison could be considered unfair. To clarify this point, we begin our experiments by measuring the source separation performance in both domains when varying respectively the number of channels and the number of microphones. This preliminary experiment is done on a small proportion of the created dataset (see below).

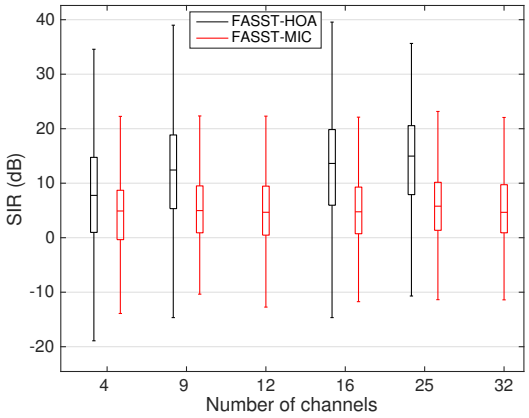
Let us explain the experiment that we conducted in order to choose the fairest number of channels and microphone to compare the performance of the multichannel sound source separation with the LGM approach. First, in the microphone domain we considered different sub antennas from the Eigenmike where the capsules were selected in order to be distributed regularly on the sphere. The considered numbers of microphones are $I = 4, 9, 12, 16, 25, 32$ (the numbers $4, 9, 16, 25$ were chosen to match the number of possible channels in the HOA domain, the number 12 is considered because the chosen capsules can be regularly distributed in the best way to cover the

[4] Whatever is the retained number $M \leqslant 25$ of channels, they are obtained by encoding the signals from the 32 microphones.

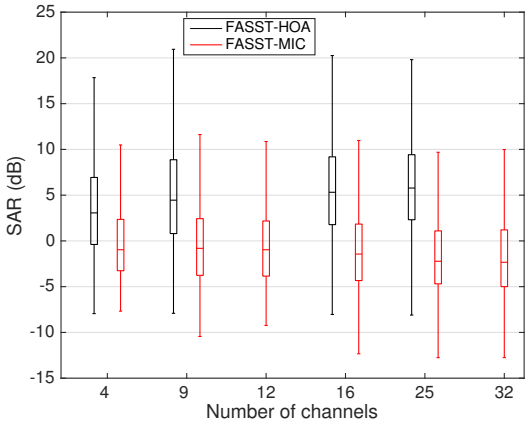
sphere). Second, HOA signals make sense if they are grouped by order L, each order L corresponding to a number of channels $M = (L+1)^2$. We have already at our disposal the 4th order signals (25 channels) by encoding the information provided by the 32 capsules of the Eigenmike. In order to have the first, the second, and the third order we have to simply truncate respectively the 25 HOA signals to the first $M = 4, 9, 16$ channels. Considering the selected capsules in the microphone domain and the truncation of the signals in the HOA domain, from our data set we considered randomly 160 mixtures. All the listed time reverberations and sound source configurations were considered.



(a) SDR



(b) SIR



(c) SAR

Fig. 5.3 Comparing FASST performance in regards of number the used microphones/channels.

We applied FASST to the different mixtures, considering the sources DoA known, the initialization and the parametrization of the toolbox are given in Section 5.4.4. The results in terms of SDR, SIR and SAR are given in Fig. 5.3.

In the microphone domain, we observe that the SIR tends to improve by 0.07 dB in average when increasing the number of microphones, the SAR tends to decrease, when it comes to the SDR we observe that it increases slightly by 0.02 dB in average until 9 microphones and drops after. In the HOA domain, we observe an improvement of all performance measures when increasing the number of channels. We can clearly see that adding more microphones doesn't improve the source separation performance in the microphone domain. As a conclusion it is unnecessary to add more microphones in the microphone domain, and therefore the comparison of FASST's performance between the HOA domain and the microphone domain is fair if the number of channels/microphones is equal to $I = M = 9$. The gap in performances between both domains is due to two main reasons. These reasons are explained in Section 5.5.2, and Section 5.5.3.

5.5.2 Extensive experiments with 9 microphones/channels

In the following the considered number of channels is equal to the considered number of microphones $I = M = 9$. In the microphone domain the selected capsules are given in Table. 5.3. More information about the angular position of the Eigenmike's capsules can be found in Appendix B.

	1	2	3	4	5	6	7	8	9
θ	0	35	-58	-31	0	-58	35	69	-32
ϕ	-32	45	0	90	212	180	135	269	-90

Table 5.3 Elevation (θ) and azimuth (ϕ), in degrees, of the selected Eigenmike microphone capsules. The radius of the microphone is 4 cm. The origin of space is the center of the Eigenmike.

In the following, comparison will be performed at a large scale, considering the whole dataset previously described (Section 5.4.1). The results of the comparison are given in Fig. 5.4. As expected the performance decreases as the reverberation and scene complexity increase, regardless of the signal domain. However, in most configurations, separating the sources in the HOA domain resulted in better performance measures compared to the microphone domain. We can clearly see a gain of 7 to 12 dB for the least challenging sound source configuration, and a gain of 1 to 6 dB for the most challenging one. Tab. 5.4 summarizes the difference in SDR values between the HOA domain and the microphone domain for configurations (A) and (D): the SDR is almost always higher in the HOA domain, regardless of the reverberation or song. As well, the gap between the performance obtained in the two domains reduces as the complexity of the scenario increases, with a more prominent influence of reverberation

time. Separating sources in the HOA domain results in a 1 to 12 dB increase in signal-to-distortion ratio, compared to the microphone domain.

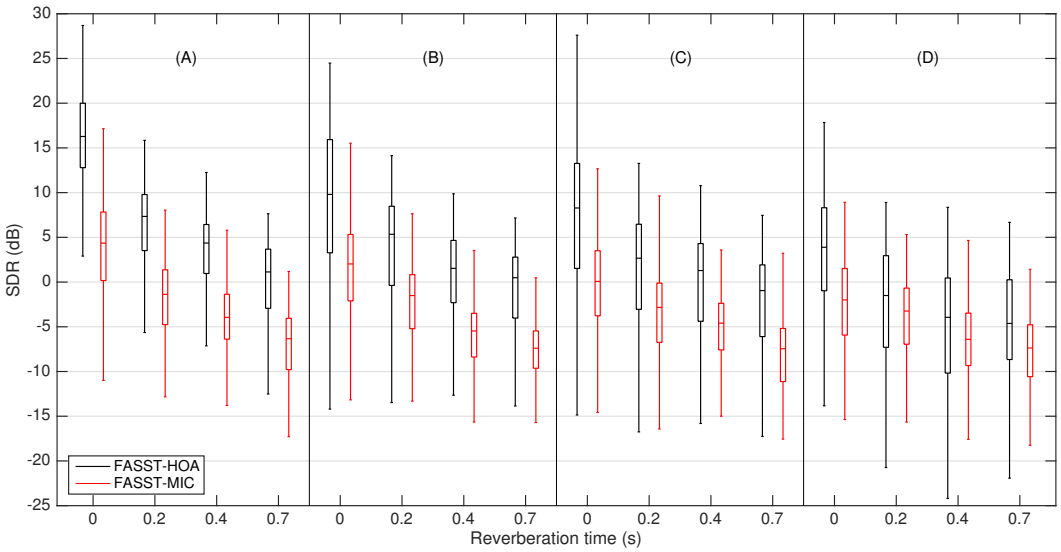


Fig. 5.4 Comparing FASST's performance in the HOA domain to FASST's performance in the microphone domain, $I = M = 9$

	$RT_{60}(s)$	0	0.2	0.4	0.7
	max	21	14.35	11.9	12
A	median	12.43	7.69	7	6.84
	min	4.3	2.52	2.5	2.9
	max	10.17	6.6	6.7	6.32
D	median	6.05	0.83	1.52	2.45
	min	-1.84	-6	-5	-4

Table 5.4 $\Delta SDR = SDR_{HOA} - SDR_{MIC}$, in dB, for scenarios A and D.

One reason may explain these results. Indeed, in FASST's EM algorithm, the empirical covariance matrix is inverted while estimating the first Wiener filter [103] and the numerical stability of this inversion differs in the two signal domains. We calculated the condition number of the empirical covariance matrix in both domains for a random example picked from the dataset. It appeared that, for frequencies below 2 kHz, the condition number was generally higher in the microphone domain than in the HOA domain, and could be about 1000 times greater for some frequency values. Therefore, the conversion of the microphone signals into HOA signals seems to act as a pre-conditioning for the EM algorithm.

Having established the interest of performing the source separation in the HOA domain with FASST, we now compare it with the reference methods. Results are presented in Fig. 5.5. FASST clearly outperforms the reference methods. This is because, contrary to the reference methods which are solely based on spatial cues, FASST also exploits spectral cues. This gives FASST an advantage when sources are close to each other and spatial information is more ambiguous. Although this fact has already been observed in microphone domain source separation [120, 39, 126], we confirm it here also on HOA-domain source separation.

Surprisingly, FASST outperforms the PIV method even in anechoic environment where the PIV method could have been expected to give the best results in terms of performance. Indeed, 9 signals should be enough to form a beam toward one source and cancel 3 interfering sources at the same time. This can be explained with the fact that encoded HOA signals don't match perfectly the theoretical signals. This imperfection is mainly caused by the physical limitations of the microphone array. Indeed, the capsules of the Eigenmike are relatively close to each other, which results in spatial aliasing and a loss of lower frequencies [76].

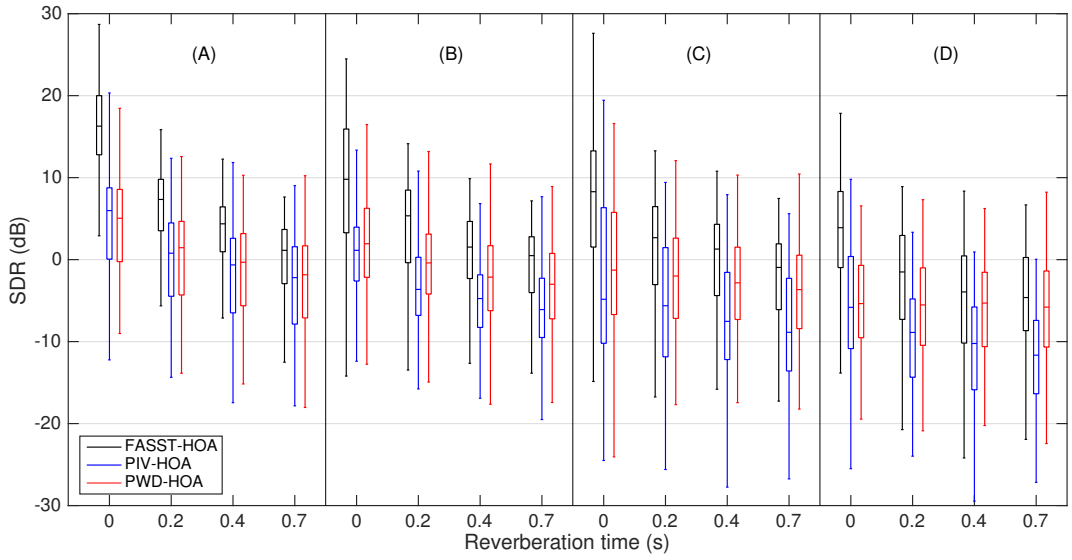
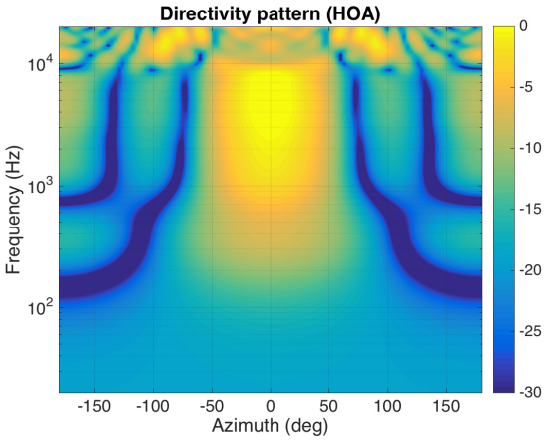


Fig. 5.5 Comparing FASST to the reference methods in the HOA domain.

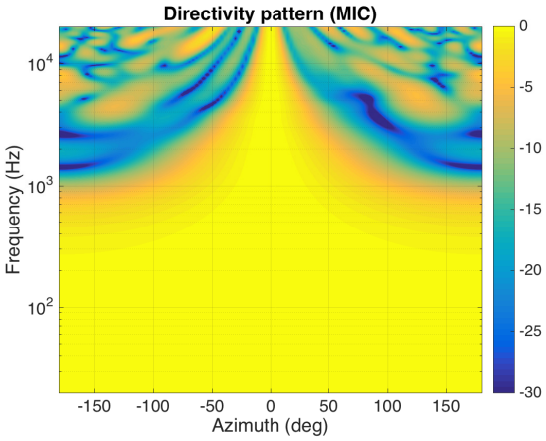
5.5.3 Comparing the PWD directivity patterns

We realized that the comparison between FASST in the microphone domain and FASST in the HOA domain in term of the direct path in Fig. 5.5 may be distorted by the difference of the PWD beamformer behavior in both domains. In order to confirm our hypothesis we analyzed the PWD directivity patterns by designing a beamformer

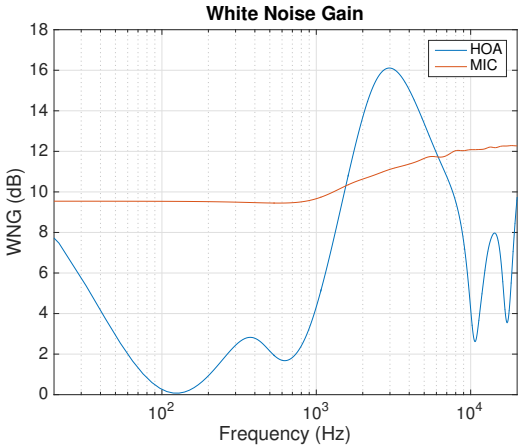
towards the direction ($\theta = 0°, \phi = 0°$), and plotting the directivity patterns in both domains, and white noise gain in regards of the frequency in Fig. 5.6. The directivity of the beamformer is frequency-dependent here because we took into consideration the issues described in Chapter 2 Section 2.3 when the ambisonic signals are acquired from encoding Eigenmike signals.



(a) Directivity pattern HOA



(b) Directivity pattern MIC



(c) White noise gain

Fig. 5.6 Comparing the directivity patterns of the PWD beamformer

When the frequency increases the PWD beamformer gets more selective in the HOA domain compared to the microphone domain. As suspected over all frequencies the PWD beamformer is much more selective in the HOA domain. This remark can be seen in terms of directivities and in terms of white noise gain in Fig. 5.5.

This does not reassesses the fact that the LGM model works in the ambisonics domain. However, it query the results in Fig. 5.4 regarding the microphone domain, which explain the poor results. There is actually no other ways to compare objectively

the performance of the sound source separation.

5.6 Conclusion

In this chapter we investigated for the first time the ability of the local Gaussian model to handle the source separation problem in the HOA domain. To this aim we have established the model's equations in the HOA domain and run numerical experiments. In **informed case** (DoA of the sound sources considered known), our simulation results show that applying a local Gaussian model-based source separation method in the HOA domain typically results in the SDR increasing by 1 to 12 dB, compared to the microphone domain with the same number of microphones/channels $I = M = 9$, including in challenging situations such as reverberant environments and complex source configurations. Although the comparison was skewed by the selectivity of the beamformer in both domains, we validated the model in the HOA domain. In the next chapter, we will explore using proximity microphones in order to guide the source separation and improve its performance, and finally employ this method to allow navigation through HOA sound scenes.

Two examples of sound sources separation with the used approach in this chapter are given in `https://hafsatimohammed.github.io/HTML_Files/Example1.html` and in `https://hafsatimohammed.github.io/HTML_Files/Example2.html`. For these examples, we simulated ambisonic recording with four sound sources in a reverberant environment ($RT_{60} = 0.35s$). We used FASST with the described parametrization in Section 5.4.4. The listening examples for this chapter are referred to as FASST in the web pages. In these web pages, we give the mixture, the first channel of the true contributions (for a comparison purpose), and the estimated contributions' first channel. We also give the first channel of estimated contributions with other algorithms discussed later in this manuscript in Chapter 6 and Chapter 7.

Chapter 6

Multichannel decomposition of ambisonic sound fields informed by spot microphones

In this chapter, we propose a workflow for a multichannel source separation on HOA mixtures, where J spot microphones provide side information with known position, one for each source in the sound field , which has been recorded live at the same time as the primary antenna. We propose to process the information of each spot microphone to estimate the power-spectral-density (PSD) of each source. These PSDs are use to initialize the expectation-maximization (EM) algorithm which estimates the multichannel Wiener filter coefficients (see Chapter 5 Section 5.2). In this chapter we investigate on the performance of the approach regarding different circumstances such as the type of sound sources, the reverberation time, the order of the ambisonic signals and the position of the sound sources. We compare the performance of the proposed workflow to the performance of the local gaussian model approach with the NMF constraint as it is in FASST. The comparison is investigated in terms of resources, time consumption and performance of the sound source separation.

6.1 Reminder on the multichannel sound source separation under the LGM assumption

As explained previously in Chapter 4 Section 4.1.2, the source separation problem in the ambisonic domain consists in estimating the contribution $\mathbf{b}_{j,t} \in \mathbb{R}^M$ of each source $j = 1, ..., J$ in each channel $m = 1, ..., M$ and at each time instant $t = 1, ..., T$. The contribution of each source in each channel in the time-frequency domain is given in Eq. (3.19).

We validated in Chapter 5 the LGM assumption in the ambisonic domain. Thereby, we model the source contributions $\mathbf{b}_{j,f,n}$ as independent of each other and following a complex-valued-zero-mean Gaussian distribution:

$$\mathbf{b}_{j,f,n} \sim \mathcal{N}_c(0, v_{j,f,n}\mathbf{R}_{\mathbf{b}_{j,f}}), \tag{6.1}$$

where $v_{j,f,n} \in \mathbb{R}_+$, $\mathbf{R}_{\mathbf{b}_{j,f}} \in \mathbb{C}^{M\times M}$ represents respectively the power spectral density and the spatial covariance matrix of the j^{th} source. Eq. (6.1) allows us to express the Wiener filter given in Eq.(4.17) as follows:

$$\mathbf{W}_{j,f,n} = v_{j,f,n}\mathbf{R}_{\mathbf{b}_{j,f}} \left(\sum_{j'=1}^{J} v_{j',f,n}\mathbf{R}_{\mathbf{b}_{j',f}} \right)^{-1}. \tag{6.2}$$

The sound source j ambisonic signals are then recovered by applying element-wise the Wiener filter on the ambisonic mixture as follows:

$$\hat{\mathbf{b}}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{z}_{fn}. \tag{6.3}$$

An EM algorithm can be used to estimate the Wiener filter coefficients. There are several variants of this algorithm. One of them is known as the full rank unconstrained model. It is given in Alg. 7. This algorithm is used in our workflow. We can say that with our workflow the NMF constraint is replaced with the side information provided by the spot microphones. We set the covariance matrix to be full rank by initialize it with the identity matrix (more details are given in Section 6.2).

We suppose that we have at our disposal an ambisonic mixture and J spot microphone signals. The position of the primary antenna and each spot microphone may not be known. We suppose that each spot microphone was close to a sound source and that they were recording live at the same time as the primary antenna. The main idea is to use the information provided by the spot microphones to guide the search of the Wiener filter coefficients.

As explained in Chapter 5, the Wiener filter coefficients are found by maximizing the log-likelihood, which happens to be non-convex function [90]. A usual way to maximize this function is using an EM algorithm. There are several variants of this algorithm. The most known one is the so-called full rank unconstrained model [36].

Algorithm 7 EM updates for full rank unconstrained model [36]

1: The number of EM iterations L
2: **for** $j = 1$ to J **do**
3: Initialize : $v_{j,fn}$
4: Initialize : $\mathbf{R}_{j,fn}$ ▷ as $I \times I$ identity matrix
5: **end for**
6: **for** $l = 1$ to L **do**
7: $\mathbf{R}_{b,fn} = \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$
8: **for** $j = 1$ to J **do**
9: $\mathbf{W}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\mathbf{R}_{b,fn}^{-1}$
10: $\hat{\mathbf{b}}_{j,fn} = \mathbf{W}_{j,fn}\mathbf{b}_{fn}$
11: $\hat{\mathbf{R}}_{bj,fn} = \hat{\mathbf{b}}_{j,fn}\hat{\mathbf{b}}_{j,fn}^{H} + (\mathbf{I} - \mathbf{W}_{j,fn})v_{j,fn}\mathbf{R}_{j,fn}$
12: $\mathbf{R}_{j,fn} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_{j,fn}}\hat{\mathbf{R}}_{bj,fn}$
13: $v_{j,fn} = \frac{1}{I}(\text{trace}(\mathbf{R}_{j,fn}^{-1}\hat{\mathbf{R}}_{bj,fn})$
14: **end for**
15: **end for**

6.2 Source separation informed by spot microphones

6.2.1 Layout

As the name of the algorithm presumes, there is no constraint on the sound source spectra $v_{j,fn}$, and the spatial covariance matrix $\mathbf{R}_{\mathbf{b}_{j,f}}$ is initialized and defined as a full rank matrix. A constraint on the sound source spectra can be added to guide the estimation of the Wiener filter parameters. For instance, As described in Chapter 5, in FASST, there is an option to set a constraint on the sound source spectra as the multiplication of two non-negative matrices. We use the full rank unconstrained model because we want to set our constraint with the help of the information provided by the spot microphones. Thus, the algorithm would be guided to find the suited parameters of the Wiener filter quickly.[1] In our workflow, we propose to estimate earlier the power spectral densities (PSD) of the sound sources using the spot microphones and give them as an initialization of the $v_{j,fn}$, line three of Alg. 7. Since our PSDs are strictly initialized with the direct path spectra, we wanted to have covariance matrices that are able to model the echoes effectively while being still flexible to each example (short or long reverberation time). Therefore, the full-rank model seems to be appropriate.

[1] with less iterations.

To this aim, the spatial covariance matrix initialization and updates will be kept the same as in Alg. 7.

We propose to preprocess the spot microphone signals for a better estimation of the sources PSDs. First, we apply an interference reduction in the spot microphone signals. Second, in case the position of the spot microphones and the main antenna are unknown, The next step is to estimate the so-called propagation parameters. Third, we apply these parameters to align in magnitude and phase the spot microphone signals with the ambisonic signals.[2] Finally, we compute the PSDs and use those to initialize Alg. 7. Usually the alignment step is included in the identification of the spatial covariance matrices, and It is quite unusual in the literature to estimate the propagation parameters and consider the sound sources spectra at the position of the microphone array. The aim of this study is to check if this can help the Wiener filter to have a better estimate of the contributions quickly (with less iterations).

Assuming that the PSDs are quite well estimated, it is possible to fix them and update only the spatial covariance matrices, by removing line 13 from Alg. 7.

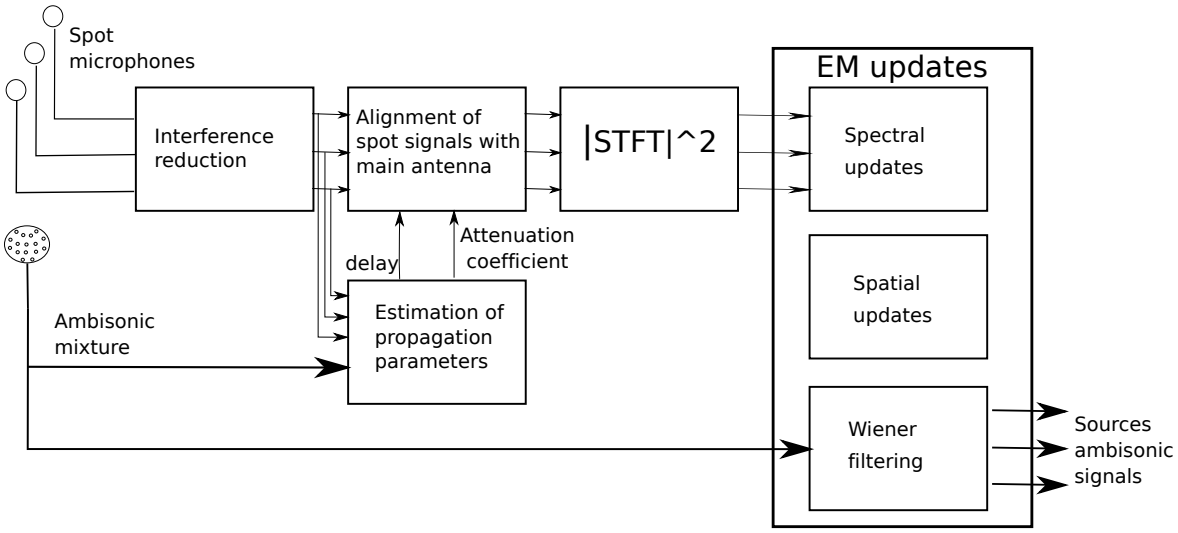


Fig. 6.1 The proposed workflow

Based on the multichannel Wiener filter approach with the local Gaussian model assumption, we propose the workflow presented in Fig 6.1.

6.2.2 Interference reduction

The main problem of the side information provided by the spot microphones are interferences. Indeed in multi-tracks live recording, each spot microphone records its

[2]We know that it is unusual to consider the PSDs at the position of the microphone array. However, we want to evaluate if such an approach can help the EM algorithm to better estimate **quickly** the spatial covariance matrices, and therefore the contributions of the sound sources.

dedicated sound source plus the contribution of the other sources. Therefore, to estimate the PSD of each source, interference reduction can be helpfull. Over the last decade, several algorithms were suggested by the literature [117, 61, 26]. We have based our workflow on the Gaussian framework proposed in [33]. This framework is similar to the multichannel sound source separation with the LGM approach. Indeed the contribution of each source in each spot microphone is assumed to follow a centered Gaussian distribution. Thus the Wiener filter mask to be applied on each spot microphone signal parameters is simplified. For each spot microphone, the Wiener filter mask depends on two main parameters, being the PSD $v'_{j,fn}$ of the signal in the spot microphone and the interference matrix $\Lambda_f = [\lambda_{ij}]$, the coefficient λ_{ij} quantifies the amount of interference of the source j in the spot microphone i (remind that in our case the number of spot microphone is equal to the number of sources $I = J$). The Wiener filter to be applied to each spot microphone signal is presented as follows:

$$\mathcal{W}_{j,fn} = \frac{v'_{j,fn}}{\sum_{k=1}^{J} v'_{k,fn}\lambda_{jk}} \tag{6.4}$$

The Wiener filter parameters can be estimated using an EM algorithm. The used algorithm is presented in [33]. To understand the process, One should consider all the distributed spot microphones as a whole antenna. The goal of the algorithm is to reduce the interferences in each channel of the antenna. In our case, each channel has a dominant sound source because the microphone in question is close to it. Considering all the channels, we are looking for the interference matrix $\Lambda_f = [\lambda_{ij}]$ that helps to minimize the mean square error at each channel. Note that the PSDs of each source in each channel $v_{i,j,f,n}$ are the same up to channel-dependent scaling (that quantify the amount of interference) factors $\lambda_{i,j,f}$:

$$v'_{i,j,f,n} = \lambda_{i,j,f} v'_{j,f,n} \tag{6.5}$$

The used EM iterative algorithm alternates between separation (reduction of interference) and re-estimation of the parameters. The re-estimation of the parameters is done by maximizing the likelihood. More information about the process are given in [33]. The signals after the reduction of interference $s'_{j,fn}$ are estimated by applying the Wiener gain given in Eq. (6.4) to the spot microphone signals $s_{j,fn}$:

$$s'_{j,fn} = \mathcal{W}_{j,fn} s_{j,fn}. \tag{6.6}$$

6.2.3 Propagation parameters

Instead of considering the sound source spectra at their position as it is usually viewed by state of the art for the EM algorithm, we want to check if it is possible to help the EM algorithm by initializing $v_{j,f,n}$ at the position of the microphone array. We want to study if it does give a better estimation of the spatial covariance matrices for a given amount of iterations, and therefore a better estimation of the contributions quickly. Usually, the propagation aspect is handled by the spatial covariance matrices. To this aim, we will fix the number of iteration and activate and deactivate this block to study the differences if there is any.

Indeed, the propagation of the sound source from the spot microphone to the main antenna involves a delay and an attenuation. We want to incorporate this aspect in the initialization of the sound sources spectra $v_{j,f,n}$. Therefore, the spot microphone signals after interference reduction, $s'_{j,t}$, must be delayed by a delay δ_j and their magnitude must be multiplied with a gain γ_j:

$$\hat{s}_{j,t} = \gamma_j s'_{j,t-\delta_j}, \tag{6.7}$$

Note that both parameters depend on the distance between the microphone j and the main antenna. Considering this distance d_j, the parameters are deduced as follows:

$$\begin{cases} \delta_j = \frac{d_j}{c} \\ \gamma_j = \frac{1}{d_j}, \end{cases} \tag{6.8}$$

where c is the speed of sound.

In the case where neither the position of the spot microphone or the main antenna are known, we propose to estimate first the delay δ_j, and deduce the distance and thereby γ_j. One way to estimate the delay δ_j is to: First, to estimate the DoA of the sound sources using one of the algorithms validated in Chapter 4. Section 4.2 (for instance DEMIX). Second, to apply a basic plane wave decomposition to estimate the j^{th} sound source at the position of the ambisonic antenna:

$$\bar{s}_{j,f,n} = \frac{\mathbf{y}_j^\top}{||\mathbf{y}_{j,f}||^2}\mathbf{z}_{f,n}. \tag{6.9}$$

where $\mathbf{y}_j$ is the spherical harmonic vector corresponding to the j^{th} sound source DoA. And finally, to compute the cross-correlation between the spot microphone signal after

interference reduction, $s'_{j,t}$, and the signal estimated in Eq. (6.9). Theoretically this coross-correlation is given by:

$$\Gamma_{s'\bar{s}}(\tau) = \mathbb{E}[s'(t)\bar{s}(t-z)], \tag{6.10}$$

practically we can compute it with a sample-based estimate. An estimation of the delay $\hat{\delta}_j$ is given by:

$$\hat{\delta}_j = \text{argmax}_\tau(\Gamma_{s'\bar{s}}(\tau)). \tag{6.11}$$

6.2.4 Wiener filtering

After prepocessing the spot microphone signal, the PSDs are initialized as the following estimation:

$$\hat{v}_{j,fn} = |\hat{s}_{j,fn}|^2. \tag{6.12}$$

The spatial covariance matrix of each source is initialized as the identity matrix. Both parameters are going to be updated by Alg. 7 up to a fixed number of iteration. The number of iterations will be chosen after a first experiment. Note that we would like to have less iterations with a performance that converges while monitoring if such an usual initialization of the sound source spectra helps to get a better performance. The contribution of each source in each channel $\mathbf{b}_{j,fn}$ is given by applying the last estimated Wiener filter in line 10 of Alg. 7, and finally, we use an inverse STFT with overlap-add to reconstruct the time-domain signals. It is possible to fix the sources PSDs by removing line 13 of Alg.7, which makes the EM algorithm to be involved with only spatial updates.

6.3 Experimental evaluation

In this section we aim to evaluate the workflow proposed in this chapter through the following experiments:

- We investigate the influence of the spectral updates on the performance of the sound source separation if the sound source spectra are well estimated (the true spectra) and fixed.
- We investigate the influence of each block on the performance of the sound source separation.
- We investigate the influence of the sound source distribution in the sound scene.

- We investigate the influence of the ambisonic order on the performance of the sound source separation.
- We investigate the influence of the type of sound sources.
- We compare it to the used approach in Chapter 5 (LGM with an NMF constraint using FASST).

6.3.1 Dataset

To evaluate the performance of the proposed workflow, we have built four different datasets. For all of them, we used MCRoomSim [130] to simulate the spot microphone recordings. We fixed both the dimension of the room to 10 m × 8 m × 3 m and the position of the ambisonic microphone array.

Note that the sound sources had the same elevation and were at a distance of 2.5 m from the main antenna for the first three datasets.

For all the datasets, we positioned each spot microphone at a distance of 0.5 m from the corresponding sound source and modeled it to be cardioid. We pointed its directivity toward the direction of the corresponding sound source.

For the first data set, we also simulated the configuration of 2 speakers. We fixed the wall absorption coefficients so that the reverberation time would be equal to $0.3s$ or $0.7s$. We considered two different angle difference, with respect to an origin corresponding to the Eigenmike location, between the two speakers, being 5° or 180°. We created the Eigenmike RIRs corresponding to each configuration using MCRoomSim. We randomly chose 10 seconds of 2×20 different sound signals from the SiSEC campaign data set [122]. We created the Eigenmike mixtures and encoded them to get the ambisonic mixtures. In the end, we got $2 \times 2 \times 20 = 80$ fourth-order ambisonic mixtures for the first dataset.

For the second data set, we simulated the configuration of 2 speakers. We fixed the wall absorption coefficients so that the reverberation time would be equal to $0.3s$. Different angle difference between the two sound sources were considered: 11 angles between 5° and 25° (2° between each) and 7 angles between 25° and 180° (25° between each). The configuration of the second data set is illustrated in Fig. 6.2 (A). We randomly chose 10 seconds of 2×10 different sound signals from the SiSEC campaign data set [122]. Similarly to previous datasets, we created the Eigenmike mixtures and encoded them to get the ambisonic mixtures. In the end, we got $10 \times (11 + 7) \times 2 = 180$ fourth-order ambisonic mixtures for the second dataset.

For the third data set, we simulated a musical content containing four sound sources (voice, bass, drums, and others). The room dimension and the position of the main

antenna are similar to the first data set. However, we considered different wall absorption coefficients, which resulted in the following reverberation times: 0 s, 0.2 s, 0.4 s, 0.8 s, 1 s, 1.2 s. We chose two sound scene configurations. Both of them are illustrated in Fig. 6.2 (B, C). We randomly chose 10 seconds of 4×10 different signals (voice, drums, bass, and others from the same song) from the DSD100.[3] In the end, we got $10 \times 2 \times 6 = 120$ fourth-order ambisonic mixtures for the third dataset. For more information about our simulated datasets please refer to Appendix B.

For the fourth dataset, we considered a room with a fixed position for the spherical microphone array and a fixed time reverberation of $RT_{60} = 0.4s$. Thirty room impulse responses corresponding to 4 sound sources with random and different positions were generated. These room impulse responses were convolved with sound sources from the DSD100. The mixtures were truncated to the first four channels. We downsampled the sampling frequency to $16KHz$. The outcome mixtures correspond to 10 seconds of the first order ambisonic sound fields.

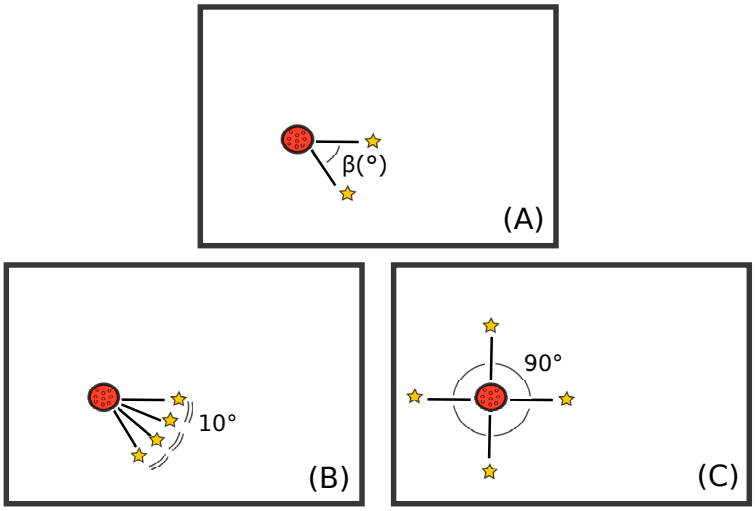


Fig. 6.2 The sound source configurations considered in our simulations. Note: stars represent sound source locations.

6.3.2 Evaluation criteria

In order to evaluate the performance of the multi-channel source separation we use the standard energy ratios [124] Chapter 3 : Signal-to-Distortion Ratio (SDR), Signal-to-Artifact Ratio (SAR), and Signal-to-Interference Ratio (SIR), computed with the BSS-eval toolbox [38].[4]

[3] https://sigsep.github.io/datasets/dsd100.html .
[4] `http://bass-db.gforge.inria.fr/bss_eval/`, version 3.0 for Matlab.

6.3.3 Experiments and results

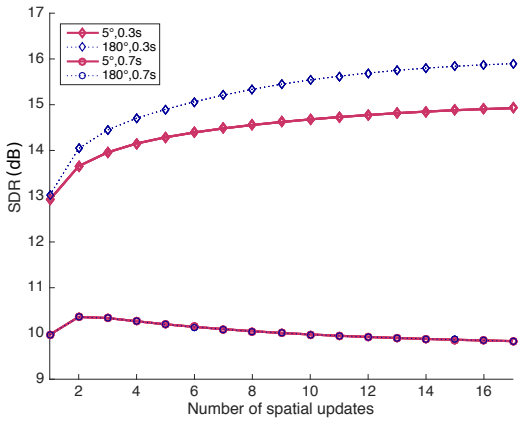
We considered five different experiments to evaluate the performance of the workflow in terms of several aspects.

6.3.3.1 Assessing the impact of the spatial updates

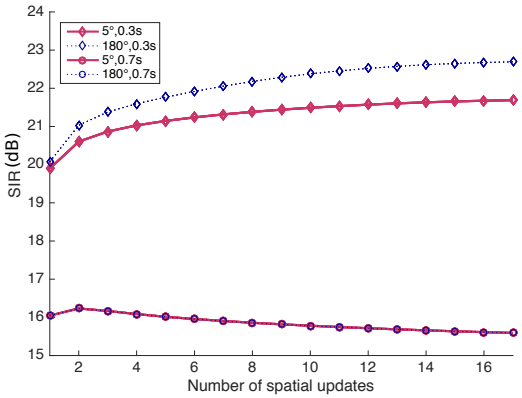
For the first experiment, we wanted to study the convergence of the EM algorithm while updating the spatial covariance matrices and fixing the sources PSDs with the ground truth values. For this, we used our first dataset. We truncated the ambisonic sound fields to the second order, which makes them have nine channels. The convergence is investigated with the chosen evaluation criteria Section 6.3.2. For each mixture, we averaged the scores over both sound sources and the 20 examples. The results are presented in Fig. 6.3. As it is explained in the legend, the color, and the line style differentiate the angle difference between both singers (bold red line for 5°, and blue dashed line for 180°) and markers for reverberation time (diamonds for $RT_{60} = 0.3s$, and circles for $RT_{60} = 0.7s$). For the rest of experiments the number of iterations of the EM algorithm was fixed to 10. Fig 6.3 shows the performance of the multichannel source separation over the number of spatial updates while fixing $v_{j,fn}$ with the ground truth magnitude sources spectra. Overall we can say that an excellent estimation of the spectral part of the Wiener filter $v_{j,fn}$ would help to get good SDR scores. In this case, we got an SDR score over $10dB$ and an SIR over $15dB$, which we consider as a great multichannel sound source separation. This is expected since the ground truth spectra PSD are used. We can see that when the time reverberation gets bigger, we need fewer spatial updates to achieve the best SDR score possible and thereby the best multichannel sound source separation. We expected this behavior since we initialize the spatial covariance matrix with the identity matrix. Note that a diagonal spatial covariance matrix describes a diffuse sound field. If ever the required number of iteration is exceeded, it seems that the performance of the sound separation decreases. However, it is not drastic; it seems to converge about $0.5dB$ below the best score. When it comes to the position of the sound sources in the sound field, if the sound source spectra are well estimated, the performances are roughly the same or within $1dB$ from each other.

These results help us to confirm that estimating the sound spectra a priori and giving them as initialization to the EM algorithm can guide the sound source separation. At this point, we just need a way to have an accurate estimation of the sound spectra. The workflow presented in Fig. 6.1 has this objective and must be characterized, which is precisely the goal of the second experiment.

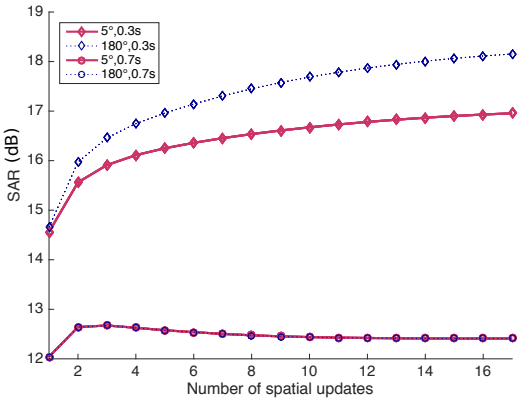
These results allow us to fix the number of iterations to 10, which seems to us



(a) SDR



(b) SIR



(c) SAR

Fig. 6.3 Performance comparison over spatial updates while fixing the true sources PSD. The scores are in dB

as a decent compromise between long and short reverberation time. For the rest of experiments, the number of iterations is fixed to 10.

6.3.3.2 Assessing the impact of each block

For the second experiment, the goal was to diagnose the impact of each block of the workflow. For this aim, we considered the second data set. We truncated the order of the ambisonic mixtures to the second-order, which resulted in considering the first 9 channels. For this experiment, we considered the position of both the spot microphones and the ambisonic antenna to be known. We investigated three processing blocks:

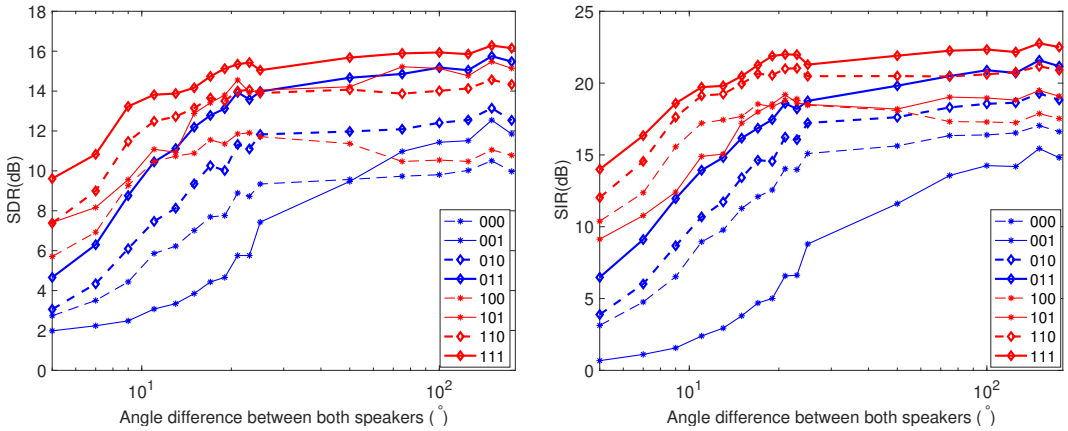
- The interference reduction Eq. (6.4). The parameters of the filter in Eq. (6.4) were estimated using the algorithm in [33]
- The application of the propagation parameters Eq. (6.7). They were considered known.
- The re-estimation of the sources PSD line 13 of Alg. 7.

The evaluation of the impact of each block was measured by judging the performance of the sound source separation. To this aim, we used the evaluation criteria discussed in Chapter 3 Section 3.4.1. Similarly to the last experiment, the scores were averaged over the sound sources and the examples. For each block, we considered two cases: turned on '1', or turned off '0'. In total, we had eight cases. We present the results in Fig. 6.4. We used binary code to simplify the description of the legend 'xyz' in which x presents the state of the first block (Interference reduction), y presents the state of the second block (alignment of the sound sources spectra), and finally, z presents the state of the third block (spectral updates along spatial updates using Alg. 7). Note that Fig. 6.4 the state of each block is specified by a plotting style:

- Two different colors for the state of the block (Interference reduction).
- Two different markers for the state of the second block (Propagation parameters).
- Two different line types for the state of the third block (PSD re-estimation).

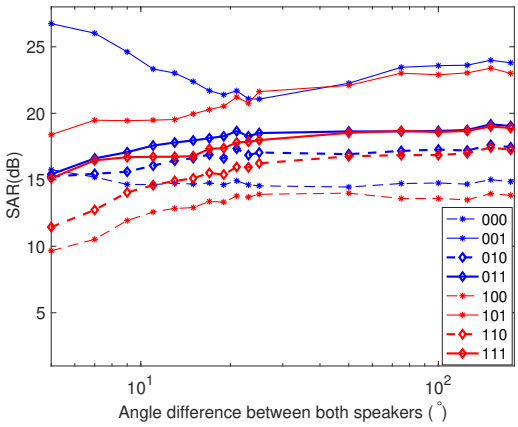
Fig 6.4 shows the results of the second experiment (the evaluation of the impact of each block on the source separation performance regarding the angle difference between both sources). We can clearly see that when the sources are close to each other (angle difference under 10°), the red lines give the best scores, which correspond to an active interference reduction. We can see also that an application of the propagation parameters helps to increase the scores (lines with diamond markers). The best score

over all the chosen angle difference range is given by activating both blocks as well as re-estimating the PSDs in the EM algorithm.



(a) SDR

(b) SIR



(c) SAR

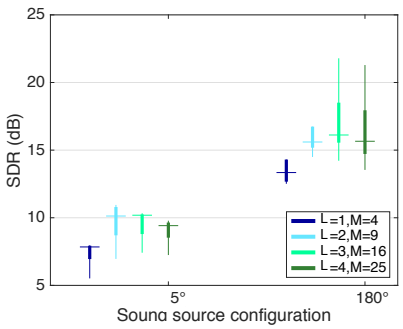
Fig. 6.4 Comparison of the workflow's performance according to each block over the distance between the sources

6.3.3.3 Assessing the impact of ambisonic order

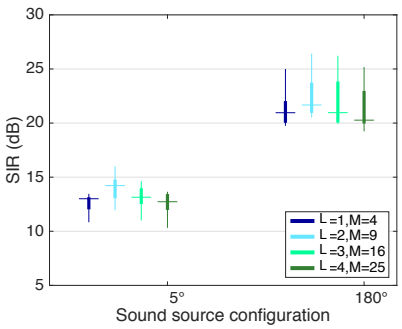
For the third experiment, we wanted to study the performance of the workflow on different ambisonic orders. We used the first data set. We truncated the mixtures to each order (first, second, and third) and applied the workflow with all the blocks being active. We did the same steps on full 25 channels of the fourth-order. We averaged the scores over both sources and present the results in the form of box plot in Fig 6.5.

When it comes to the impact of the order of the ambisonic sound field on the workflow, we observe in Fig 6.5 that when the order gets more significant (from the first-order $m = 1, M = 4$ until the third order $m = 3, M = 16$), the SDR score increases and slightly decreases after the third-order. This behavior was different from the last

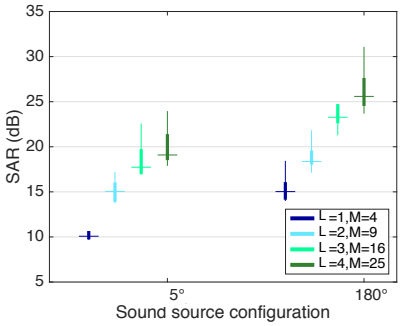
chapter where an LGM approach with a NMF constraint was used. In the last chapter in Fig. 5.3 the SDR increases while the number of channels gets more significant. In terms of SAR, we observe an important improvement when the number of channels is more significant. In terms of SIR, we observe a slight decrease when the number of channels gets larger than $M > 9$. It seems that with the interference reduction block, we get close to the sound source spectra, and adding more than 9 channels does not help to improve the performance of the sound source separation. We observe that the SAR increases when the number of channels gets larger.



(a) SDR



(b) SIR



(c) SAR

Fig. 6.5 Comparison of the performance according to the ambisonic order over the type of the sound scene configuration. L and M denotes the order of the ambisonic signals and the number of channels, respectively.

6.3.3.4 Assessing the performance of the approach on complex sound scenes

For the fourth experiment, we investigate the impact of the environment and the type of sound sources by considering complex sound scenes with much longer reverberation times and different types of sound sources, including instruments such as drums, bass, guitars, etc. Therefore, we used the third dataset. In this data set, we had 4 sound sources. Each sound source represents a different type of instrument, singing voice, drums, bass, others (piano, or violin, guitar, etc). We used the same evaluation cri-

teria as the last three experiments. We observed different behavior for each type of instrument regarding the activation of the blocks. First, we report the impact of the blocks over the configuration of the environment in Fig. 6.6. For this figure, the scores were averaged over all the sound sources. Second, we report the SDR for each sound source. In Fig. 6.7 for the voice. Fig. 6.8 for the bass. Fig. 6.10 for the drums. Fig. 6.9 for others.

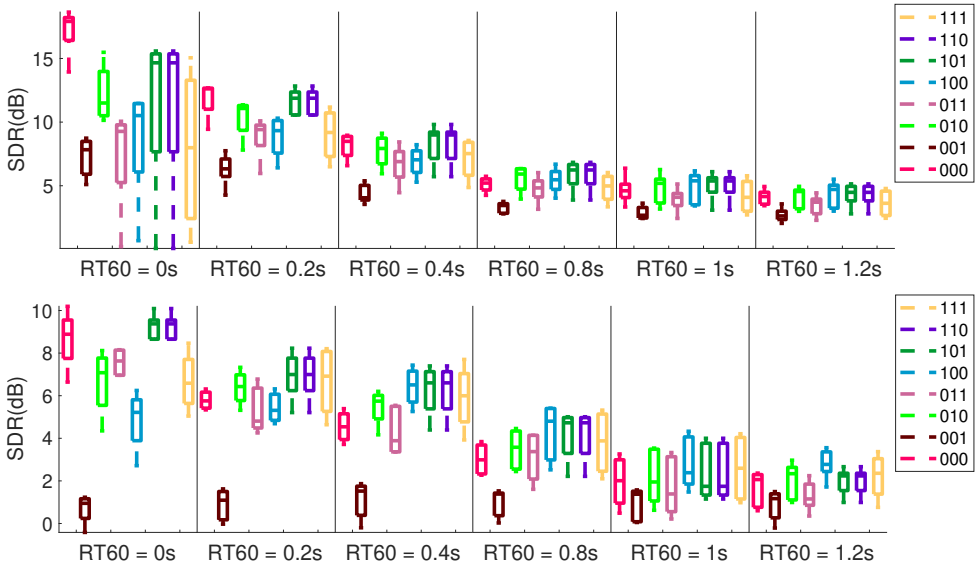


Fig. 6.6 SDR scores for complex sound fields. The top corresponds to configuration B and the bottom to configuration C.

The performance of the workflow considering complex sound scenes is presented in Figs.6.6 to 6.9. it seems that the behavior of the workflow is not similar to when the sound fields contain only speech sound sources. It is not always the best strategy to activate all the blocks. Overall, the best strategy seems to activate the first and the last blocks or the first and the second blocks. However, on the one hand, once the reverberation time gets more significant, the best strategy seems to be activating all the blocks for the voice which is not surprising as we already have the same conclusion in the second experiment, and the drums. On the other hand, it is not the case for the bass and others where the best strategy seems to be fixing the PSDs after applying the first two blocks.

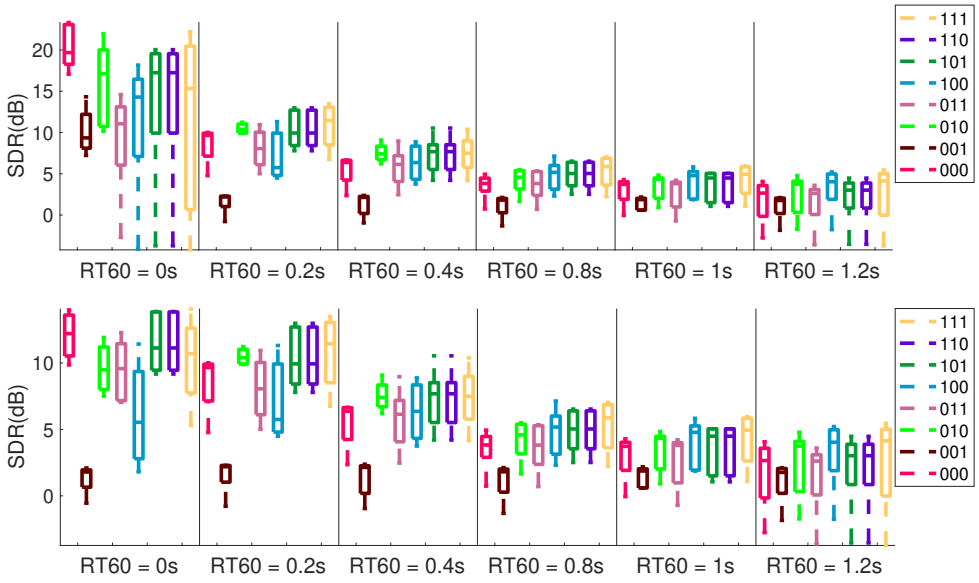


Fig. 6.7 SDR scores for the voice. The top corresponds to configuration B and the bottom to configuration C.

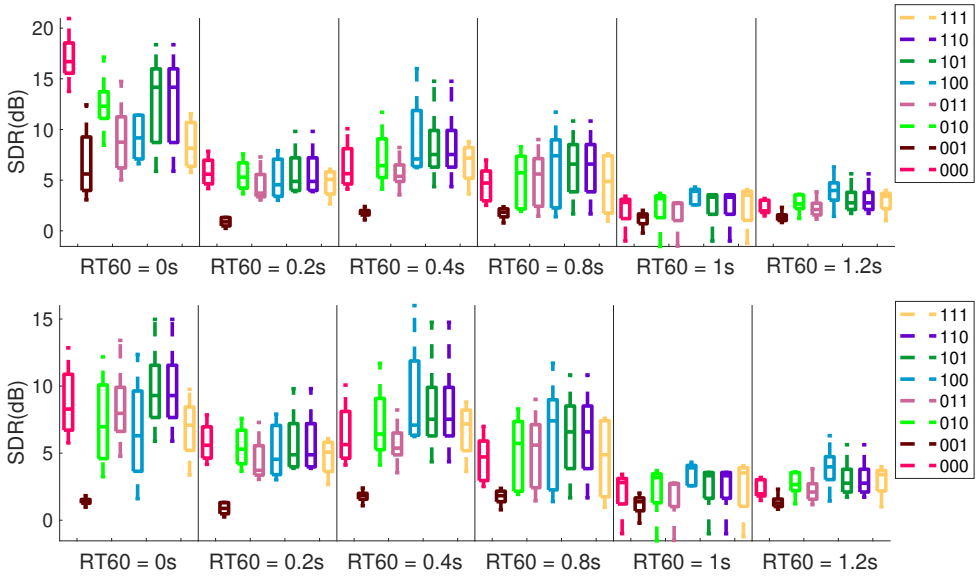


Fig. 6.8 SDR scores for the bass. The top corresponds to configuration B and the bottom to configuration C.

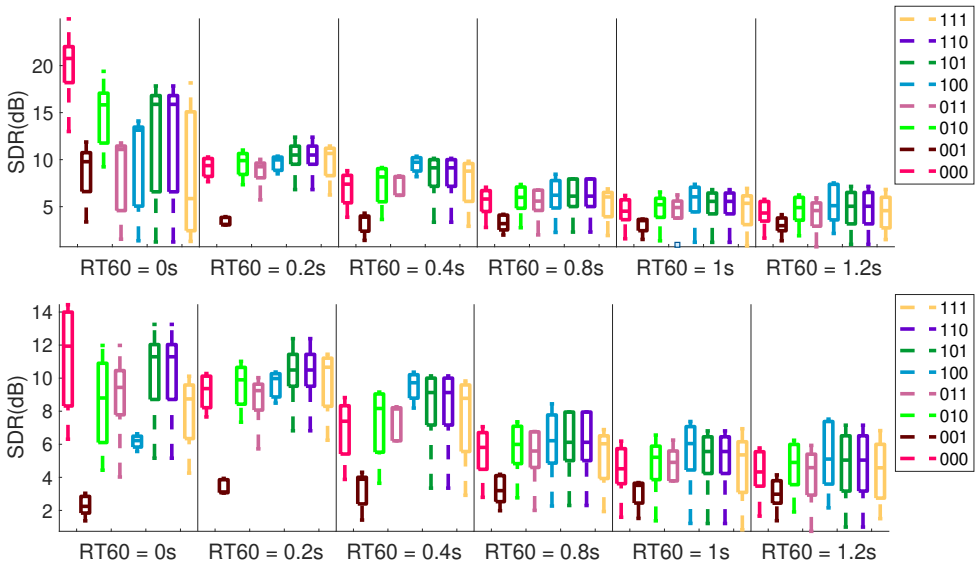


Fig. 6.9 SDR score for others. The top corresponds to configuration B and the bottom to configuration C.

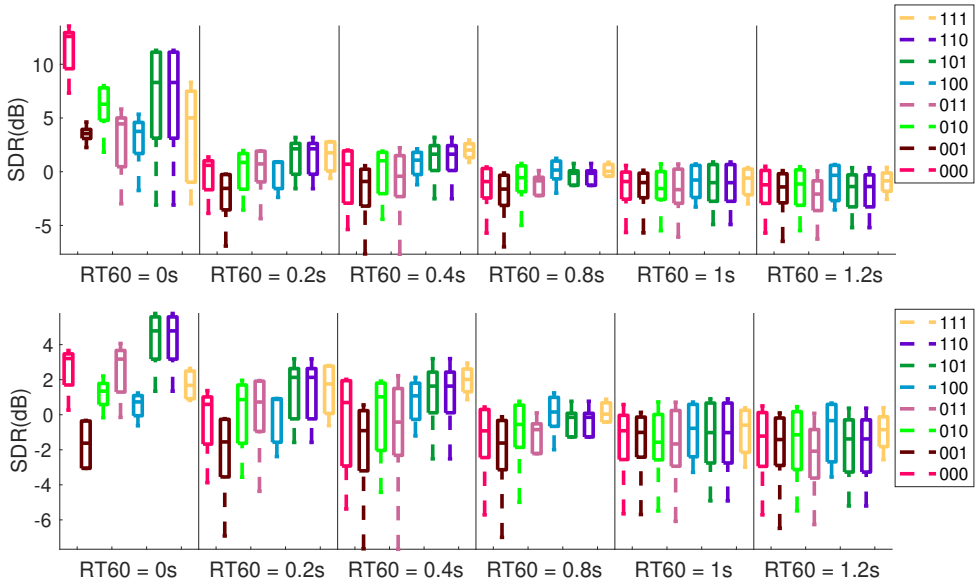


Fig. 6.10 SDR scores for the drums. The top corresponds to configuration B and the bottom to configuration C.

For the drums, the scores are lower than that obtained for the other instruments. An informal listening test revealed that for the drums some attacks were missing. It seems that the application of the Wiener filter smoothed the sound source separation to the point that some attacks of the drums went missing. Note that with the Wiener filter the summation of all the contributions gives the exact mixture. The missing drums

attacks were found in the estimation of others. Similar behavior will be remarked in the next chapter.

6.3.3.5 Comparing the performance of the approach to a an LGM approach with an NMF constraint

Since we fixed the number of iterations for our approach to 10 iterations, it is clear that our approach will be faster if we compare it to the approach in Chapter 5 (LGM approach with an NMF constraint) with 150 iterations. To this aim, we conducted a first experiment to showcase the fairness of our comparison in terms of time consumption between our approach and the one in Chapter 5. For the first experiment, we compared the performance of FASST with 10 iterations to FASST with 150 iterations. We present the results in Fig. 6.11. As it is presented, FASST requires more than 10 iterations for better performance. We chose 150 iterations because it seems to be the number of iterations where the maximum likelihood begins to converge.

Therefore, for the last experiment, we fixed the number of iterations to 150 for FASST and 10 for the EM algorithm for our approach. Both algorithms were run on the same machine. For our approach, considering the results in Section. 6.3.3.1 and Section. 6.3.3.2 ,we chose to activate all the blocks. In Fig. 6.12, we present the comparison in terms of SDR between FASST and our approach, which is referred to it as SpotMic. The Δ_{SDR} is given in Table 6.1. The processing time is given in Table 6.2.[5]

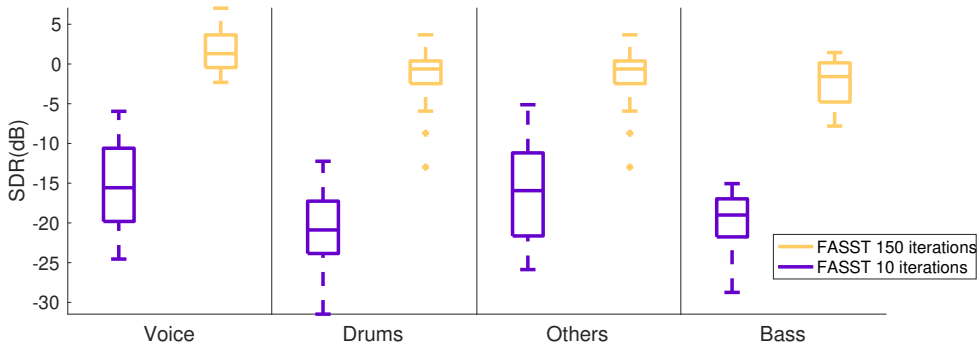


Fig. 6.11 FASST 10 iterations Vs 150 iterations.

[5]MacBook pro with 2,2 GHz Intel Core i7 processor and 16Go of Ram

	FASST	SpotMic
Number of iterations	150	10
Sampling frequency	16 kHz	16kHz
Duration of mixtures	10s	10s
Number of channels	4	4
Number of sound sources	4	4
Time of processing	775.84s	56.42s

Table 6.2 FASST and our approach time of processing.

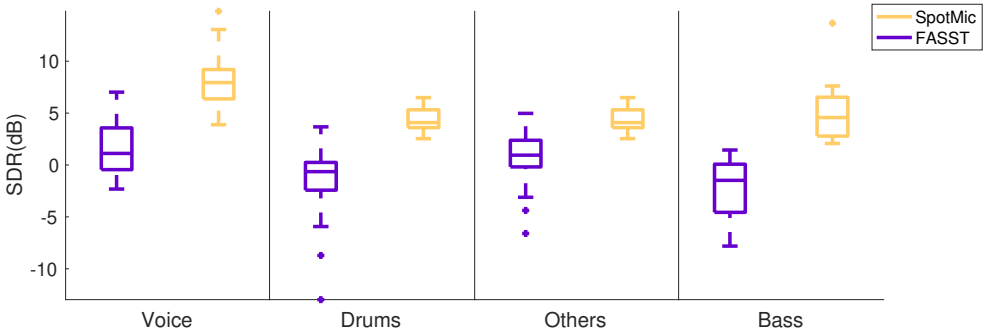


Fig. 6.12 Comparison between FASST and the proposed workflow.

	Instrument	max	median	mean	min
$\Delta_{SDR}(dB)$	Voice	7.79	6.82	6.54	6.20
	Drums	2.81	4.72	6.06	15.5
	Others	6.14	4.67	5.62	9.31
	Bass	12.22	6.05	7.29	9.88

Table 6.1 Δ_{SDR} between FASST and our approach.

We can see that with our workflow, the performance of the sound source separation is significantly better than the approach in Chapter 5 (LGM approach with an NMF constraint), according to Fig.6.12). We observe a gain of at least 4.5dB according to Table 6.1 with fifteen times less iteration for the EM algorithm. Along with the improvement of the sound source separation performance, we gain time as well according to the time measurement reported in Table 6.2. With our approach we can be thirteen times quicker than the approach in Chapter 5 (LGM approach with an NMF constraint) with a gain of 4.5dB at least in terms of SDR. The main disadvantage with our approach is that we need to have a spot microphone for each sound source along with

the spherical microphone array. However, spot microphones are not very expensive, and therefore we recommend to use them for our application.

6.4 Conclusion

In this chapter, we proposed a workflow based on using side information provided from spot microphones to guide the EM algorithm and improve the estimation of the Wiener filter parameters. First, we checked the proper functioning of the idea. Second, we evaluated the impact of each block on speech separation. Third, we evaluated the performance of the separation for each ambisonic order. Fourth, we checked the performance of the workflow on complex ambisonic sound fields. Finally, we compared our workflow to the LGM approach with the NMF constraint approach. We can conclude that:

- In the case of speech, our approach seems to give an SDR of at least 10dB when all the blocks are active even when the sound sources are close to each other in the presence of moderate reverberation (in the case of our experiment we got an SDR of 10dB with an angle difference of 5°).[6] Therefore our unusual way of incorporating the propagation parameters in the sound sources PSDs as an initialization can be beneficial to the sound source separation, which was the case for speech at least.

- The number of channels influences the performance of the sound source separation. On the one hand, it seems that the performance of the sound source separation gets better when the number of channels increases until the third order, where the performance seems to be the best. On the other hand, the SAR continue to increase while increasing the number of channels.

- The reverberation time influences the performances as well. The larger the reverberation time, the lower the scores. However, the SDR scores on average are larger than 0dB, for every sound source configuration and reverberation time. We observed that the activation of the block depends on the reverberation time. For anechoic and shallow reverberated environments, we recommend deactivating all the blocks, whatever is the type of sound source in the sound field. For typical reverberated environments (such as conference rooms) or highly reverberated environments, we recommend activating all the blocks if the sound field contains speech only, and the two first blocks if the sound field contains music along speech signal.

[6] By moderate reverberation or common environments we mean rooms with a reverberation time between $0.3s$ and $0.4s$.

- The type of sound sources influences performance. It seems that it works better for speech and worst for instruments such as drums. However, it gets great results compared to the used approach in Chapter 5 (LGM approach with an NMF constraint and rank-1 model). With our approach, we can gain at least 4.5dB in terms of SDR compared to the approach in Chapter 5 (LGM approach with an NMF constraint) with fewer iterations. With the parameters that we considered, we got significantly better performances thirteen quicker.

Two examples of sound sources separation with the approach proposed in this chapter are given in **https://hafsatimohammed.github.io/HTML_Files/Example1.html** and **https://hafsatimohammed.github.io/HTML_Files/Example2.html**. For these examples, we simulated ambisonic recording with four sound sources in a reverberated environment $RT_{60} = 0.35s$ with four-spot microphones that were close to the sound sources. We used the approach proposed in this chapter. All the blocks were active. The position of the microphones was considered known, and the number of iterations was fixed to 10. In these web pages, we give the mixture first channel, the true contribution in the first channel (for a comparison purpose), and the estimated contribution in the first channel. Along with this listening example, you can find the one from Chapter 5 and compare the performance of the sound source separation. An example of navigation with the approach proposed in this chapter is given in **https://hafsatimohammed.github.io/HTML_Files/Example_Navigation.html**. For this experience, we simulated three sound sources in a reverberant environment that were recorded live with an ambisonic antenna and three spot microphones. Each one was close to a given sound source. Using this chapter's approach, we applied this chapter's approach to perform the multichannel sound source separation to decompose the ambisonic sound scene. To each output, we applied the matching pursuit algorithm in Alg. 5 to decompose it into plane waves (as described in Chapter 4, Section 4.1.2). All of this was done before launching the demo. During the demo in real-time, we reconstruct the ambisonic sound field regarding the user's position, which is manipulated using the keyboard.

Chapter 7

Multichannel music separation using neural networks in the ambisonic domain

In this chapter, we propose to replace the spot microphones and estimate the sound source spectra by neural networks. As a proof of concept we chose to restrict our study to the separation of music signals. Therefore, we worked on musical ambisonic mixtures that contain 3 instruments (drums, bass, others) and a singer (male or female). We refer to the singer in this chapter as an instrument as well.

In the case of sound fields containing only speech, there are some neural network solutions for speech enhancement in the ambisonic domain, such as [94] that can be integrated into our approach. In this chapter we assess the proposed strategies and neural network architectures. We compare them to each other and investigate their efficiency. At the end we compare all the proposed approaches in the previous chapters (Chapter 5 and Chapter 6) to each other.

7.1 The proposed approach

We consider that our ambisonic mixtures contain a singer female or male and three instruments being the drums, the bass, and others. The term "others" refer to another type of instrument that is different from bass and drums. Our approach aims to recover the contribution of each instrument, including the voice in each ambisonic channel.

In the literature, several articles propose to use neural networks to perform a sound source separation for single-channel mixture such as in [44, 96, 106]. The goal of the neural networks in sound source separation problems in general is to recover the spectra of the sound sources from the spectrum of the mixture. In other articles such as in

[80], the authors recommend to recover the masks (see Chapter 4 Section 4.3.1.1) and apply them to the mixture. With this technique, the training is smoother because the estimated output is always between zero and one, instead of being in an extensive range of values.

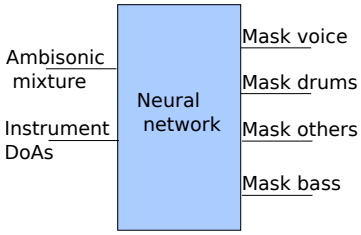


Fig. 7.1 Inputs and outputs of the used neural networks.

With the neural networks, our goal is quite similar to that found in the literature. We aim to recover the instrument spectra. We chose to recover the masks that we can apply to the first channel of the mixture to get the spectra. In contrary to most previous works [44, 96, 106], we have several channels that we can exploit to guide the neural network in finding the spectra. In [94], the author proposed a neural network approach for ambisonic speech enhancement. She proposed to use beamforming towards the sound sources as an input along with the first channel of the ambisonic mixture. We considered similar approach for our inputs. In our case, we used the first channel of the ambisonic mixture along with the beamforming toward the instruments. We discuss the features in more detail later in Section 7.1.2. As an output of the neural network, we considered the masks of the instruments. Fig. 7.1 illustrates, in general, the input and the output of our neural network.

We then use the masks to compute the instrument spectra, which are used to estimate the Wiener filter coefficients (covariance matrices) as it is shown in Fig. 7.2.

Once the masks are estimated, we consider three approaches to estimate the covariance matrices in order to compute the Wiener filter coefficient:

- Estimate the covariance matrices from a rough first estimation of the contribution of each source in each channel. It will be refer to as “Approach 1”. Note that this approach can be considered as an adaptation of the approach presented for speech in [94] for music separation.

- Estimate the covariance matrices with the EM algorithm in Alg. 7, while fixing the spectra with the estimated one from the neural networks. It will be refer to as “Approach 2”.

- Estimate the covariance with the EM algorithm in Alg. 7, but this time the spectra are initialized with the estimated one from the neural networks and not

fixed. It will be refer to as "Approach 3".

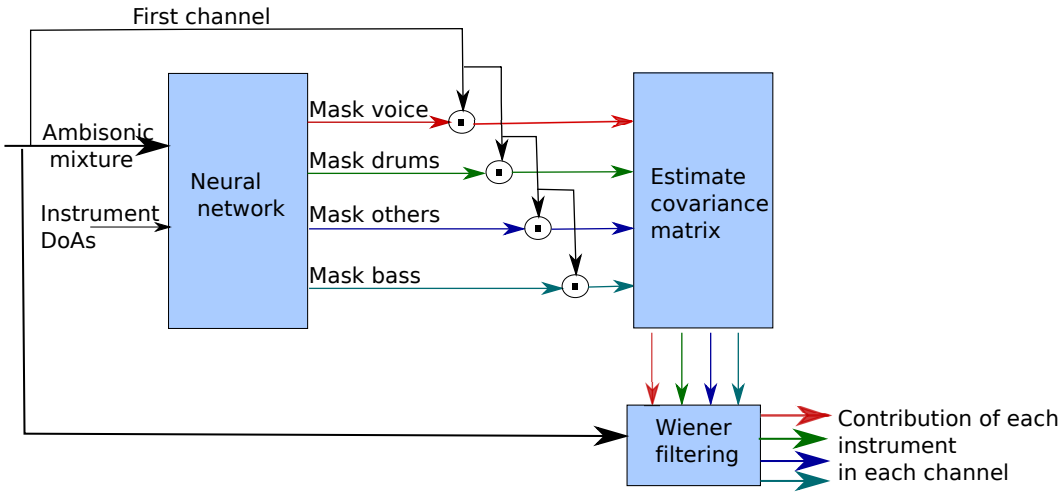


Fig. 7.2 Overview of multichannel music separation with neural networks

We illustrate "Approach 1" in Fig. 7.3a. The idea is to estimate the covariance matrix from a first estimation of the instrument contributions, which are given by applying element-wise the estimated masks to the mixture channels:

$$\bar{\mathbf{s}}_{j,f,n} = m_{j,f,n}\mathbf{z}_{f,n}. \tag{7.1}$$

The covariance matrix of each instrument are then given by:

$$\mathbf{\Sigma}_{j,f} = \frac{1}{N}\sum_{n=1}^{N} \bar{\mathbf{s}}_{j,f,n}\bar{\mathbf{s}}_{j,f,n}^{H}. \tag{7.2}$$

A second estimation of the contributions is given by applying a Wiener filter to the mixture, the coefficients of which are computed as follows: :

$$\mathbf{w}_{j,f} = (\mathbf{\Sigma}_{j,f} + \mathbf{\Sigma}_{j',f})^{-1}\mathbf{\Sigma}_{j,f}, \tag{7.3}$$

with $\mathbf{\Sigma}_{j',f}$ being the covariance of the noise, which is computed as in Eq. (7.2), but with an estimated noise signal. The noise signal can be estimated using one of the following equations:

$$\bar{\mathbf{s}}_{j',f,n} = \sum_{j'=1,j'\neq j}^{J} m_{j',f,n}\mathbf{z}_{f,n} \tag{7.4}$$

$$\bar{\mathbf{s}}_{j',f,n} = (1 - m_{j,f,n})\mathbf{z}_{f,n}, \tag{7.5}$$

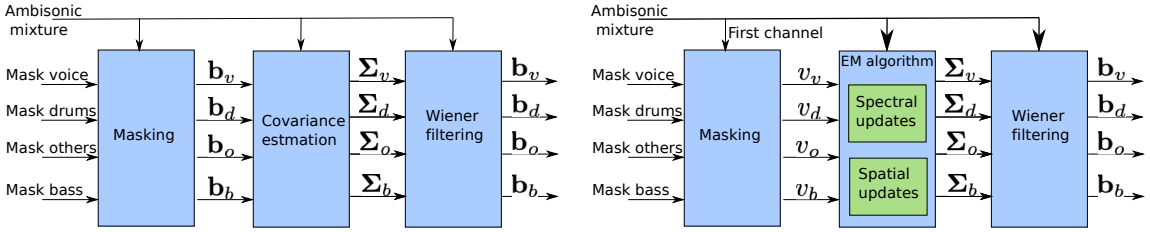
by the noise signal, we mean interference signals (other instruments), diffuse noise, *etc.*

In reality, Eq. (7.4) would consider the interferences as the rest of the instruments and will not consider the other problems such as the diffuse noise. Eq. (7.5) estimates the noise accurately as it considers the entire mixture besides the instrument that we are estimating. Therefore, we recommend using Eq. (7.5) to estimate the noise.

For "Approach 2" and "Approach 3", we estimate the spectra as the square module of the mask applied to the first channel of the mixture:

$$v_{j,f,n} = |m_{j,f,n}\mathbf{z}_{w,j,f,n}|^2. \tag{7.6}$$

As in Chapter 6, we initialized the EM algorithm with these spectra. We used Alg. 7. For "Approach 2", we fixed the spectra and only performed spatial updates. For "Approach 3", we performed both spectral and spatial updates. We set the number of iterations to 10 for both "Approach 2" and "Approach 3" approaches.



(a) "Approach 1" (b) "Approach 2" and "Approach 3"

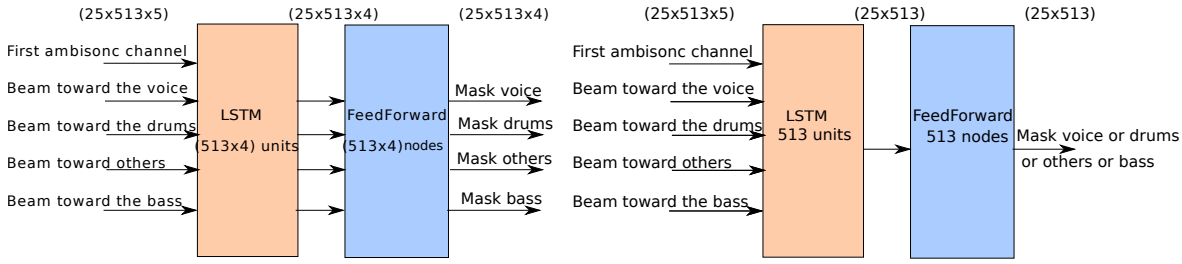
Fig. 7.3 The different approaches. For the second approach we do not consider the spectral updates.

7.1.1 Architectures

There are several types of neural networks. Since we are treating sound data that evolves with time, we chose to use recurrent neural network as an architecture. Especially, Long Short-Term Memory LSTM, in which the output of the previous input is saved and used for the prediction of the current output. For our neural network, the first layer is going to be an LSTM, followed by a Feedforward neural network with a Sigmoid as an activation function, so that the output is between zero and one. We considered two approaches:

- Estimate the spectra of all instruments at the same time with one architecture, referred to it as OneForAll (Fig. 7.4a).

- Estimate the spectra of each instrument with its dedicated architecture, referred to it as EachForOne (Fig. 7.4b).



(a) OneForAll (b) EachForOne. Four times one for each instrument

Fig. 7.4 The chosen architectures

7.1.2 Features and training parameters

All the mixtures were resampled at 16 kHz. The STFT of all signals were computed using a window of 1024 samples, with 50% overlap. The recovery of the temporal signals was done with an overlap-add process using a similar window.

As inputs, we considered 25 consecutive frames from the mixture's first channel spectrum and from the beamforming towards the instrument spectra. As an output, we considered the masks of each instrument corresponding to the 25 selected frames. For the training, we computed OSMs using Eq. (3.66) (for more information about OSMs, please refer to Chapter 3 Section 3.4.2). Note that for the training and validation data, the true sound source signals were known. Fig. 7.5 illustrates how features are extracted.

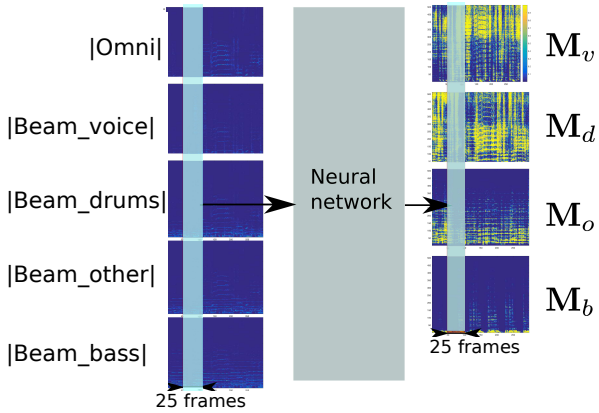


Fig. 7.5 Feature extraction. The features are highlighted in blue, which correspond to 25 temporal frames.

The first input is the magnitude of the mixture first channel $|\mathbf{z}_w|$. The four other inputs are the magnitude of the sound sources estimation $|\hat{\mathbf{s}}_j|$, with a matched filter on

the direction of each instrument. They are given by:

$$\hat{s}_{j,f,n} = \frac{\mathbf{y}_j^H}{||\mathbf{y}_{j,f}||^2}\mathbf{z}_{f,n}. \tag{7.7}$$

The outputs are in the form of OSMs, which are computed from the real sound sources signals at the ambisonic microphone position s_j as follows:

$$m_{j,f,n} = \frac{|s_{j,f,n}|^2}{\sum_{j=1}^{J} |s_{j,f,n}|^2}. \tag{7.8}$$

Once the order of both inputs and outputs is chosen, it must be respected for both the training and the test stages.

For the first neural network (AllInOne), the LSTM layer contains 513×4 hidden units, and the Feedforward layer contains 513×4 nodes. Every 513 output of the output layer are the mask of each instrument spectrum. For the second neural network (OneForEach), the LSTM layer contains 513 hidden units, and the Feedforward layer contains 513 nodes. The 513 nodes of the output layer are the mask of a given instrument spectrum. Given several article [79, 4, 54, 94], we choose the following training parameters for both neural networks:

- We activated the Feedforward layer with a Sigmoid function, which allows us to have numbers in the output nodes between zero and one, since the OSMs are between zero and one.
- We used the mean square error as cost function, with an L2 regularization of 10^{-5}.
- We used Nadam [34] as a type of optimization for the gradient descent.
- We initialized the learning rate to 10^{-4}.
- We used a dropout of 50% on the first layer weights.
- We fixed the number of Epochs to 100.
- We used an early stopping mechanism, which happens if ever the validation error is not decreasing during 10 epochs.

7.2 Experimental protocol

The objectives of the experiments are to investigate the performance of each architecture as well as the three different ways to estimate the Wiener filter. In the end,

Data dedicated for	Number of RIR	RT_{60}	Number of created mixtures
Training	4×50	0.25s	2500 of 10s. They were created by picking randomly from the 4×50 created RIRs, and the 50×4 sound source signals dedicated to the training. Note that, in order to have 2500 different mixture each time, we randomly mixed instruments from different songs.
Validation	4×25	0.3s	50 of 10s. They were created by picking randomly from the 4×50 created RIRs, and the 25×4 sound source signals dedicated to the validation. Note that, in order to have 50 different mixtures each time, we randomly mixed instruments from different songs.
Testing	4×25	0.35s	25 of 10s. created by picking randomly from the 25×4 created RIRs and the 25×4 sound source signals dedicated to the test.

Table 7.1 Training, validation and test datasets.

we compare all the approaches studied or proposed as a solution to the sound source separation problem.

7.2.1 Training, validation, and test datasets

We created the training, the validation, and the test datasets using sound source signals from the DSD100 dataset. Note that this dataset consists from 100 stereo songs. Each one of the songs contains a singing voice, drums, a bass, and another instrument, which their signals are provided as well. We used these sound source signals to create our ambisonic mixtures. We divided these songs into 50×4 sound source signals to use for our training dataset, 25×4 sound source signals for our validation dataset, and the rest 25×4 sound source signals for our test data set.

More information about the datasets are given in Table. 7.1. For each RIRs, the position of the sources was randomly chosen and ensured to be different each time with an angular distances of at least $\theta = 5^\circ$ and $\phi = 5^\circ$.

7.2.2 Results and discussion

After training both neural networks, we ran both of them on the test dataset, which resulted in estimating the masks of each instrument in each mixture. The estimated

masks are used to compute the instrument spectra, and to compute the instrument contribution in each ambisonic channels using whether "Approach 1" or "Approach 2" or "Approach 3". For the "Approach 2" or "Approach 3", we chose to run the EM algorithm with 10 iterations due to the conclusion of the investigation in Chapter 6 Section 6.3.3.1.[1]

7.2.3 Comparison of all the neural network approaches

We present the results in plot box style in terms of SDR, SIR and SAR in Fig. 7.6, Fig. 7.8, and Fig. 7.7, respectively. Comparing both neural networks, it seems that training a neural network for each instrument gives the best results. This result was not expected. Technically the OneForAll neural network has more information about all the sound sources during the training process since it has the OSM of each instrument compared to EachforOne. However, we can explain this result with the fact that the OneForAll neural network may have been a bit shallow for the amount of constraint given as outputs. The results might have been different if more layers have been added.

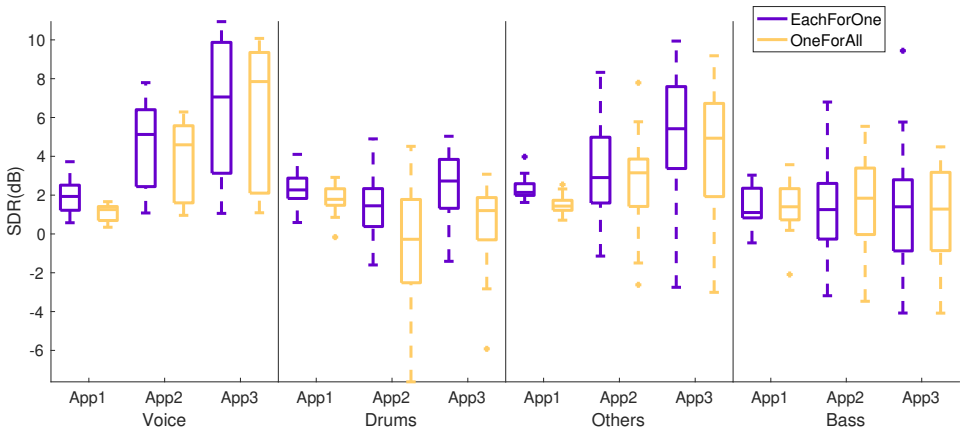


Fig. 7.6 Comparison in terms of SDR of all the approaches using one of the proposed architectures.

For "Approach 1", the SDR is always above zero for each instrument. It seems that the drums and others gave the best scores. For "Approach 2" and "Approach 3", the SDR scores are mostly above zero, but this time the singing voice and others are the best one estimated.

[1]Indeed, in this section we concluded on the fact that when the sound source spectra are well estimated we need around 10 iterations of the EM algorithm for the spatial updates in order to have the best separation performance in terms of SDR. We assume that with the neural networks we are able to get a great estimation of the source spectra.

	Instrument	max	median	mean	min
$\Delta_{SDR}(dB)$	Voice	8.41	5.21	4.90	0.74
	Drums	0.162	-0.46	-1.08	-5.76
	Others	6.63	3.34	2.90	-3.71
	Bass	0.95	0.09	-0.50	-1.90

Table 7.2 Δ_{SDR} between "Approach 3" and "Approach 1" using EachForOne architecture.

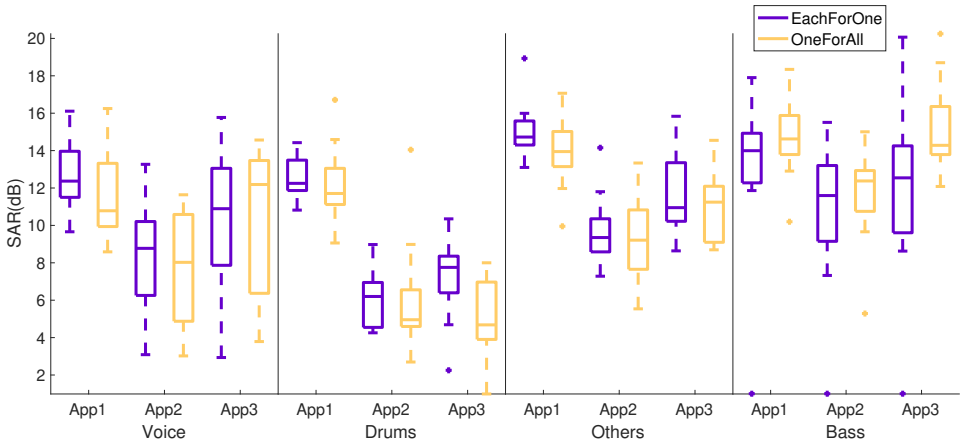


Fig. 7.7 Comparison in terms of SAR of all the approaches using one of the proposed architectures.

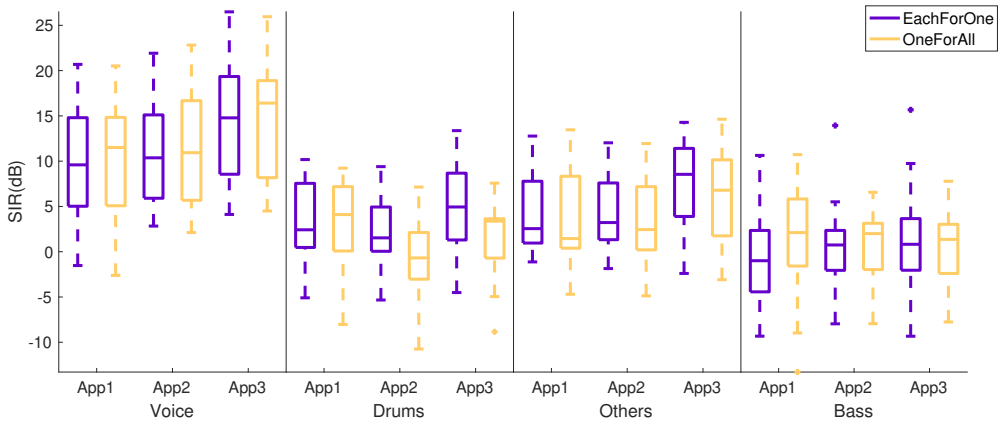


Fig. 7.8 Comparison in terms of SIR of all the approaches using one of the proposed architectures.

Using an EM algorithm in "Approach 2" and "Approach 3", unlike in "Approach 1", increases the SDR and the SIR scores significantly for the singing voice and others. However, it does estimates the drums and the bass poorly compared to "Approach 1". The difference of SDR between "Approach 1" and "Approach 3" for the EachForOne neural network is given in Table 7.2. For the drums, we already had similar behavior in the last chapter as in "Approach 2" and "Approach 3" .

7.2.4 Comparison of neural network approaches to the previous studied approaches in Chapter 5 and Chapter 6

To compare these approaches to the one proposed in Chapter 6 and the one we studied in Chapter 5, we used the same dataset created for the fifth experiment in Chapter 6. Note that this dataset is very different from that used in the training and the validation data.[2] For comparison, we run the data on "Approach 3", and both neural architectures (OneForAll and EachForOne). We fixed the number of iteration to 10 similarly to the method in Chapter 6 (The approach with the spot microphones). For the approaches in Chapter 5 (A systematic use of the LGM approach referred to as "FASST") and Chapter 6 (LGM approach guided with the spot microphones referred to as "SpotMic"), we used the same parameters as in the fifth experiment in Chapter 6. As a reminder we chose 150 iterations for "FASST" because we observed that it is the needed amount of iteration for the convergence of the likelihood for a systematic use of the LGM approach.

We compared the performance of the sound source separation in terms of the SDR and time consumption. We presents the results in Fig. 7.9 and Table 7.5 . The Δ_{SDR} between the spot microphones approach and the neural network approach, and the one between FASST and the neural network approach are given in Table 7.3 and Table 7.4, respectively.

	Instrument	max	median	mean	min
$\Delta_{SDR}(dB)$	Voice	-0.16	1.16	0.83	2.38
	Drums	0.02	0.48	1.21	8.15
	Others	-0.28	0.61	0.4	0.11
	Bass	5.28	2.77	3.63	6.21

Table 7.3 Δ_{SDR} between FASST and $NN_{OneForAll}$.

[2]The used dataset for the comparison was constructed from different sound sources and different room dimensions with way more considerable reverberation time. We recall that the RT_{60} was equal to 0.25s for the training and to 0.30s for the validation, and it is equal to 0.4s for the dataset used for the comparison between all the approaches.

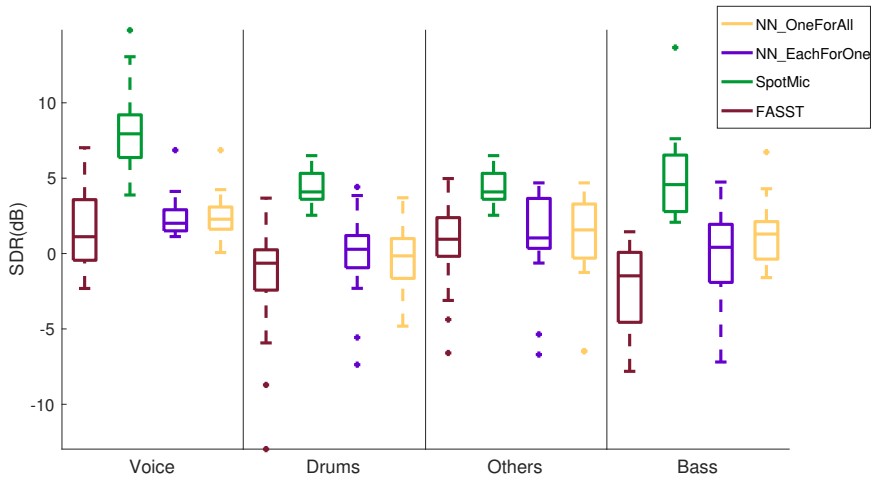


Fig. 7.9 Comparison between all the studied sound source separation in this Ph.D. in terms of SDR.

	Instrument	max	median	mean	min
$\Delta_{SDR}(dB)$	Voice	7.95	5.66	5.70	3.82
	Drums	2.79	4.24	4.48	7.35
	Others	6.43	4.05	5.20	9.20
	Bass	6.93	3.27	3.60	3.67

Table 7.4 Δ_{SDR} between SpotMic and $NN_{OneForAll}$.

	FASST	SpotMic	$NN_{OneForAll}$	$NN_{EachForOne}$
Number of iterations	150	10	10	10
Sampling frequency	16 kHz	16kHz	16kHz	16kHz
Duration of mixtures	10s	10s	10s	10s
Number of channels	4	4	4	4
Number of sound sources	4	4	4	4
Time of processing	775.84s	56.42s	53.13s	53.13s

Table 7.5 Comparison of time of processing between all the approaches. The used computer set up is MacBook pro with 2,2 GHz Intel Core i7 processor and 16Go of Ram.

Regarding the performance of the sound source separation in terms of SDR we can say that the SpotMic approach proposed in Chapter 6 is the best. It gives the best scores for every type of sound source, as we show it in Fig. 7.9, Table 7.4 and

Table 7.5. Replacing the spot microphones with a trained neural network can be beneficial in terms of computational time as we show it in Table 7.5, we saved almost 3.3 seconds. We managed to get a better time than the SpotMic approach because of the first estimation of the sound source spectra, which is in real-time with the neural network methods. In the case of the SpotMic method, the interference reduction block takes a little bit of time. However, a neural network are restricted to the training data (type of sound sources, number of sound sources, type of environment). It can be challenging to generalize with all the different situations (related to the environment) that we can face, such as longer or smaller reverberation time or different types of sound sources. If we compare the performances of the neural networks approaches in Fig. 7.9 and Fig. 7.6 (Approach 3), we can see that a longer reverberation time results in reduction of the performance of the sound source separation. With the neural network approaches, we managed to get better scores in terms of SDR while saving a lot of time compared to FASST, as we show it in Fig. 7.9, Table 7.3 and Table 7.5. The neural network methods seem to estimate the contributions of the bass more accurately than FASST. We are aware of the fact that the comparison isn't fare due to the fact that we are running the EM algorithm for "FASST" with 150 iterations and 10 iterations for the other approaches. The comparison was done in order to show case the fact that we need less iterations for a better performance with the approaches that we are suggesting.

7.3 Conclusion

In this chapter, we propose to replace the proposed method in Chapter 6 by a trained neural network. As a proof of concept we chose to restrict our study to the separation of music signals, which can be useful for our application. The proposed approaches in this chapter work on ambisonic sound scenes that contains a singer, drums, a bass, and any other instrument.

In this chapter, we used neural networks to perform for the first time music separation in the ambisonic domain (as far as we know). Similarly to the state of the art for speech enhancement in the ambisonic domain, we proposed to estimate the OSMs of each instrument using neural networks. We investigated two different architectures: one where a single network is used to to compute all the masks at the same time, and an other one where each instrument mask is estimated with its corresponding neural network.

We investigate to use the estimated mask to compute the Wiener filter coefficient in three different ways. We compared the different strategies with each other. The best strategy seems to be training a neural network for each instrument to estimate

the OSMs, compute the instrument spectra, and use them to initialize the EM algorithm. This helps the EM algorithm to have a effective estimation of the Wiener filter coefficient with less iteration (in our case, we chose ten iterations).

We compared this approach to the one proposed in Chapter 5 and Chapter 6. We concluded, on the one hand, the efficiency of the neural network approach compared to the LGM approach with an NMF constraint. We obtained better SDR scores as well as saving time and resources. On the other hand, the SpotMic approach gives better SDR scores than the neural network approach. However, we managed to gain 25% less computational time with the neural network approach.

Neural networks are quite revolutionary in sound source separation problems. However, they come with several constraints, such as the number of sound sources in the sound field, their types. Another problem that we can face using such an approach is that the order of the inputs must be respected. Indeed, if a neural network is trained to have the first input as in our case, the first ambisonic channel and the last one as the beamforming towards the drums, the order of the inputs should be the same while using it. For our case, we can have the DoA using one of the algorithms tested in Chapter 4 Section 4.2 and apply the beamformer towards the instruments, but we can be clueless about their types. In the frame of our application, this problem can be solved by using the cameras to label each direction.

Two examples of sound sources separation with both proposed approaches in this chapter are given in **`https://hafsatimohammed.github.io/HTML_Files/Example1.html`** and **`https://hafsatimohammed.github.io/HTML_Files/Example2.html`**. The same mixtures are used as the examples given in the conclusion of Chapter 5 and Chapter 6 (for a comparison purpose). Just note that these mixtures were pulled from the test data set.

Chapter 8

Conclusion

8.1 Context and summary

This Ph.D. thesis focuses on the problem of navigating with 6DoF in the 3D sound fields that are acquired from a live recording. To this aim, we use ambisonics as 3D audio technology. Nowadays, there is a focus around virtual reality content that allows the user to move freely with 6DoF. Most of the proposed contents these past decades were purely synthetic images and sound. By synthetic, we mean they do not present a real environment. If ever we want to navigate virtually in a real environment, one must first record it, and therefore, use cameras and microphones. My Ph.D. work treats the audio aspect of this application. We chose to work with ambisonics as a 3D technology due to its several advantages for this application.

The only problem with ambisonics is the difficulty in changing the point of view. Indeed, if ever a sound field is recorded and represented in the ambisonic domain, the representation of the entire sound field is given at the recording position. In order to simulate a movement from a point to another, the point of view must be changed.

To respond to the problem, we developed a navigation strategy that is based on sound source localization and sound source separation. For our strategy, we proposed two variants. For the first variant (Chapter 4 Section 4.1.1), we applied a simple plane wave decomposition using full-band beamforming techniques, and a reconstruction of the sound field according to the current user position. For the second variant (Chapter 4 Section 4.1.2) , we proposed to decompose the ambisonic sound field using a multichannel sound source separation, followed by a plane wave decomposition, and a reconstruction of the ambisonic sound field according to the current user position. For the sound source localization, we surveyed several approaches in the ambisonic domain and the microphone domain. We adapted the microphone domain methods to the ambisonic domain and assessed them through some numerical experiments. The results were satisfying in terms of localizing the sound sources. Our strategy was tested

using an objective metric. The numerical experiment showed that the second variant of our strategy was efficient compared to one of the best approaches in state of the art. The second variant of our strategy is based on multichannel sound source separation of ambisonic sound fields. However, such techniques are lacking in terms of research. In the microphone domain, there is a multichannel sound separation approach that was proposed ten years ago and had never been used in the ambisonic domain yet. The approach is known as the multichannel sound source separation based on the local Gaussian model. We derived the equations of the model from the microphone domain to the ambisonic domain. It turned out that such a technique can be used in the ambisonic domain. We validated the approach with some numerical experiments.

We proposed to use the local Gaussian model approach along with some side information that is coming from live recording spot microphones. We proposed a method to help the pre-processing of the side information and validate the efficiency of each block of the workflow through some numerical experiments.

We finally proposed to use neural networks in the place of the spot microphones for musical content. We developed two different neural networks and compared the performances of the multichannel sound source separation through numerical experiments.

Two examples of all the studied sound source separation approaches are given in this web page: "`https://hafsatimohammed.github.io`".

8.2 Contributions and conclusions

This Ph.D. thesis has concluded with the following contributions:

- We adapted some sound source localization approaches from the microphone domain to the ambisonic domain and validated their operation through some numerical experiments. The performance of these approaches gives excellent values that are in the same order of magnitude as sophisticated approaches from state of the art.

- We proposed to navigate in ambisonic sound fields with two different strategies. The first strategy (Chapter 4 Section 4.1.1) is not different from what exists in state of the art. The idea is to use plane-wave decomposition to deconstruct the sound field, followed by a reconstruction of it that depends on the user movements. The only difference compared to existing methods is that we recommend during the decomposition of the sound field to take advantage of the number of channels and use the proposed beamformer (regularized pseudo-inverse). The second strategy (Chapter 4 Section 4.1.2) has not been proposed before. We aimed

to: First, decompose the ambisonic mixture into J different mixtures using multichannel sound source separation techniques. Second, decompose each mixture with a plane wave decomposition followed by a reconstruction that depends on the user movement.

- To check the proper functioning of our strategies, we compared our navigation strategy to the best existing navigation approach according to [3]. We took into consideration the same objective metric as in [3].We concluded on the outperformance of our second strategy to the reference.
- We considered time-frequency mask methods as a solution to the multichannel sound source separation problem. These approaches do a great job in terms of separating the sound sources from each other. However, they introduce lot of artifacts, which is not great for our application. Therefore, we had to disregard them
- We verified and validated the ability for the Local Gaussian model approach to handle the multichannel sound source separation in the ambisonic domain [50]. We run some experiments in which we compare the performance of such decomposition. We investigated the influence of the number of channels. Indeed, on the one hand, in the ambisonic domain, the larger the number of channels, the better the performance. On the other hand, it is not the case in the microphone domain, for which the performance did not improve when there were more than 9 microphones. To this aim, we chose to compare the performance in both domains with the same number of channels/microphones, which were equal to nine. In the ambisonic domain, we obtained better performance. We concluded on the validation of the LGM approach in the ambisonic domain. We noticed that the reverberation time influences the performance of the separation, which drops if the reverberation time gets more significant.
- We proposed to add some spot microphones to guide the multichannel sound source separation that is based on the LGM approach. To this aim, we proposed a workflow. Through some experiments, we learned that our method works very well on speech, we obtained an SDR of 10dB even though we had two sound sources that were very close to each other with an angle difference of 5° and in a room with a moderate reverberation time $RT_{60} = 0.3s$. We investigated the influence of the ambisonic order on the performance, which does not seem to improve for orders. The reverberation time still has an impact on our approach. The type of sound sources has an influence on performance as well. We compared our approach to a conventional LGM method, and we gain at least 4.5dB in terms of SDR with much less time (13 times quicker).

- We proposed to replace the spot microphones and the method proposed in Chapter 6 by a trained neural network. We chose to work on music separation. Although there are a lot of articles about it (multichannel music separation using neural networks) in the microphone domain, it had never been used on ambisonics yet. We proposed two architectures and three different strategies for the Wiener filter computation. We compared all the approaches to each other and concluded on the outperformance of one of the strategies. We compared the neural network approaches to the approaches discussed in Chapter 5 and Chapeter 6. We learned that we could save a lot of time as well as gain performance compared to a conventional LGM approach. However, compared to the spot microphones approach, we can gain 3.2 seconds, but the performance of the separation is significantly worse.

8.3 Publications

M. Hafsati, N. Epain and J. Daniel. Editing ambisonic sound scenes. In *International Conference on Spatial Audio*. Graz, 2017.

M. Hafsati, N. Epain, R. Gribonval, N. Bertin. sound source separation in the higher ambisonics domain on *DAFx*. Birmingham, 2019.

8.4 Perspectives

First, the objective metric used to judge the efficiency of our strategy in Chapter 4 Section 4.4 does not allow to conclude on the outperformance of our approach compared to state of the art. Indeed, this metric does not take into consideration the spatial aspect of the navigation, which is very important. The only way to compare and conclude on the outperformance of an approach is through some subjective tests with enough test subjects. We propose to use a MUSHRA test with the ambisonic representation of sound fields at a given position as a reference. Compare the reference to mixtures that were required from our algorithms. At the end of the campaign, we will first discard the answers of the candidates who didn't give a high score to the truth. We will process the rest and conclude on the efficiency of our approach compared to state of the art.

Second, for the multichannel sound source separation based on the local Gaussian model, we wonder if it is possible to add a constraint on the spatial covariance matrix. We didn't focus on this aspect. Indeed, with this we can help the EM algorithm to find quickly Wiener filter coefficient. We can for example set this matrix with a constraint to fit a specific type of environment, a given reverberation time, or further the position of the sound sources regarding the main antenna.

Third, we propose to add more layers to our neural network and look for a way to extend the use to more than musical content. We wonder if it is possible to incorporate the phase and be able to estimate the coefficient of the Wiener filter directly using a deep neural network.

Fourth, we considered that we had a distance map matrix to solve the problem of the sound source distances from the user current position and movement. It would be more interesting to locate sound sources in terms of DoAs and distances.

Finally, for our navigation strategy and the used multichannel sound source separation, the sound sources were considered to be static. There is a way to treat the problem of moving sources by treating little frames. However, we wonder if it is possible to have a spatial covariance matrix that is time dependent, which will help to solve the problem of dynamic sound sources.

References

[1] V. R. Algazi, R. O. Duda, and D. Thompson. Dynamic binaural sound capture and reproduction, Feb. 19 2008. US Patent 7,333,622.

[2] A. Allen. Ambisonics sound field navigation using directional decomposition and path distance estimation, Jan. 17 2019. US Patent App. 15/647,741.

[3] A. Allen and W. B. Kleijn. Ambisonic soundfield navigation using directional decomposition and path distance estimation. In *International Conference on Spatial Audio*. Graz, 2017.

[4] S.-i. Amari, A. Cichocki, and H. H. Yang. Recurrent neural networks for blind separation of sources. In *Proc. Int. Symp. NOLTA*, pages 37–42. Citeseer, 1995.

[5] S. Araki, H. Sawada, and S. Makino. K-means based underdetermined blind speech separation. In *Blind speech separation*, pages 243–270. Springer, 2007.

[6] S. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. *Signal Processing*, 87(8):1833–1847, 2007.

[7] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1):121–133, 2009.

[8] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot. Blind spectral-gmm estimation for underdetermined instantaneous audio source separation. In *International Conference on Independent Component Analysis and Signal Separation*, pages 751–758. Springer, 2009.

[9] G. B. Arfken and H. J. Weber. Mathematical methods for physicists, 1999.

[10] M. Baque. *Analyse de scène sonore multi-capteurs: un front-end temps-réel pour la manipulation de scène*. PhD thesis, 2017.

[11] N. Barrett and S. Berge. A new method for b-format to binaural transcoding. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.

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

[13] A. J. Berkhout. A holographic approach to acoustic control. *Journal of the audio engineering society*, 36(12):977–995, 1988.

[14] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[15] J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In *Microphone arrays*, pages 19–38. Springer, 2001.

[16] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[17] M. M. Boone, E. N. Verheijen, and P. F. Van Tol. Spatial sound-field reproduction by wave-field synthesis. *Journal of the Audio Engineering Society*, 43(12):1003–1012, 1995.

[18] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, et al. Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[19] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[20] M. Bruneau. *Manuel d'acoustique fondamentale*. Hermes, 1998.

[21] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[22] J.-F. Cardoso. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 4, pages 1941–1944. IEEE, 1998.

[23] J.-F. Cardoso, H. Snoussi, and J. Delabrouille. Blind separation of noisy gaussian stationary sources. application to cosmic microwave background imaging. In *2002 11th European Signal Processing Conference*, pages 1–4. IEEE, 2002.

[24] C. Cerles and J. Daniel. Subjective and objective evaluation of a hoa processing chain. *International Conference on Spatial Audio (ICSA)*, 2015.

[25] E. C. Cherry and W. Taylor. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 26(4):554–559, 1954.

[26] A. Clifford, J. D. Reiss, et al. Microphone interference reduction in live sound. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[27] P. Coleman, P. Jackson, and J. Francombe. Audio object separation using microphone array beamforming. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[28] P. D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60(3):302, 1963.

[29] J. Daniel. Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. *Ph. D. Thesis, University of Paris VI, France*, 2000.

[30] J. Daniel and S. Moreau. Further study of sound field coding with higher order ambisonics. In *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004.

[31] J. Daniel, J.-B. Rault, and J.-D. Polack. Ambisonics encoding of other audio formats for multiple listening conditions. In *Audio Engineering Society Convention 105*. Audio Egineering Society, 1998.

[32] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[33] D. Di Carlo, K. Déguernel, and A. Liutkus. Gaussian framework for interference reduction in live recordings. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, 2017.

[34] T. Dozat. Incorporating nesterov momentum into adam.(2016). *Dostupné z: http://cs229. stanford. edu/proj2015/054_report. pdf*, 2016.

[35] N. Q. Duong, E. Vincent, and R. Gribonval. Spatial covariance models for under-determined reverberant audio source separation. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 129–132. IEEE, 2009.

[36] N. Q. Duong, E. Vincent, and R. 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, 2010.

[37] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency gaussian source models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 78–81. IEEE, 2005.

[38] C. Févotte, R. Gribonval, and E. Vincent. BSS_EVAL toolbox user guide–revision 2.0. 2005.

[39] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[40] M. A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10, 1973.

[41] M. A. Gerzon. Practical periphony: The reproduction of full-sphere sound. In *Audio Engineering Society Convention 65*. Audio Engineering Society, 1980.

[42] M. A. Gerzon. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871, 1985.

[43] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

[44] E. M. Grais, M. U. Sen, and H. Erdogan. Deep neural networks for single channel source separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3734–3738. IEEE, 2014.

[45] H. Groemer. *Geometric applications of Fourier series and spherical harmonics*, volume 61. Cambridge University Press, 1996.

[46] S. Hafezi, A. H. Moore, and P. A. Naylor. Multi-source estimation consistency for improved multiple direction-of-arrival estimation. In *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, pages 81–85. IEEE, 2017.

[47] S. Hafezi, A. H. Moore, and P. A. Naylor. Multiple doa estimation based on estimation consistency and spherical harmonic multiple signal classification. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1240–1244. IEEE, 2017.

[48] S. Hafezi, A. H. Moore, and P. A. Naylor. Robust source counting and acoustic doa estimation using density-based clustering. In *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 395–399. IEEE, 2018.

[49] M. Hafsati, N. Epain, and J. Daniel. Editing ambisonic sound scenes. In *International Conference on Spatial Audio*. Graz, 2017.

[50] M. Hafsati, N. Epain, R. Gribonval, and N. Bertin. Sound source separation in the higher order ambisonics domain. In *Digital Audio Effects Conference (DAFx)*, 2019.

[51] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–18, 2015.

[52] I. J. Hirsh. The influence of interaural phase on interaural summation and inhibition. *The Journal of the Acoustical Society of America*, 20(4):536–544, 1948.

[53] I. J. Hirsh. The relation between localization and intelligibility. *The Journal of the Acoustical Society of America*, 22(2):196–200, 1950.

[54] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

[55] Y. Izumi, N. Ono, and S. Sagayama. Sparseness-based 2ch bss using the em algorithm in reverberant environment. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 147–150. IEEE, 2007.

[56] E. Jan, P. Svaizer, and J. L. Flanagan. Matched-filter processing of microphone array for spatial volume selectivity. In *Proceedings of ISCAS'95-International Symposium on Circuits and Systems*, volume 2, pages 1460–1463. IEEE, 1995.

[57] D. P. Jarrett, E. A. Habets, and P. A. Naylor. 3d source localization in the spherical harmonic domain using a pseudointensity vector. In *2010 18th European Signal Processing Conference*, pages 442–446. IEEE, 2010.

[58] J. Jouhaneau and M. Rossi. *Notions élémentaires d'acoustique, électroacoustique: les microphones et les haut-parleurs: exercices et problèmes résolus*. Tec & Doc, 1994.

[59] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 5, pages 2985–2988. IEEE, 2000.

[60] W. Koenig. Subjective effects in binaural hearing. *The Journal of the Acoustical Society of America*, 22(1):61–62, 1950.

[61] E. K. Kokkinis and J. Mourjopoulos. Unmixing acoustic sources in real reverberant environments for close-microphone applications. *Journal of the Audio Engineering Society*, 58(11):907–922, 2010.

[62] S. Koyama, S. Shimauchi, and H. Ohmuro. Sparse sound field representation in recording and reproduction for reducing spatial aliasing artifacts. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4443–4447. IEEE, 2014.

[63] M. Kronlachner. Spatial transformations for the alteration of ambisonic recordings. *Master's thesis, Graz University of Technology*, 2:6, 2014.

[64] M. Kronlachner and F. Zotter. Spatial transformations for the enhancement of ambisonic recordings. In *International Conference on Spatial Audio*, 2014.

[65] M. Kühne, R. Togneri, and S. Nordholm. Time-frequency masking: Linking blind source separation and robust speech recognition. In *Speech Recognition*. IntechOpen, 2008.

[66] M. Kühne, R. Togneri, and S. Nordholm. A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation. *Signal Processing*, 90(2):653–669, 2010.

[67] P. Lecomte. *Ambisonie d'ordre élevé en trois dimensions: captation, transformations et décodage adaptatifs de champs sonores*. PhD thesis, Conservatoire national des arts et metiers-CNAM, 2016.

[68] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[69] Z. Li and R. Duraiswami. Flexible and optimal design of spherical microphone arrays for beamforming. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):702–714, 2007.

[70] J. Licklider. The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20(2):150–159, 1948.

[71] D. G. Malham. Higher order ambisonic systems for the spatialisation of sound. In *ICMC*, 1999.

[72] N. Mariette, B. F. Katz, K. Boussetta, and O. Guillerminet. Sounddelta: a study of audio augmented reality using wifi-distributed ambisonic cell rendering. In *Audio Engineering Society Convention 128*. Audio Engineering Society, 2010.

[73] D. R. Moore and A. J. King. Auditory perception: The near and far of sound localization. *Current Biology*, 9(10):R361–R363, 1999.

[74] S. Moreau. Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore higher order ambisonics: microphone 3d et contrôle de distance. *Ph.D. thesis University of Maine, Le Mans, France*, 2006.

[75] S. Moreau and J. Daniel. Study of higher order ambisonic microphone. In *7ème Congrès Français d'Acoustique (Joint congress CFA-DAGA'04)*, 2004.

[76] S. Moreau, J. Daniel, and S. Bertet. 3D sound field recording with higher order ambisonics–Objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the AES*, pages 20–23, 2006.

[77] P. H. Myers. Three-dimensional auditory display apparatus and method utilizing enhanced bionic emulation of human binaural sound localization, Mar. 28 1989. US Patent 4,817,149.

[78] O. Nadiri and B. Rafaely. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1494–1505, 2014.

[79] G. Naithani, T. Barker, G. Parascandolo, L. Bramsl, N. H. Pontoppidan, T. Virtanen, et al. Low latency sound source separation using convolutional recurrent

neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 71–75. IEEE, 2017.

[80] A. Narayanan and D. Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7092–7096. IEEE, 2013.

[81] R. O. Neubauer. Estimation of reverberation time in rectangular rooms with non-uniformly distributed absorption using a modified fitzroy equation. *Building Acoustics*, 8(2):115–137, 2001.

[82] R. Nicol. *Représentation et perception des espaces auditifs virtuels*. PhD thesis, Université du Maine, 2010.

[83] R. Nicol. Sound spatialization by Higher Order Ambisonics: Encoding and decoding a sound scene in practice from a theoretical point of view. In *International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[84] R. Nicol, J. Daniel, M. Emerit, G. Pallone, D. Virette, N. Chetry, P. Guillon, and S. Bertet. Le son 3d dans toutes ses dimensions. *Acoustique & techniques*, (52):43–50, 2008.

[85] J. Nikunen and A. Politis. Multichannel nmf for source separation with ambisonic signals. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 251–255. IEEE, 2018.

[86] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich. A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[87] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2009.

[88] A. Ozerov and C. Févotte. Multichannel Nonnegative Matrix Factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[89] A. Ozerov, C. Févotte, and E. Vincent. An introduction to multichannel nmf for audio source separation. In *Audio Source Separation*, pages 73–94. Springer, 2018.

[90] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on audio, speech, and language processing*, 20(4):1118–1133, 2011.

[91] Y. Peled and B. Rafaely. Study of speech intelligibility in noisy enclosures using optimal spherical beamforming. In *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, pages 285–289. IEEE, 2008.

[92] Y. Peled and B. Rafaely. Study of speech intelligibility in noisy enclosures using spherical microphones arrays. In *2008 Hands-Free Speech Communication and Microphone Arrays*, pages 160–163. IEEE, 2008.

[93] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. Crnn-based joint azimuth and elevation localization with the ambisonics intensity vector. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 241–245. IEEE, 2018.

[94] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40. IEEE, 2018.

[95] L. Perotin, R. Serizel, E. Vincent, and A. Guérin. Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):22–33, 2019.

[96] P. Pertilä and J. Nikunen. Distant speech separation using predicted time–frequency masks from spatial features. *Speech communication*, 68:97–106, 2015.

[97] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. Habets. Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[98] V. 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*. Audio Engineering Society, 2006.

[99] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[100] B. Rafaely. Spherical microphone array with multiple nulls for analysis of directional room impulse responses. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 281–284. IEEE, 2008.

[101] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher. Spherical microphone array beamforming. In *Speech Processing in Modern Communication*, pages 281–305. Springer, 2010.

[102] S. Rickard. The duet blind source separation algorithm. In *Blind speech separation*, pages 217–241. Springer, 2007.

[103] Y. Salaün, E. Vincent, N. Bertin, N. Souviraa-Labastie, X. Jaureguiberry, D. T. Tran, and F. Bimbot. The Flexible Audio Source Separation Toolbox Version 2.0. In *ICASSP*, 2014.

[104] C. Schörkhuber, P. Hack, M. Zaunschirm, F. Zotter, and A. Sontacchi. Localization of multiple acoustic sources with a distributed array of unsynchronized first-order ambisonics microphones. In *Congress of Alps-Adria Acoustics Assosiation, Graz, Austria*, 2014.

[105] F. Schultz and S. Spors. Data-based binaural synthesis including rotational and translatory head-movements. In *Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception*. Audio Engineering Society, 2013.

[106] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann. Spatial diffuseness features for dnn-based speech recognition in noisy and reverberant environments. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4380–4384. IEEE, 2015.

[107] B. Shore, I. Leiper, and B. Hiles. Stereo origination, sound pick-up and stereo post production techniques. In *IEE Colloquium on Stereo Sound for Television*, pages 2–1. IET, 1990.

[108] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, 2006.

[109] A. Southern, J. Wells, and D. Murphy. Rendering walk-through auralisations using wave-based acoustical models. In *Signal Processing Conference, 2009 17th European*, pages 715–719. IEEE, 2009.

[110] T. Stanojevic. Virtual sound sources in the total surround sound system. In *Proceedings 137th SMPTE Technical Conference and World Media Expo*, pages 405–421. SMPTE, 1995.

[111] E. W. Start. Direct sound enhancement by wave field synthesis. 1997.

[112] S. S. Stevens and E. B. Newman. The localization of actual sources of sound. *The American journal of psychology*, 1936.

[113] G. Theile. Multichannel natural music recording based on psychoacoustic principles. In *AES 19 th International Conference*, 2001.

[114] O. Thiergart, G. Del Galdo, M. Taseska, and E. A. Habets. Geometry-based spatial sound acquisition using distributed microphone arrays. *IEEE transactions on audio, speech, and language processing*, 21(12):2583–2594, 2013.

[115] J. G. Tylka and E. Choueiri. Comparison of techniques for binaural navigation of higher-order ambisonic soundfields. In *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[116] J. G. Tylka and E. Choueiri. Soundfield navigation using an array of Higher-Order Ambisonics microphones. In *Audio Engineering Society conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2016.

[117] C. Uhle and J. Reiss. Determined source separation for microphone recordings using IIR filters. In *in 129th Convention of the Audio Engineering Society*. Citeseer, 2010.

[118] H. L. Van Trees. *Optimum array processing: part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[119] J. Vilkamo, T. Lokki, and V. Pulkki. Directional audio coding: Virtual microphone-based synthesis and subjective evaluation. *Journal of the Audio Engineering Society*, 57(9):709–724, 2009.

[120] E. Vincent. Contributions to audio source separation and content description. *HDR, Université Rennes 1, France*.

[121] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):91–98, 2005.

[122] E. Vincent, S. Araki, and P. 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.

[123] E. Vincent, S. Arberet, and R. Gribonval. Underdetermined instantaneous audio source separation via Local Gaussian Modeling. In *International Conference on Independent Component Analysis and Signal Separation*, pages 775–782. Springer, 2009.

[124] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[125] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies. Probabilistic modeling paradigms for audio source separation. In *Machine Audition: Principles, Algorithms and Systems*, pages 162–185. IGI global, 2011.

[126] E. Vincent, T. Virtanen, and S. Gannot. *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[127] T. Virtanen. Unsupervised learning methods for source separation in monaural music signals. In *Signal Processing Methods for Music Transcription*, pages 267–296. Springer, 2006.

[128] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007.

[129] P. Vogel. Application of wave field synthesis in room acoustics. 1995.

[130] A. Wabnitz, N. Epain, C. Jin, and A. Van Schaik. Room acoustics simulation for multichannel microphone arrays. In *Proceedings of the International Symposium on Room Acoustics*, pages 1–6. Citeseer, 2010.

[131] D. Wang. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in amplification*, 12(4):332–353, 2008.

[132] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830–1847, 2004.

[133] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov. Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays. *IEEE transactions on audio, speech, and language processing*, 18(1):2–16, 2010.

Appendix A

The ambisonic formalism

We use the spherical coordinates (r, θ, ϕ) to describe the space, They are represented in Fig.A.1 and are related to the Cartesian coordinates with the following equations:

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

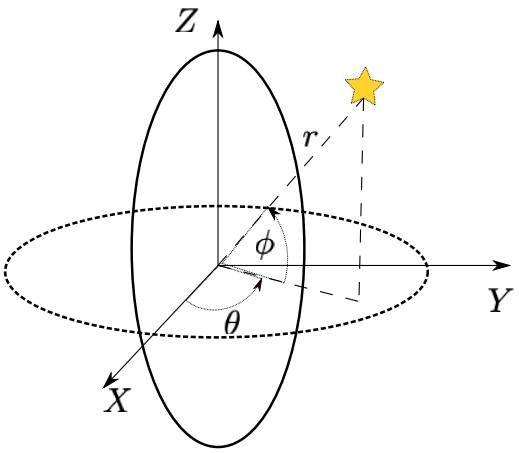


Fig. A.1 Spherical coordinate system. A given point in space is describe by radius r, azimuth θ, and elevation ϕ.

A.1 Spherical harmonic functions

In the context of the HOA, the real values spherical harmonics are used, which are different from the classical spherical harmonics used for examples in atomic physics.

Spherical harmonics $y_{le}(\theta,\phi)$[1] are directional functions of (θ,ϕ). They are defined with their degree $l \in \mathbb{N}$ and order $e \in \{-l, -l+1, \ldots, l-1, l\}$[45]:

$$y_{le}(\theta,\phi) = \sqrt{(2l+1)\epsilon_l \frac{(l-e)!}{(l+e)!}} P_{le}(sin(\phi)) \times \begin{cases} cos(e\theta) & if\ e > 0 \\ sin(e\theta) & if\ e < 0 \end{cases} \tag{A.2}$$

with $\epsilon_l = 0$ if $e = 0$ and $\epsilon_l = 2$ if $l > 0$, the functions P_{le} are the associated Legendre polynomial, they are given for x in $[-1, 1]$:

$$\begin{cases} P_{le}(x) = (1-x^2)^{\frac{e}{2}} \frac{\partial^e}{\partial x^e} P_l(x) \\ P_0(x) = 1 \\ P_1(x) = x \\ (l+1)P_{l+1}(x) = (2l+1)xP_l(x) - lP_{l-1}(x), \qquad l > 1 \end{cases} \tag{A.3}$$

If we consider all the directions, we can plot the spherical harmonic functions Fig.A.2.

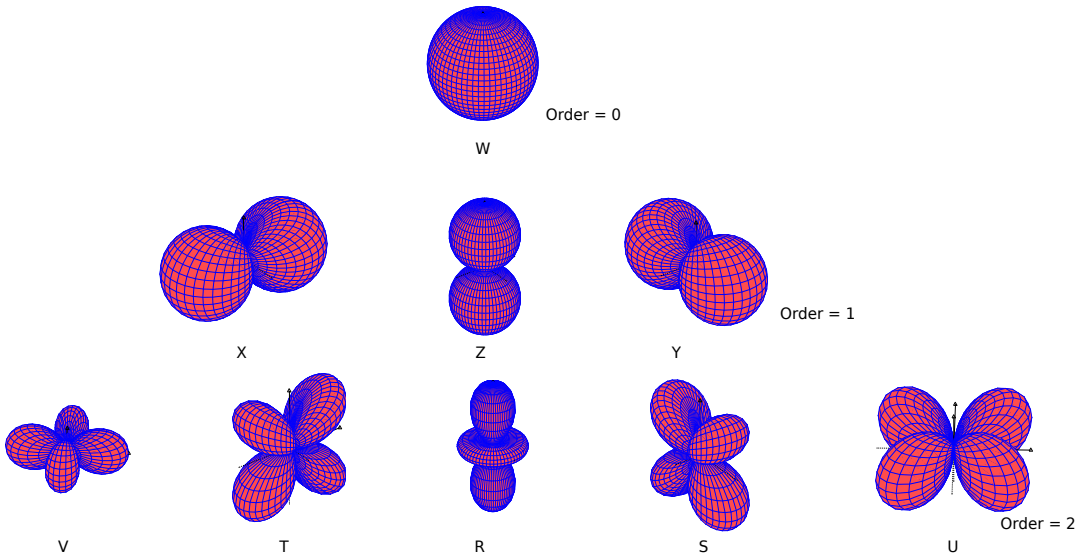


Fig. A.2 Spherical harmonic functions for orders up to l = 2

[1] $y_{le} \in \mathbb{R}$ is an element from the vector $\mathbf{y}$.

A.2 Decomposition of a sound field in the spherical harmonic basis

The idea behind the ambisonic format is to represent a sound field at a particular point in the spherical harmonic functions. The starting point behind this decomposition was the expression of the acoustic wave equation in the spherical coordinates (r, θ, ϕ):

$$\nabla^2 p(k, r, \theta, \phi) - \frac{1}{c^2} \frac{\partial^2 p(k, r, \theta, \phi)}{\partial t^2} = 0, \tag{A.4}$$

where $c \simeq 340 m/s$ represents the speed of sound.

According to [31], the solution to the acoustic wave equation leads to the decomposition of the sound pressure into Fourier-Bessel series, which is expressed in accordance with the spherical harmonic functions $y_{le}(\theta, \phi)$, and the spherical Bessel functions of the first kind $j_l(kr)$, and a weighting coefficients z_{le}. The expression of the solution is given by:

$$p(k, r, \theta, \phi) = \sum_{l=0}^{\infty} i^l j_l(kr) \sum_{-l \leqslant e \leqslant l} z_{le} y_{le}(\theta, \phi), \tag{A.5}$$

where $k = \frac{2\pi f}{c}$ represents the wavenumber.

We give in the following the first three analytical expressions of the spherical Bessel functions of the first kind $j_l(kr)$:

$$\begin{cases} j_0(x) = \frac{sin(x)}{x} \\ j_1(x) = \frac{sin(x)}{x^2} - \frac{cos(x)}{x} \\ j_2(x) = (\frac{3}{x^3} - \frac{1}{x}) sin(x) - \frac{3}{x^2} cos(x). \end{cases} \tag{A.6}$$

In Fig. A.3 we plotted the first three of the spherical Bessel functions of the first kind.

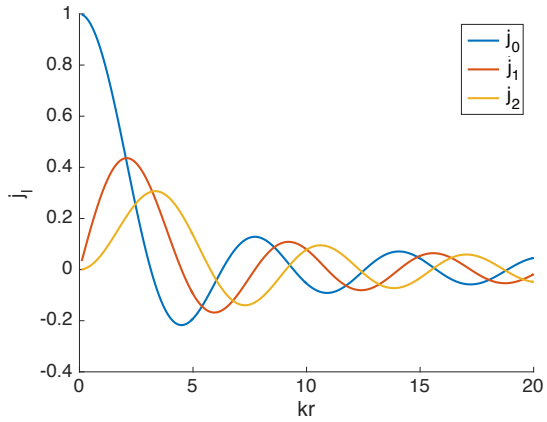


Fig. A.3 The first three spherical Bessel functions of the first kind

Physically it is impossible to represent the sound pressure as in Eq. (A.5). The equation must be truncated to a certain order L known as the ambisonic order:

$$p(k, r, \theta, \phi) \simeq \sum_{l=0}^{L} i^l j_l(kr) \sum_{-l \leqslant e \leqslant l} z_{le} y_{le}(\theta, \phi). \tag{A.7}$$

The truncation provides an approximation of the sound field pressure in regards to a limited number of spherical Fourier coefficients $\mathbf{z}_{le}$. Each order contains $2l + 1$ coefficients. The total number of coefficients for a given order L is:

$$M = \sum_{l=0}^{L} 2l + 1 = (L + 1)^2. \tag{A.8}$$

The spherical Fourier coefficients are the ambisonic signals. In order to understand how the ambisonic signals are found, we consider a plane wave coming from the direction (θ_p, ϕ_p). This plane wave carries a signal with an amplitude of s_p. The pressure at the point (r, θ, ϕ) is therefore given by:

$$p(k, r, \theta, \phi) = s_p e^{ikrcos(\gamma)}, \tag{A.9}$$

with γ being the angle between the observation direction (θ, ϕ) and the source direction (θ_p, ϕ_p). The pressure can be expressed in regards to the Legendre polynomial, and the spherical Bessel functions of the first kind [20]:

$$p(k, r, \theta, \phi) = s_p \sum_{l=0}^{L} (2l+1) i^l j_l(kr) P_l(cos(\gamma)). \tag{A.10}$$

A key point is to take into consideration the addition theorem [9]. Considering two distinct points (θ_1, ϕ_1), and (θ_2, ϕ_2), and γ as the angle between the two points we can write:

$$\sum_{e=-l}^{l} y_{le}(\theta_1, \phi_1) y_{le}(\theta_2, \phi_2) = (2l+1) P_l(cos(\gamma)). \tag{A.11}$$

Given the above theorem in Eq. (A.11), Eq. (A.10) can be written as follows:

$$p(k, r, \theta, \phi) = s_p \sum_{l=0}^{L} i^l j_l(kr) \sum_{e=-l}^{l} y_{le}(\theta, \phi) y_{le}(\theta_p, \phi_p). \tag{A.12}$$

By term identification with Eq. (A.7), we can identify the ambisonics signals as:

$$z_{le} = s_p y_{le}(\theta_p, \phi_p) \tag{A.13}$$

For simplicity, we drop the indexes l and e. Considering a sound field that contains J plane waves, and each plane wave j is carrying at the time t a signal of an amplitude s_j at the observation point, and coming from the direction (θ_j, ϕ_j), we can write the ambisonic signals $\mathbf{z}_t$ in $\mathbb{R}^M$, of an order L with $M = (L+1)^2$, as follows:

$$\mathbf{z}_t = \sum_{j=1}^{J} s_{j,t} \mathbf{y}(\theta_j, \phi_j), \tag{A.14}$$

we present the expression of $\mathbf{y}(\theta_j, \phi_j)$ for an order $L \leqslant 2$ in the following:

$$\mathbf{y}(\theta, \phi) = \begin{bmatrix} y_{00}^1(\theta, \phi) \\ y_{11}^1(\theta, \phi) \\ y_{11}^{-1}(\theta, \phi) \\ y_{10}^1(\theta, \phi) \\ y_{22}^1(\theta, \phi) \\ y_{22}^{-1}(\theta, \phi) \\ y_{21}^1(\theta, \phi) \\ y_{21}^{-1}(\theta, \phi) \\ y_{20}^1(\theta, \phi) \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{3} cos(\theta) cos(\phi) \\ \sqrt{3} sin(\theta) cos(\phi) \\ \sqrt{3} sin(\phi) \\ \sqrt{\frac{5}{12}} 3 cos(2\theta) cos^2(\phi) \\ \sqrt{\frac{5}{12}} 3 sin(2\theta) cos^2(\phi) \\ \sqrt{\frac{5}{3}} 3 cos(\theta) cos(\phi) sin(\phi) \\ \sqrt{\frac{5}{3}} 3 sin(\theta) cos(\phi) sin(\phi) \\ \sqrt{5} \frac{3 sin^2(\phi) - 1}{2} \end{bmatrix} \tag{A.15}$$

Eq. (A.14) can be written also as follows:

$$\mathbf{z}_t = \mathbf{s}_t \mathbf{Y}, \tag{A.16}$$

with the vector $\mathbf{s}_t \in \mathbb{R}^J$ and the matrix $\mathbf{Y} \in \mathbb{R}^{(J \times M)}$. This description can be applied to any sound field consisting of incoming waves with the hypothesis that the sources are far from the observation point.

Appendix B

Sound scenes simulations

B.1 Modeling of HOA microphone array

In order to obtain ambisonic signals of a specific sound scene, one must first record it at a particular point using a specific type of microphone array. The most natural analogy to represent a sound field in a spherical harmonics basis is to record it with an array of coincident microphones. However, it is physically impossible to have such microphone for a representation above the first order $L > 1$. A solution to the problem is to use a spherical microphone array. This is explained in more detail in Chapter 2 Section 2.3.

The first step in simulating sound scenes is to model a spherical microphone array. There are several commercially available spherical microphone arrays such as Eigenmike,[1] Zylia (ZM-1),[2] and Ambeo,[3] with which it is possible to produce ambisonic signals up to orders $L = 4$, $L = 3$, and $L = 1$, respectively. Other prototypes of spherical microphone arrays were proposed in the literature, such as the Orange Labs Prototype in [74], the University of Maryland prototype in [133], and the CNAM prototypes the MemsBedev and the SpherBedev in [67].

In our case, we based our simulations on the Eigenmike microphone array.[4] The modeling of the microphone was done by taking into consideration the geometrical and the acoustical description of the microphone array. Indeed the Eigenmike is in a spherical form with a radius $r_{Eigenmike} = 0.04$m, and 32 omnidirectional microphones distributed around the sphere. The angular position of the microphones are presented in Table. B.1. The acoustic impedance of the rigid sphere was taken into consideration. Indeed the sphere was considered to be perfectly rigid, and thereby, the acoustic

[1] https://mhacoustics.com/products
[2] https://www.zylia.co/
[3] https://fr-fr.sennheiser.com/microphone-3d-audio-ambeo-vr-mic
[4] `https://mhacoustics.com/products`

impedance was supposed to be infinite.

The impulse responses of the Eigenmike's 32 signals were then computed by considering 642 plane waves coming from a regularly sampled sphere[5] with the same origin as the modeled Eigenmike and a radius of 10m. The impulse responses corresponding to the plane wave coming from the direction ($\theta = 31,71°, \phi = 0°$) of each capsule of the Eigenmike are presented in Fig. B.1.

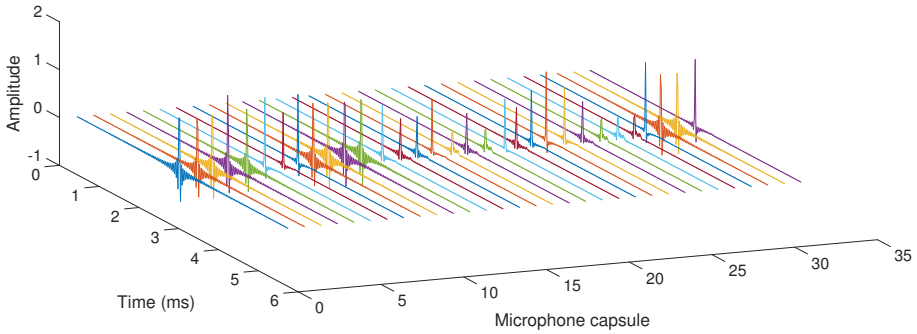


Fig. B.1 Impulse responses of the modeled Eigenmike corresponding to the plane wave coming from the direction ($\theta = 31,71°, \phi = 0°$).

number	θ	ϕ	number	θ	ϕ	number	θ	ϕ
1	0	21	12	-91	35	23	-225	0
2	-32	0	13	-90	69	24	-180	-35
3	0	-21	14	-90	32	25	-135	-58
4	32	0	15	-89	-31	26	-111	-35
5	-45	58	16	-180	-69	27	-135	0
6	-69	35	17	-212	21	28	-135	35
7	-45	0	18	-180	0	29	-269	69
8	0	-35	19	-148	-21	30	-90	32
9	45	-58	20	-180	0	31	90	-32
10	69	-35	21	-225	58	32	89	-69
11	45	0	22	-249	35			

Table B.1 Elevation (ϕ) and Azimuth (θ), in degrees, of the Eigenmike microphone capsules. The radius of the microphone is 4 cm. The origin of space is the center of the Eigenmike.

B.1.1 Simulations of room impulse responses

For a realistic simulation of indoor recordings, one must consider how sound waves reflect on the walls. Indeed in indoor recordings, sound waves coming from a given

[5]The samples represent a regular polyhedron of 642 points

sound source (depending on the type of source and the environment where the source is located and the position of the recording device and the sources) get reflected when they meet boundaries such as walls, floor ceiling, and objects.

To this aim we used a room acoustics simulation software called Multichannel Room Acoustics Simulator (MCRoomSim) [130]. With this software, we were able to configure the physical characteristics of the room, sources, and receivers, which supports any type of microphone array if the impulse responses were measured and given. In our case, the impulse responses of a real Eigenmike microphone (Section B.1) were used. This specific room acoustics simulation software was chosen because it handles ambisonics as well. Note that for our simulations, we wanted to be as realistic as possible. Therefore, only encoded ambisonic signals were used.[6]

The software models reflections of sound waves in a shoebox-shaped room, and gives the impulse responses of multiple numbers of sensors if defined. A receiver's impulse responses in a given room simulated based on several inputs; First, the user has the ability to control the physical characteristics of the room such as the dimensions of the room, the amount of scattering occurring when waves reflect off the walls, the frequency-dependent absorption of the boundaries (walls, ceiling, and floor) *et cetera*. Second, the user have control over the sources setup such as the number of sources, their position in the room, their type, such as omnidirectional, male or female speech, and their orientation. Third, the user have control over the receivers setup[7] which involves the type of the receiver (omnidirectional, spherical harmonic, impulse responses, cardioid ..., in our case we used most of the times the impulse response option to specify the modeled Eigenmike microphone), and its position and orientation in the room. Fourth, the simulation options, with the ability to control several features of the simulator.

Let us consider the fact that we want to simulate the a room with a given reverberation time. For instance a room of 10m×8m×3m, and a reverberation time of $RT_{60} = 0.4s$, which describes the required time for a sound to decay by 60 dB in close spaces. We know that the reverberation time is an image of the boundaries absorptions/reflections of the room. We may use the equation in [81], which is related to the Eyring-Kuttruff formula, in order to set these parameters (absorptions/reflections of walls, celling, floor), which is given by:

$$Absorption = 1 - \exp(-0.1611 \frac{V_{room}}{S_{room} RT_{60}}), \tag{B.1}$$

where V_{room} and S_{room} represent the volume and the total surface of the room respec-

[6] by encoded ambisonic signals we mean that the ambisonic signals were derived from a spherical microphone array signals

[7] The receiver corresponds to the output

tively. The room boundaries absorption were all then set to 0.3028. The reverberation time in regards to the frequencies is represented in Fig. B.2.

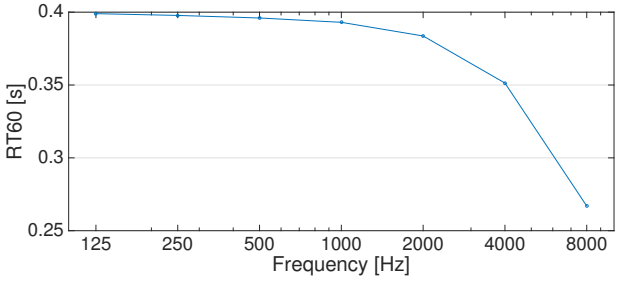


Fig. B.2 The predicted Eyring-Kuttruff Reverberation Time.

Let us consider the modeled Eigenmike in Section B.1, and one source placed at $(x = 3, y = 3, z = 1)$, and $(x = 0.5, y = 3, z = 1)$, respectively. In MCRoomSim, we can choose different types of sources. For this simulation, the sound source was configured to be a male speaker, which makes it a directional sound source. The orientation of the sound source was set to be toward the origin of the Eigenmike. A graphical representation of the described sound scene is represented in Fig. B.4.

The software provides as an output the impulse responses of the described room on each microphone capsule of the Eigenmike. The impulse responses of the described room in regards to our microphone array are represented in Fig. B.3.

If multiple receivers are present in the sound field (I receivers), such as in our case with several capsules in the spherical harmonic, we can compute the contribution of each sound source in each capsule:

$$c_{i,j,t} = [\alpha_i * s_j]_t, \tag{B.2}$$

where (*) denotes the convolution operator. The mixture $\mathbf{x}_t$ therefore given by:

$$\mathbf{x}_t = \sum_{j=1}^{J} \mathbf{c}_{j,t}, \tag{B.3}$$

where $\mathbf{c}_{j,t} = [c_{ij,t}]_{i=1...I}^{T}$ is a vector that contains the contribution of the j^{th} source in each microphone, and $\mathbf{x}_t = [x_{i,t}]_{i=1...I}^{T}$ the microphone array mixture.

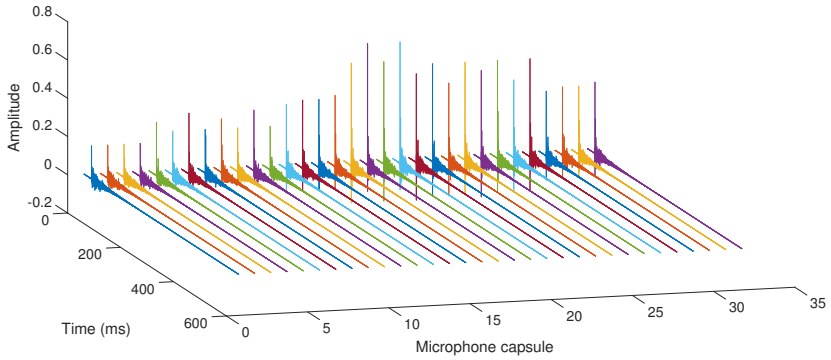


Fig. B.3 The described room impulse responses of the modeled Eigenmike.

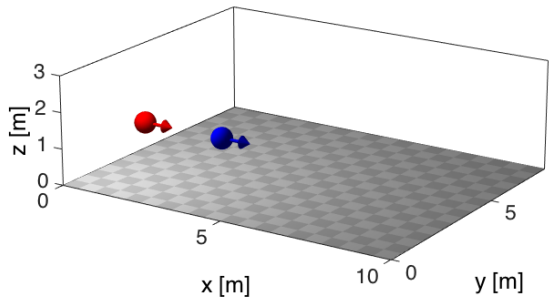


Fig. B.4 Graphical representation of the described shoebox with the receiver (Eigenmike) in blue and the source in red.

B.1.2 Encoding microphone signals

As explained in Section B.1.1, in order to take into consideration the imperfection of the encoded ambisonic sound scenes instead of generating directly the room impulse responses of the theoretical ambisonic signals:

- We generate the room impulse responses of the spherical microphone array.
- We convolve the capsules room impulse responses with the desired sound source signals Eq. (B.2) and sum them to create the mixture Eq. (B.3).
- We encode the spherical microphone array mixture $\mathbf{x}_t$ into ambisonic signals $\mathbf{z}_t$ using Eq. (2.16).

In the end, we get a fourth-order ambisonic mixture, which corresponds to 25 channels.

UNIVERSITE BRETAGNE LOIRE MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre: **Décomposition de scène sonore HOA pour navigation en six degés de liberté**

Mot clés : Ambisonics, Navigation, 6DoF, Séparation de source sonore, Wiener filter, Localisation de source sonore.

Resumé : Cette thèse s'inscrit dans le contexte multimedia dont le sujet technique est la navigation dans des champs sonores 3D. Contrairement aux contenus de réalité virtuelle, notre application vise les contenus issus de captations réelles. Nous utilisons l'ambisonique comme technologie d'audio 3D. Le problème d'utiliser ce genre de représentation de champ sonore réside dans la difficulté d'avoir 6 degrés de liberté, avec la possibilité de changer de point de vue. Afin de contourner ce problème, nous recommandons de faire une décomposition du format ambisonique en ondes planes. Cela a été déjà proposé dans plusieurs contributions dans l'état de l'art en utilisant des techniques de formation de voies en pleine bande. La particularité d'une de nos méthodes est d'utiliser des techniques de séparations de sources sonores multicanaux, avec laquelle nous cherchons les contributions de chaque source dans chaque canal ambisonique. Cela n'a jamais été utilisé auparavant pour faire de la navigation dans des contenus ambisoniques. Dans cette thèse, nous proposons différentes manières pour faire la séparation de source multicanaux dans le domaine ambisonique.

Title: **Higher order ambisonics sound scene decomposition for six degree of freedom navigation**

Keywords : Ambisonics, Navigation, 6DoF, Sound source separation, Wiener filter, Sound source localization.

Abstract : This Ph.D. thesis focuses on the problem of navigating with 6DoF in the 3D sound fields that are acquired from a live recording. We use ambisonic as a 3D sound technology. The problem with ambisonics is the difficulty in changing the point of view. Indeed, If ever a sound field is recorded and represented in the ambisonic domain, the representation of the entire sound field is given at the recording position. In order to simulate a movement from a point to another, the point of view must be changed. To respond to the problem, we recommend decomposing the ambisonic sound field into plane waves. This has already been proposed by several approaches in state of the art. However, the particularity of one of our methods is to use multi-channel sound source separation by looking for the contribution of each source in each channel. This has never been proposed before to navigate in ambisonic sound field. In this thesis, we propose several approaches to apply multichannel sound source separation in the ambisonic domain.