Ambisonic domain methods for immersive audio production enhancement

Author: Andrés Pérez López

TESI DOCTORAL UPF / year 2020

THESIS SUPERVISORS
Dr. Adan Garriga Torres
Dr. Emilia Gómez Gutiérrez
Department of Information and Communication Technologies

To my cat

Thanks thanks to.... thanks

Abstract

Recent years have witnessed a rise in immersive multimedia content, as a result of the developments in Virtual and Augmented Reality, where spatial audio plays a crucial role for the creation of a realistic experience. Ambisonic microphones allow the recording of sound scenes while preserving their spatial characteristics. For that reason, ambisonic has turned into the *de facto* standard for immersive audio transmission. Yet, few production tools exploit all the information available in such media. In the present thesis we raise the research question of obtaining relevant information from ambisonic recordings, in a way that is meaningful for producers. In order to answer the question, we have developed several methods and tools applied to different levels of the immersive sound production workflow; with each of those elements representing a novel contribution to the state-of-the-art in their respective fields.

Resum

En els últims anys hem assistit a l'auge del contingut multimèdia immersiu, afavorit pels avanços en Realitat Virtual i Augmentada. En aquest mitjà, el so espacial juga un paper molt important per a la creació d'una sensació de realisme. Els micròfons ambisonics permeten gravar escenes sonores preservant les seues característiques espacials. A causa d'això, ambisonics s'ha convertit en l'estàndard de facto per a la transmissió d'àudio immersiu. No obstant això, encara son poques les eines de producció que permeten explotar tota la informació disponible en aquest mitjà. En la present tesi plantegem la pregunta de com obtindre informació rellevant de gravacions ambisonics de manera automàtica, d'una forma que siga significativa per als productors. Per a això, hem desenvolupat una sèrie de mètodes i eines que s'apliquen a diferents nivells de la cadena de producció de so immersiu; cadascun d'aquests elements representa una contribució nova a l'estat de l'art en aquest camp.

Preface **is that really needed?**

Contents

List of figures **xvi**

List of tables **xvii**

1 Introduction **1**
1.1 Problem description and motivation 1
1.1.1 3D arrays 1
1.1.2 spherical microphone arrays 1
1.1.3 ambisonics 1
1.1.4 Current limitations of vr/ar production . . . 2
1.2 Goals 2
1.3 Context 2
1.4 Outline and main contributions 3
1.4.1 Thesis structure 3
1.4.2 List of Publications 5

2 Scientific Background **7**
2.1 Conventions 7
2.1.1 Reference system 7
2.1.2 Nomenclature 9
2.2 Spherical Harmonics 9
2.2.1 Definition 9
2.2.2 Spherical array processing 11
2.3 Ambisonics 12

2.3.1 Ambisonics Theory 12
2.3.2 Practical considerations 17
2.4 Parametric Spatial Audio Analysis 21
2.5 Spatial Coherence Analysis 25
2.6 Reverberation 26
2.6.1 SOFA Conventions 30
2.7 Signal Models 31

3 Blind reverberation time estimation **33**
3.1 Introduction 33
3.2 Signal Model 34
3.3 Baseline method 35
3.4 Proposed method 36
3.4.1 Dereverberation 37
3.4.2 System Identification 39
3.5 Experimental setup 40
3.5.1 Dataset 40
3.5.2 Setup 41
3.5.3 Evaluation metrics 43
3.6 Results 43
3.7 Conclusion 47

4 Coherence Estimation **49**
4.1 Introduction 49
4.1.1 Problem definition 50
4.2 Methods 50
4.2.1 Simulation 50
4.2.2 Recording 51
4.2.3 Data processing and metrics 51
4.3 Results and discussion 52
4.3.1 A-Format 52
4.3.2 B-Format 53
4.4 Conclusions 56

5 Sound Event Localization and Detection **57**
5.1 Introduction 57
5.2 Method 59
5.2.1 DOA estimation 59
5.2.2 Association 60
5.2.3 Beamforming 62
5.2.4 Deep learning classification back-end 62
5.3 Experiments 63
5.3.1 Dataset, evaluation metrics and baseline system 63
5.3.2 Parametric front-end 65
5.3.3 Deep learning classification back-end 65
5.4 Results and Discussion 68
5.5 Conclusion 71

6 Data generation and storage **73**
6.1 Ambiscaper 73
6.2 Ambisonic SOFA convention, pysofaconventions . . 73
6.3 masp library 73

7 Conclusions **75**
7.1 Summary of Contributions 75
7.2 Conclusion 76
7.3 Future work 76

List of Figures

1.1 todo caption . . . 3
1.2 todo caption . . . 5

2.1 Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values. . . . 11
2.2 Directive patterns of first-order ambisonic decoding. 16
2.3 Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from [Carpentier, 2017]. . . . 19
2.4 Parametric time-frequency spatial audio analysis of a first order ambisonic recording. From top to bottom: 1.) Magnitude spectrogram of the omnidirectional channel. 2.) and 3.) Azimuth and elevation of the estimated instantaneous narrowband DOAs $\Omega(k, n)$. 4.) Instantaneous narrowband diffuseness $\Psi(k, n)$. . 24
2.5 Room impulse response model, from [Murphy et al., 2017]. . . . 27
2.6 Room impulse response model, adapted from `http://www.bnoack.com/`. . . . 30

3.1 Experiment results for *speech* (left column) and *drums* (right column) datasets. Estimation error computed for each audio clip. . . . 45

3.2 Experiment results for *speech* (left column) and *drums* (right column) datasets. Total estimation error across audio clips and acoustic conditions. Top: boxplot. Bottom: histogram and density plot 46

4.1 *A-Format* coherence between microphone signals. Left: MSC as a function of the frequency of theoretical, simulated and recorded *((BLD,BRU)*, $N = 5, I = 64$) signals. Right: mean error $\bar{\varepsilon}$ of the recorded signals' MSC *(BLD,BRU)* compared to the simulated values, for all values of N and I.REDO FIGURE WITH N AND I 52

4.2 Estimated *B-Format* coherence (Δ) of a simulated diffuse sound field, as a function of the temporal averaging vicinity radius r. Left: $\Delta(k)$ for different values of r, with (coarse) and without (fine) application of radial filters. Right: mean and standard deviation of $\Delta(k)$ as a function of r. 53

4.3 *B-Format coherence* between microphone signals. Left: Δ of simulated and recorded ($N = 5, I = 64$) signals. Right: $\bar{\varepsilon}$ of the recorded signals coherence across all values of N and I.REDO FIGURE 55

5.1 System architecture. 59
5.2 DOA estimation architecture. 59
5.3 Association architecture. 61
5.4 Back-end architecture. 64
5.5 DCASE2019 Challenge Task 3 results, evaluation set. 70

List of Tables

2.1 Cartesian and spherical representation of characteristic points along the unit sphere. 8
2.2 Ambisonic decoding: standard values of $alpha_n$ weightings. Adapted from [Daniel, 2000]. 15
2.3 Reverberation time computation: usual reference levels 29

3.1 Baseline system: linear regression parameters 42
3.2 Experiment results 43

5.1 Parameter values for the selected configuration. Top: *DOA analysis* parameters. Bottom: *Association* parameters. 66
5.2 Results for development (top) and evaluation (bottom) sets. 68

Chapter 1

Introduction

1.1 Problem description and motivation

1.1.1 3D arrays

Sound propagates in 3D: need for 3D mic arrays to capture spatial properties

1.1.2 spherical microphone arrays

- even distribution of capsules
- mathematical convenience: spherical harmonics

1.1.3 ambisonics

advantages on the vr/ar context

- device independent
- intermediate storage format
- signal-independent transformations are easy
- de-facto standard for vr

1.1.4 Current limitations of vr/ar production

1.2 Goals

Research question: How can we exploit the characteristics of ambisonic recordings in order to extract signal-dependent, meaningful information from them?

1.3 Context

Different levels of applications:

- Acoustic Parameter Estimation (low level, audio2data)
  - Direction of Arrival estimation
  - Coherence analysis
  - Acoustic description (RT60, etc)
  - Source counting
- Signal Enhancement (high level, audio2audio)
  - Source Separation
  - Dereverberation / denoising
  - IR estimation
- Scene Description (high level, audio2data)
  - Event Detection
  - Acoustic Scene Classification

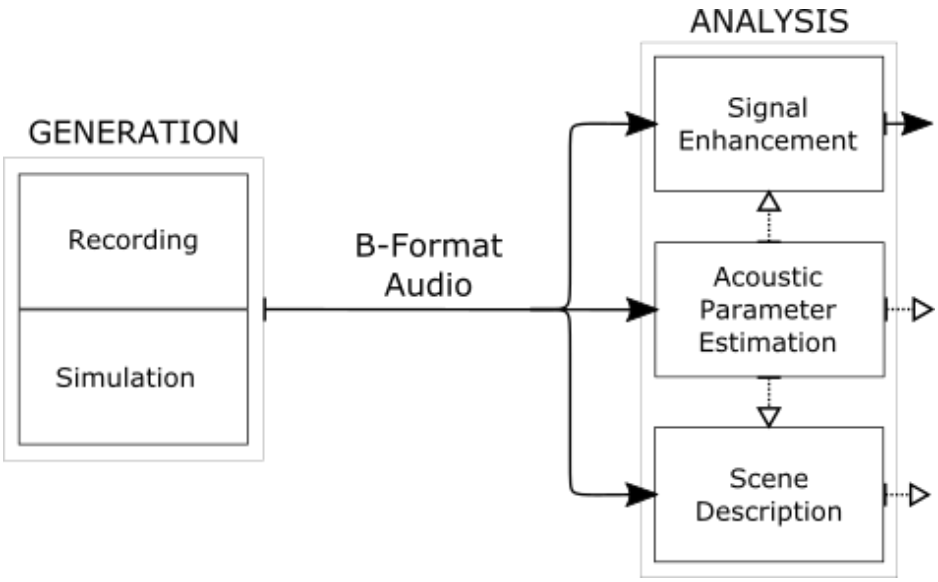


Figure 1.1: todo caption

1.4 Outline and main contributions

1.4.1 Thesis structure

The present Thesis is organised as follows.

Chapter 2 introduces the basic concepts that will be developed throughout the Thesis, including spherical harmonics and ambisonics, coherence estimation, parametric analysis or room acoustics. The Chapter also defines the signal models and the mathematical terminology.

Chapters 3, 4 and 5 develop the most significant academic contributions of the Thesis. **Chapter 3** presents a novel method for blind reverberation time in ambisonic recordings. To the best of our knowledge, this is the first method proposal specifically focusing on that problem. The method is based on a Multichannel Auto-Regressive model of the late reverberation, which allows for an effective dereverberation of the ambisonic sound scene, and en-

ables computation of the reverberation time from an estimation of the room impulse response. The evaluation metrics show a method performance similar to other state-of-the-art methods. **Chapter 4** analyses the response of tetrahedral microphone arrays, which are the simplest and most common form of ambisonic microphones, under spherically isotropic sound field. The analysis is performed using both simulated and recorded diffuse field, and the results quantify the differences between ideal and real values under a variety of conditions and estimators. In **Chapter 5**, a complete system for Sound Event Localization and Detection of ambisonic sound scenes is described. The algorithm comprises two different parts. First, a parametric analysis is performed on the ambisonic signal. The analysis yields spatial localization and temporal activities of the sound events present in the scene. Then, each of those events is assigned to a class label by means of a deep-learning classifier. The method is able to perform in a similar way to the baseline system, while greatly improving its localization capabilities.

Finally, **Chapter 6** presents some libraries and software utilities developed throughout the Thesis. All the code has been publicly released under open source licenses. The libraries include utilities for the creation of datasets, the storage and exchange of impulse response files in a standard way, and the implementation of convenience tools for acoustic and microphone array signal processing analysis. Although the libraries do not directly involve any scientific contribution, they can be a great help for scientific and innovative purposes; given the industial nature of the candidate's doctoral program, we have considered relevant to include them in the present Thesis. check

In order to place the different chapters within the problem context described in Section 1.3, we add to Figure 1.1 the Chapter numbers with our contributions, obtaining Figure 1.2.

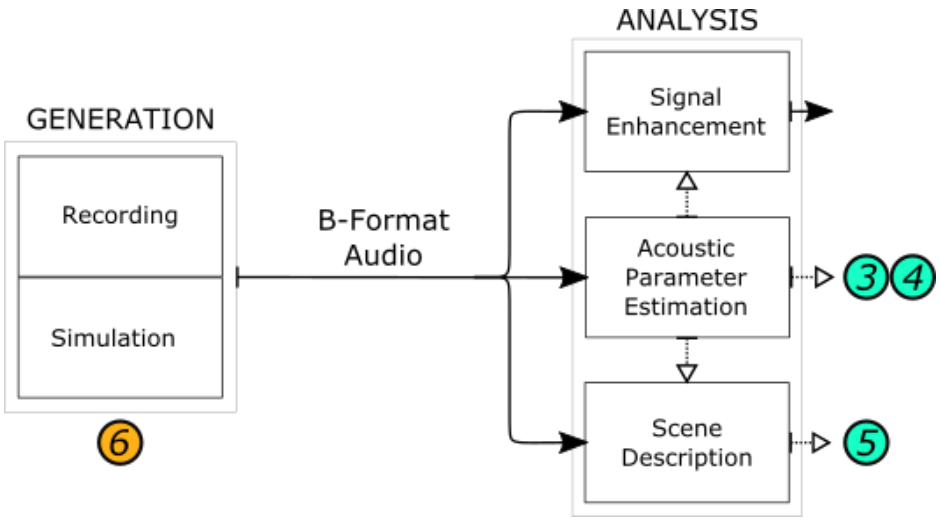


Figure 1.2: todo caption

1.4.2 List of Publications

In the following list we show the main scientific contributions of the Thesis, organised by Chapters:

- Chapter 3:
  "Blind reverberation time estimation from ambisonic recordings". A. Pérez-López, A. Politis and E. Gómez. Submitted to *IEEE 22nd International Workshop on Multimedia Signal Processing, 2020.*

- Chapter 4:
  "Analysis of spherical isotropic noise fields with an A-Format tetrahedral microphone". A. Pérez-López and N. Stefanakis. *The Journal of the Acoustical Society of America 146.4 (2019): EL329-EL334.*

- Chapter 5:
  A hybrid parametric-deep learning approach for sound event

localization and detection. A. Pérez-López, E. Fonseca and X. Serra. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).*

- Chapter 6:
  Ambiscaper: A Tool for Automatic Generation and Annotation of Reverberant Ambisonics Sound Scenes.
  A. Pérez-López. In *16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018.*

  Ambisonics directional room impulse response as a new convention of the spatially oriented format for acoustics. A. Pérez-López and J. De Muynke. In *Audio Engineering Society Convention 144. Audio Engineering Society, 2018.*

  pysofaconventions, a Python API for SOFA. A. Pérez-López. In *Audio Engineering Society Convention 148. Audio Engineering Society, 2020.*

  A Python library for Multichannel Acoustic Signal Processing. A. Pérez-López and A. Politis. In *Audio Engineering Society Convention 148. Audio Engineering Society, 2020.*

Moreover, as a result of the development of the Thesis, the following open-source libraries have been implemented and released. All of them are available through the author's GitHub page, `https://github.com/andresperezlopez`: describir cada uno? con respecto a paper o chapter?

- rt60 estimation: todo
- DCASE2019_task3
- ambiscaper
- pysofaconventions
- masp: Multichannel Acoustic Signal Processing library

Chapter 2

Scientific Background

2.1 Conventions

2.1.1 Reference system

In what follows, we will make use of a right-handed coordinate system, where the positive x-axis points towards the *front*, the positive y-axis points towards the *left*, and the positive z-axis points towards the *zenith* (North Pole).

Any position in the unit sphere may be described in spherical coordinates by two angles: the *inclination* angle ϑ, which accounts for the aperture with respect to the z-axis, and the *azimuth* angle φ, which represents the counter-clockwise angle with respect to the x-axis from the top-view. The value ranges are $0 \leq \vartheta \leq \pi$ for the inclination, and $0 \leq \varphi \leq 2\pi$ for the azimuth.

Table 2.1 shows the spherical coordinate values for some reference points on the unit sphere. Notice that the poles ($\vartheta = \pm\pi$) are a special case for the spherical coordinate system – in that case, the azimuth angle is not defined.

The transformation between spherical and cartesian coordinate

Table 2.1: Cartesian and spherical representation of characteristic points along the unit sphere.

Position	Cartesian	ϑ	φ
front	$[1, 0, 0]$	$\pi/2$	0
back	$[-1, 0, 0]$	$\pi/2$	π
left	$[0, 1, 0]$	$\pi/2$	$\pi/2$
right	$[0, -1, 0]$	$\pi/2$	$-\pi/2$
zenith	$[0, 0, 1]$	0	*
nadir	$[0, 0, -1]$	π	*

systems is given by the following relationship:

$$\begin{aligned} x &= \cos\varphi \sin\vartheta \\ y &= \sin\varphi \sin\vartheta \\ z &= \cos\vartheta \end{aligned} \tag{2.1}$$

The *elevation* angle θ provides an alternative way of describing the relationship with respect to the z-axis. θ is defined as the aperture with respect to the xy-plane, with positive values towards the positive z-axis. The relationship between elevation and inclination angles is:

$$\theta = \pi/2 - \vartheta \tag{2.2}$$

For the sake of compactness, a point in the unit sphere will be often represented by $\Omega = (\vartheta, \varphi)$.

Given the periodic nature of the azimuth angle, the descriptive statistic operations applied to ϑ will refer to the 2π*-periodic* version or the operator; this situation does not affect the inclination/elevation coordinate.

2.1.2 Nomenclature

is this name correct? Throughout the Thesis, we refer to time-domain signals with lowercase, e.g. $x(t)$, with t as the time index.

Time-domain signals transformed by the Short-Time Fourier Transform (STFT) are represented with uppercase, e.g. $X(k, n)$, where k is the frequency bin index, and n the time frame index.

Multichannel signals are in general denoted by a subscript variable index, usually with the letter m; for example, $x_m(t)$ or $X_m(k, n)$. Signals with an integer subscript index, such as $x_0(t)$, represent a specific channel of the corresponding multichannel signal.

In the context of ambisonic, subscripts and superscripts are used in signal names with a specific meaning; check Section 2.3 for a detailed explanation.

Vector notation is represented with boldface characters, e.g. $\boldsymbol{X}(k, n)$. When used, the way to construct the vectors will be specified.

2.2 Spherical Harmonics

2.2.1 Definition

Spherical harmonics are continuous functions defined on the sphere surface. Due to their mathematical properties, any spherical function can be decomposed as a combination of spherical harmonics, in what is known as the *Spherical Harmonics Expansion* [Jarrett et al., 2017].

Many different spherical harmonic definitions exist in the literature, with minor variations among them. In the following, we will use the real-valued, fully normalized spherical harmonics as defined by [Zotter and Frank, 2019]:

$$Y_n^m(\varphi, \vartheta) = N_n^{|m|} P_n^{|m|} \cos(\vartheta) \Phi_m(\varphi), \tag{2.3}$$

where the *normalization factor* N_n^m is:

$$N_n^m = (-1)^m \sqrt{\frac{2n+1}{2}\frac{(n-m)!}{(n+m)!}} \tag{2.4}$$

the *Legendre polynomials* P_n^m are defined as:

$$P_{n+1}^m = \begin{cases} \frac{2n+1}{n-m+1} x P_n^m, & \text{for } n = m, \\ \frac{2n+1}{n-m+1} x P_n^m - \frac{n+m}{n-m+1} P_{n-1}^m & \text{else}, \end{cases} \tag{2.5}$$

with $P_n^n = \frac{(-1)^n (2n)!}{2^n n!} \sqrt{1-x^{2^n}}$ and the initial term $P_0^0 = 1$, and Φ_m is the azimuthal part of the spherical harmonics:

$$\Phi_m(\varphi) = \frac{1}{\sqrt{2\pi}} \begin{cases} \sqrt{2}\sin(|m|\varphi), & \text{for } m < 0, \\ 1, & \text{for } m = 0, \\ \sqrt{2}\cos(m\varphi), & \text{for } m > 0. \end{cases} \tag{2.6}$$

One of the properties of the spherical harmonics is orthonormality on the sphere surface:

$$\int_{\mathbb{S}^2} Y_n^m(\varphi, \vartheta) Y_{n'}^{m'}(\varphi, \vartheta) \, \mathrm{d}cos\vartheta \, \mathrm{d}\varphi = \delta_{nn'}\delta_{mm'}, \tag{2.7}$$

where δ_{xy} represents the Kronecker delta operator:

$$\delta_{xy} = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{else}. \end{cases} \tag{2.8}$$

The spherical harmonics depend on the *order* $n \geq 0$ and the *degree* m, $|m| \leq n$ for each value of n. In practice, the maximum order N, $n \leq N$ determines the spatial resolution of the sound field expansion.

Through the spherical harmonic expansion, any sound field may be represented with a limited spatial resolution by the finite combination of all spherical harmonics up to order N. For a given order n, the number of spherical harmonic functions is $2n+1$. With

the accumulation of all orders up to N, the total number of spherical harmonics is given by $M = (N+1)^2$. Figure 2.1 depicts all spherical harmonics from orders 0 to 3.

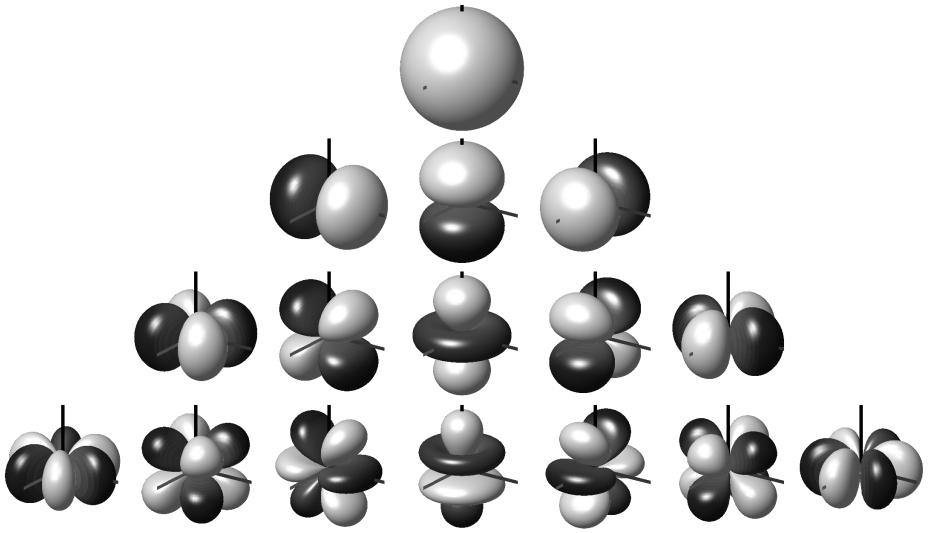


Figure 2.1: Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values.

2.2.2 Spherical array processing

Let us consider a sound field captured with a spherical microphone array, which contains Q capsules distributed around a spherical surface of radius R at the positions $\Omega_q, 1 \leq q \leq Q$. The captured frequency-domain signals $X_q(k)$ can be represented as the spherical harmonic domain signals $X_n^m(k)$ through the spherical harmonic transform of order n and degree m [Moreau et al., 2006]:

$$X_n^m(k) = \sum_{q=1}^{Q} X_q(k) Y_n^m(\Omega_q) \Gamma_n(kR), \tag{2.9}$$

is that actually valid? or Y should be the complex-valued spherical harmonics?

where the term $\Gamma_n(kR)$ models the radial transfer function, and depends on the microphone geometry. There are several possible sampling schemes of capsules along the sphere, each one having different properties; the reader is referred to [Rafaely, 2004] for a deeper insight.

By using this model, the maximum spherical harmonic order N that can be retrieved with negligible spatial aliasing depends on the number of microphone capsules [Moreau et al., 2006]:

$$N \geq (Q+1)^2. \tag{2.10}$$

Furthermore, the sphere radius R has also an effect on the operational bandwidth of the microphone. More precisely, for a given spherical harmonic order n, the maximum aliasing-free operational frequency is given by:

$$f_{max} = \frac{nc}{2\pi R}, \tag{2.11}$$

with c being the sound speed.
moreau has c/2Rgamma.

2.3 Ambisonics

2.3.1 Ambisonics Theory

Ambisonics is a spatial sound recording and playback technology initially developed during the 1970s [Gerzon, 1973], and further expanded into its modern formulation around the 2000s [Daniel, 2000]. Ambisonics is based on the idea of decomposing a sound field into its spherical harmonic representation.

Originally, the decomposition was limited to first-order spherical harmonics, as the so-called *First Order Ambisonics* (FOA); mainly

because of practical limitations. The technique was later formalized for arbitrary spherical harmonic orders, known as *Higher Order Ambisonics* (HOA). In general, with the term *ambisonics* we will be referring to the latter definition.

Ambisonic encoding

Let us consider a sound field composed of a point sound source S located in far-field at the angular position Ω_s. The sound pressure at the coordinate origin P can be expressed in terms of the spherical harmonic expansion of order N as: check equation, find references, how to explain the domain? extend also to multiple sources by superposition

$$P = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_s) S \tag{2.12}$$

The ordered set of values of all spherical harmonics up to order N, evaluated at the source position, is known as the *ambisonic coefficients*:

$$Y_n^m(\Omega_s) = [Y_0^0(\Omega_s), Y_1^{-1}(\Omega_s), \ldots, Y_N^N(\Omega_s)] \tag{2.13}$$

Furthermore, the process of multiplying the signal S by the ambisonic coefficients is known in the literature as the *ambisonic encoding*. The resulting signal vector is usually referred to as the *ambisonic* (or *B-Format*) signal S_n^m:

$$S_n^m = Y_n^m(\Omega_s) S \tag{2.14}$$

Although the term *B-Format* was initially introduced as an alternative name for first-order ambisonic signals [gerzon, tesis de daniel], it is nowadays common to use it as a synonim of ambisonic signals, without any order restriction. We will use the latter acception in what follows.

Historically, the name *B-Format* was used as an opposite of *A-Format*, which describes the signals recorded by a tetrahedral microphone array [Gerzon, 1975]. The tetrahedron is the simplest and most common form of spherical microphone arrays (indistinctly referred to as ambisonic microphones) with uniform capsule distribution. Again, the term *A-Format* is also currently employed for referring to the signals recorded by any spherical microphone array, regardless of the number or arrangement of capsules.

Likewise, the process of signal conversion from the spatial domain (microphone capsules) to the spherical harmonic domain (ambisonic signals), as in Eq. 2.9, is known as *A-B conversion*. A number of different approaches have been developed for this process, and the interested reader is referred to [Moreau et al., 2006] for more information.

In practice, there are two alternative ways to generate ambisonic signals. The first one is the *synthesis*, based on the direct application of ambisonics encoding (Eq. 2.12) to a monophonic signal. The second one is the *recording* with a spherical microphone array, followed by the aforementioned domain conversion.

Ambisonic Decoding

Conversely, the sound field reconstruction is performed by the *ambisonic decoding* operation. This process is equivalent to weight-and-sum beamforming in the spherical harmonic domain, and it is sometimes also referred to as the *virtual microphone* technique [Zotter and Frank, 2019].

Let us consider a loudspeaker located at the angular position Ω_p. In accordance with Eq. 2.12, the signal feed P is *decoded* from the ambisonic signal as:

$$P = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_s) S Y_n^m(\Omega_\ell) \alpha_n \tag{2.15}$$

Table 2.2: Ambisonic decoding: standard values of $alpha_n$ weightings. Adapted from [Daniel, 2000].

Decoding	N	n			
		0	1	2	3
basic	0	1			
	1	1	1		
	2	1	1	1	
	3	1	1	1	1
max-rE	0	0.577			
	1	0.775	0.4		
	2	0.861	0.612	0.305	
	3	0.906	0.732	0.501	0.246
in-phase	0	0.333			
	1	0.5	0.1		
	2	0.6	0.2	0.029	
	3	0.667	0.286	0.071	0.008

where α_n is a weighting factor which accounts for the beam directivity. There are several standard weightings used for different purposes; their values are shown in Table 2.2, and the first-order directive patterns are plotted in Figure 2.2.

The decoding equation 2.15 can be written in matrix form as:

$$P = S_n^m Y_n^m(\Omega_p)^T \alpha_n \tag{2.16}$$

where the superscript T represents the matrix transposition. This equation can be extended to the usual case of decoding to a loudspeaker array, comprised of L loudspeakers located at the positions $\Omega_L = [\Omega_{p_1}, \ldots, \Omega_{p_L}]$. In such case, the loudspeaker feed

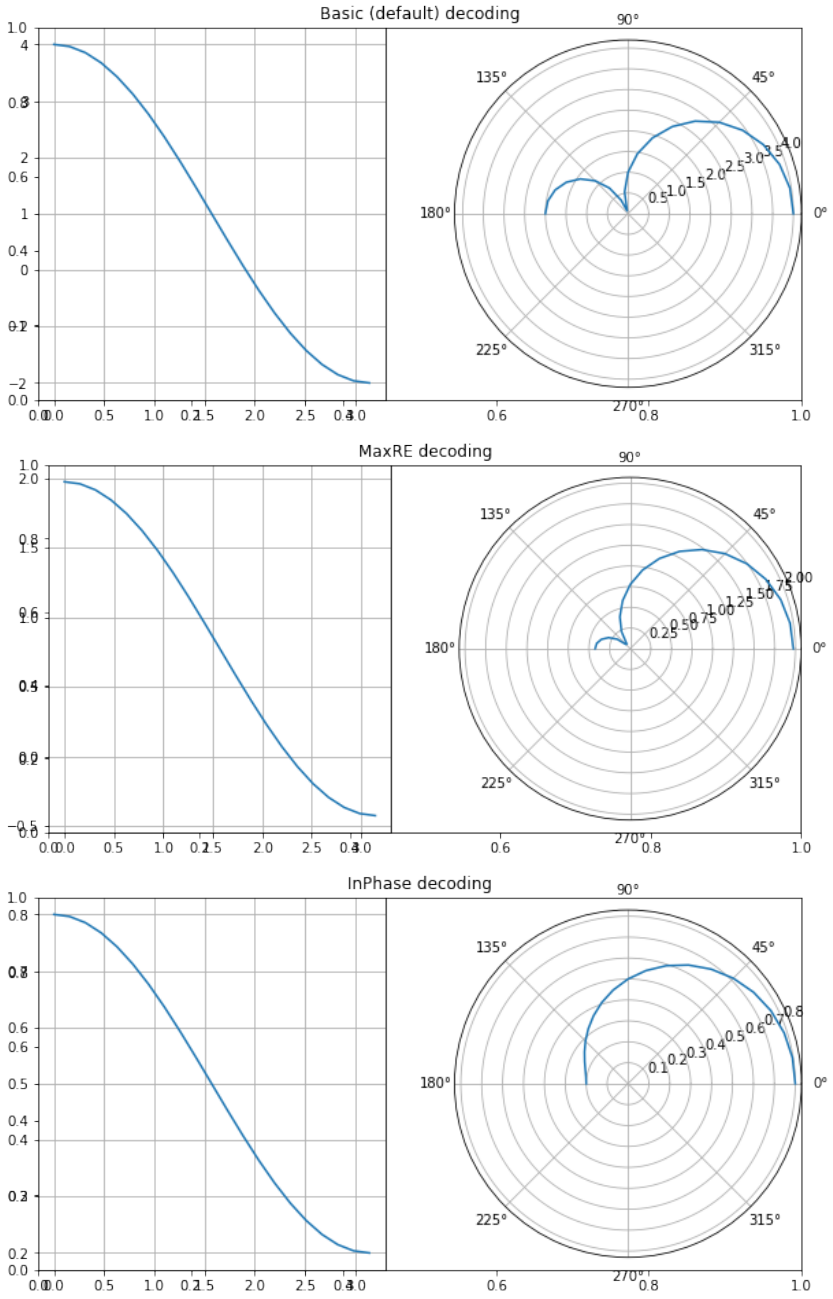


Figure 2.2: Directive patterns of first-order ambisonic decoding.

vector P_L can be written as:

$$P_L = S_n^m D, \tag{2.17}$$

where

$$D = \text{diag}(\alpha_n)[Y_n^m(\Omega_{p_1})^T, \ldots, Y_n^m(\Omega_{p_L})^T] \tag{2.18}$$

is a $M \times L$ matrix known as the *decoding matrix*, and $\text{diag}(\alpha_n)$ is a diagonal matrix of size M containing the values of α_n along the main diagonal. Although the matrix D is frequency-independent and depends solely on the loudspeaker array geometry, in practical scenarios it is usual to include frequency-dependent weightings, $\alpha_n(k)$, to improve the broadband sound field reconstruction [Daniel, 2000].

Furthermore, sound field reconstruction with Eq. 2.17 is only possible when the loudspeakers are evenly located on the 3D space; in other words, the speaker layout must take the form of one of the five *Platonic solids*: tetrahedron, cube, octahedron, dodecahedron or icosahedron. Provided that this condition is usually difficult to fulfil in real scenarios, there are several methods which allow ambisonic decoding for such *irregular* layouts. One of the most commonly used is the AllRAD method cite zotter. AllRAD proposes a two step decoding: first, the ambisonic signal is decoded to a nearly-uniform layout of virtual speakers. Then, the signals of the virtual speakers are further distributed into the real speakers by the *Vector-Based Amplitude Panning (VBAP)* method cite pulkki.

2.3.2 Practical considerations

Due to historical and practical reasons, there are two aspects that must be taking into account when working with ambisonic signals: *channel normalization* and *channel ordering*. In the following, the term *channels* will be used as a synonym for spherical harmonics,

as they are usually referred to in sound engineering contexts[1].

Channel normalization

Let us consider the spherical harmonics $Y_n^m(\Omega)$ as defined in Eq. 2.3. Due to the orthonormal property showed in Eq. 2.7, they follow the *fully 3d normalized* or *N3D* channel normalization convention. what about the 1/sqrt(4pi)???.

Alternatively, the *Schmidt 3d semi-normalized* or *SN3D* [daniel] convention is also of widespread usage. The conversion between *N3D* and *SN3D* is driven by the following expression:

$$Y_n^m(\Omega)^{(\text{N3D})} = \sqrt{2n+1} Y_n^m(\Omega)^{(\text{SN3D})} \tag{2.19}$$

MaxN is another existing convention. It defines all spherical harmonics as having a maximum absolute value of 1:

$$\max_{\Omega} |Y_n^m(\Omega)^{(\text{MaxN})}| = 1, \forall (n,m) \tag{2.20}$$

Finally, the *Furse-Malham* (or *FuMa*) normalization only differs from *Max-N* in the scaling of the zero-th order component:

$$Y_n^m(\Omega)^{(\text{FuMa})} = \begin{cases} 1/\sqrt{2}, & \text{if } n = 0, \\ Y_n^m(\Omega)^{(\text{MaxN})}, & \text{else.} \end{cases} \tag{2.21}$$

Each of the normalization schemes has its own particularities. For instance, *N3D* is the most mathematically straightforward, and spherical harmonics defined in that way can be directly used for both encoding and decoding (as in Eqs 2.12 and Eq. 2.15) – however, from a sound engineer point of view, other normalization schemes with maximum values below the unity might be preferred, such

[1] In fact, ambisonic signals are inherently multichannel, even though each channel corresponds to a spherical harmonic, and not to a loudspeaker feed as in traditional *channel-based* audio.

as *SN3D*. Besides this, *FuMa* has been historically the default normalization [Gerzon, 1985], while the more modern *N3D* and *SN3D* were popularized after [Daniel, 2000].

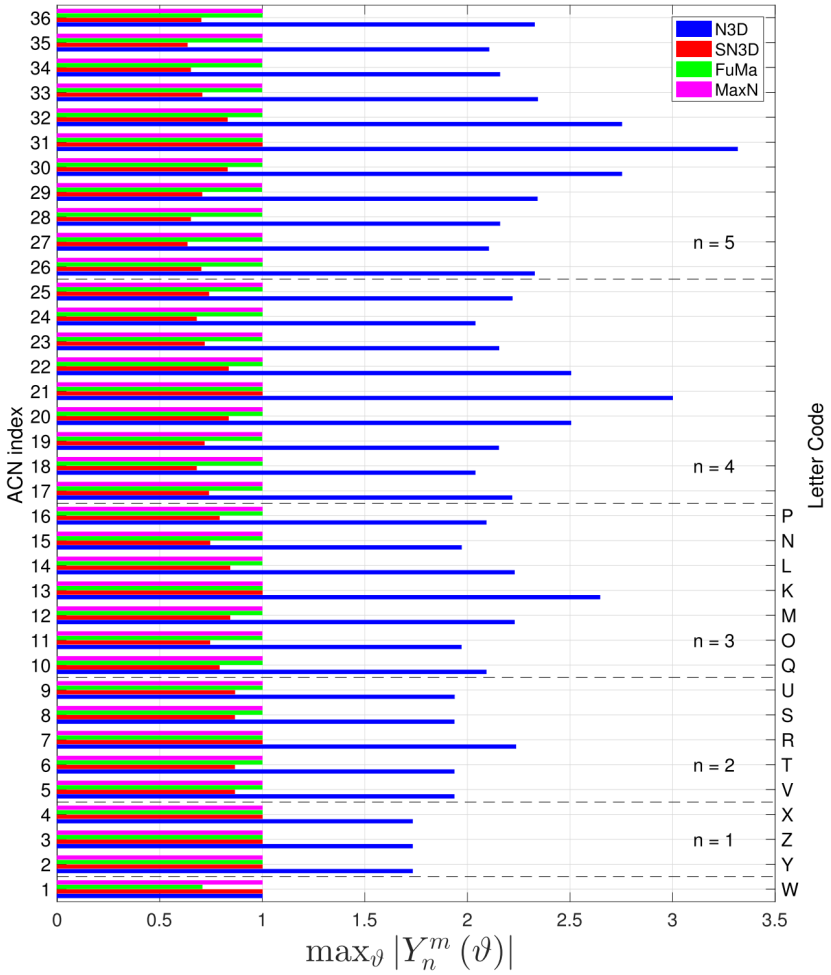


Figure 2.3: Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from [Carpentier, 2017].

As a summary, Figure 2.3 displays the different normalization schemes. The reader is referred to [Carpentier, 2017] for an extensive review on the topic.

Channel ordering

Channel ordering refers to the manner in which spherical harmonics, inherently organized in the 2D space by dimensions n and m, are sorted into a one-dimensional vector.

The *ACN* (from *Ambisonic Channel Number*) scheme follows from the mathematical description given in Eq. 2.13. The spherical harmonics are first ordered by ascending order n and, inside each order, by ascending degree m. The index of a given channel $i \in [0 \dots M-1]$ can be thus obtained by the following relationship:

$$i = n^2 + n + m \tag{2.22}$$

Historically, first-order ambisonic audio has followed what it might be called *traditional B-Format* channel ordering [Gerzon, 1985]. By this scheme, the four channels of a FOA signal S_n^m are referred to by the axis where the corresponding spherical harmonic steers, plus the name W for the zeroth order component:

$$S_n^m(\Omega)^{\text{(FuMa CO)}} = [W, X, Y, Z] \tag{2.23}$$

where:

$$\begin{aligned} W &= S_0^0(\Omega) \\ X &= S_1^1(\Omega) \\ Y &= S_1^{-1}(\Omega) \\ Z &= S_1^0(\Omega) \end{aligned} \tag{2.24}$$

This nomenclature was extended to second and third order, and is currently known as the *Furse-Malham* or *FuMa* channel ordering. The channel names use all english alphabet letters from K to Z in third order and, although there would be enough letters to go up

to fourth order, the unconvenience of the system was clear [Malham, 2003]. Figure 2.3 shows the equivalence between *FuMa* (*"letter code"*) and *ACN* channel names.

In practice, there exist two main combinations of channel normalization and ordering schemes:

- The *classical* approach, usually limited to first-order ambisonics, which uses *FuMa* normalization and channel ordering[2].
- The *modern* approach, inspired by the *ambix* file format [cite ambix], with *SN3D* normalization and *ACN* channel ordering.

Anyhow, the *classical B-Format* channel naming and ordering is still widely used when referring to first-order ambisonics.

2.4 Parametric Spatial Audio Analysis

Trough parametric analysis, sound fields may be described in terms of a small amount of sound sources and associate parameters. Such representation might reduce to a great extent the complexity of processing methods [Jarrett et al., 2017].

One of the most successful sound field parametric models is DirAC [Pulkki, 2007], which was originally conceived as a method for impulse response processing and spatial sound reproduction [Merimaa and Pulkki, 2005].

DirAC (acronym for *Directional Audio Coding*) is a perceptually motivated time-frequency (TF) domain method, based on the assumption that any sound field may be reproduced with high perceptual quality by considering two parameters: the sound field diffuseness and the most prominent sound *Direction-of-Arrival* (DOA) [Pulkki et al., 2018].

[2]In general, it may be expected that *early* ambisonic material follow these conventions without any explicit mention to them.

Let us consider a *SN3D*-normalized first-order ambisonic signal in time-frequency domain, $S_n^m(k,n)$. For the sake of clarity, we will use in this section *FuMa* channel notation and ordering (Eq. 2.23):

$$S_n^m(k,n) = [W(k,n), X(k,n), Y(k,n), Z(k,n)] \tag{2.25}$$

Given this representation, we can express the *pressure* $P(k,n)$ of the sound field as:

$$P(k,n) = W(k,n) \tag{2.26}$$

as well as the sound *pressure-gradient* (or *velocity*) $\boldsymbol{U}(k,n)$ as:

$$\boldsymbol{U}(k,n) = -\frac{1}{\rho_0 c}[X(k,n), Y(k,n), Z(k,n)], \tag{2.27}$$

where ρ_0 is the mean density of the medium, and c is the speed of sound.

The *active intensity* $\boldsymbol{I}(k,n)$, defined as the amount of transmitted acoustic energy, can be expressed in terms of sound pressure and velocity [Fahy and Salmon, 1990]:

$$\begin{aligned} \boldsymbol{I}(k,n) &= \Re\{P^*(k,n)\boldsymbol{U}(k,n)\} \\ &= -\frac{1}{\rho_0 c}\Re\{W^*(k,n)[X(k,n), Y(k,n), Z(k,n)]\}, \end{aligned} \tag{2.28}$$

where * represents the complex conjugate operator.

An estimate of the instantaneous DOA $\Omega(k,n)$ can be extracted from the intensity vector, interpreting each of its time-frequency bins as a point in the cartesian space. Effectively, the sound propagation direction is the opposite to the observed arrival direction.

$$\Omega(k,n) = \angle(-\boldsymbol{I}(k,n)), \tag{2.29}$$

with $\angle$ representing the spherical angle operator of a cartesian vector. The result of this computation must be understood as the

direction of the net energy flow, which in the case of a single plane-wave will correspond to the source position.

Another useful parameter is the *energy density* $E(k,n)$ [Stanzial et al., 1996]:

$$\begin{aligned} E(k,n) &= \frac{1}{2\rho_0 c^2}|P(k,n)|^2 + \frac{1}{2}\|\boldsymbol{U}(k,n)\|^2 \\ &= \frac{1}{2\rho_0 c^2}\Big(|W(k,n)|^2 + \|[X(k,n),Y(k,n),Z(k,n)]\|^2\Big). \end{aligned} \tag{2.30}$$

Finally, the *diffuseness* $\Psi(k,n)$ can be computed from the sound intensity and energy density [Merimaa and Pulkki, 2005]:

$$\begin{aligned} \Psi(k,n) &= 1 - \frac{\|\langle \boldsymbol{I}(k,n)\rangle\|}{c\langle E(k,n)\rangle} \\ &= 1 - 2\frac{\|\langle \Re\{W^*(k,n)[X(k,n),Y(k,n),Z(k,n)]\}\rangle\|}{\langle |W(k,n)|^2 + \|[X(k,n),Y(k,n),Z(k,n)]\|^2\rangle}, \end{aligned} \tag{2.31}$$

where the symbols $\langle\cdot\rangle$ represent the expectation operator, which is usually implemented as time-domain averaging.

Even though Eq. 2.31 (known as *DirAC's diffuseness*) is one of the most common ambisonic diffuseness estimators, several alternative formulations exist. Other diffuseness estimation procedures include the *coefficient of variation method* [Ahonen and Pulkki, 2009] and the more recent *COMEDIE* estimator [Epain and Jin, 2016]. In any case, in what follows, the term *diffuseness* and the symbol Ψ will refer by default to Eq. 2.31.

As a mathematical convenience, we will define the *B-Format coherence* as the complement of the diffuseness:

$$\Delta(k,n) = 1 - \Psi(k,n) \tag{2.32}$$

In conclusion, Figure 2.4 plots the spectrograms of the DOA $\Omega(k,n)$ and diffuseness $\Psi(k,n)$ of a FOA recording, which consists

of a sound source located at the front, plus a moderate amount of reverberation and background noise.

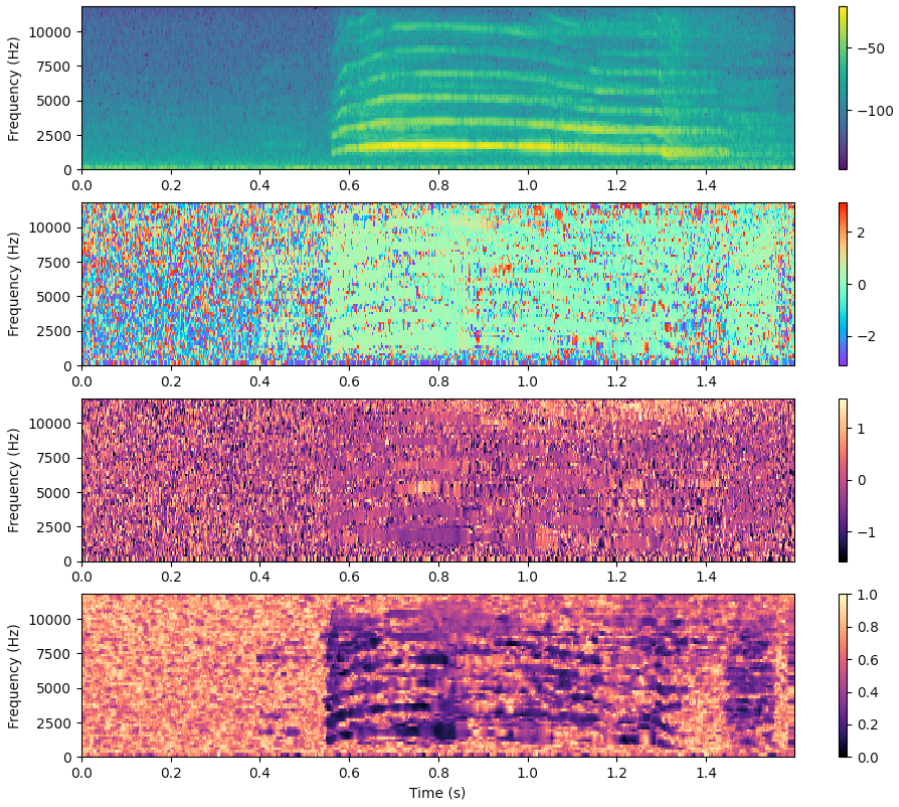


Figure 2.4: Parametric time-frequency spatial audio analysis of a first order ambisonic recording. From top to bottom: 1.) Magnitude spectrogram of the omnidirectional channel. 2.) and 3.) Azimuth and elevation of the estimated instantaneous narrowband DOAs $\Omega(k, n)$. 4.) Instantaneous narrowband diffuseness $\Psi(k, n)$.

2.5 Spatial Coherence Analysis

put this chapter in context or something

In the context of microphone array signal processing, diffuseness is commonly estimated through the *Magnitude Squared Coherence* (MSC) [Elko, 2001] between two frequency-domain signals S_1 and S_2, as a function of the *wavenumber* k and the microphone distance r:

$$\text{MSC}_{12}(kr) = \frac{|\,\langle S_1(kr) S_2(kr)^* \rangle\,|^2}{\langle |S_1(kr)|^2 \rangle\, \langle |S_2(kr)|^2 \rangle}, \tag{2.33}$$

where the $\langle \cdot \rangle$ operator represents the temporal expected value, and * defines the complex conjugate operator. In the case of spherical isotropic noise fields, Eq. (2.33) can be expressed in terms of microphone directivity patterns $T(\phi, \theta, kr)$ as [Elko, 2001]:

$$\begin{aligned}\text{MSC}_{12}(kr) &= \frac{|N_{12}(kr)|^2}{|D_{12}(kr)|^2} \\ &= \frac{|\int_0^{\pi}\int_0^{2\pi} T_1(\phi,\theta,kr) T_2^*(\phi,\theta,kr) e^{-jkr cos\theta} sin\theta d\theta d\phi|^2}{|\sqrt{\int_0^{\pi}\int_0^{2\pi} |T_1(\phi,\theta,kr)|^2 sin\theta d\theta d\phi}\sqrt{\int_0^{\pi}\int_0^{2\pi} |T_2(\phi,\theta,kr)|^2 sin\theta d\theta d\phi}|^2}.\end{aligned} \tag{2.34}$$

Moreover, the general expression of the directivity of a first-order differential microphone is given by the following relationship:

$$T_i(\Omega_i) = \alpha_i + (1 - \alpha_i)\cos\Omega_i, \tag{2.35}$$

where $i \in [1, 2]$ is the microphone index, Ω_i is the angle between wave incidence and microphone orientation axis, and $\alpha_i \in [0, 1]$ is the directivity parameter of the microphone i, which ranges from bidirectional ($\alpha_i = 0$) to omnidirectional ($\alpha_i = 1$).

For first-order differential microphones, there is a closed-form expression for the numerator and denominator of Eq. (2.34):

$$
\begin{aligned}
N_{12}(kr) &= \frac{\alpha_1\alpha_2 sin(kr)}{kr} \\
&+ \frac{(1-\alpha_2)(1-\alpha_2)(x_1x_2+y_1y_2)}{(kr)^3}(sin(kr)-krcos(kr)) \\
&+ \frac{z_1z_2}{kr^3}[((kr)^2sin(kr)+2krcos(kr))(1-\alpha_1)(1-\alpha_2)+2sin(kr)(1-\alpha_1)(1-\alpha_2)] \\
&+ \frac{z_1}{(kr)^3}[j(kr)^2\alpha_2 cos(kr)(\alpha_1-1)+jkr\alpha_2 sin(kr)(1+\alpha_1)] \\
&+ \frac{z_2}{(kr)^3}[j(kr)^2\alpha_1 cos(kr)(\alpha_2-1)+jkr\alpha_1 sin(kr)(1+\alpha_2)], \\
D_{12}(kr) &= \frac{\sqrt{3\alpha_1^2+(1-\alpha_1)^2}\sqrt{3\alpha_2^2+(1-\alpha_2)^2}}{3},
\end{aligned}
\tag{2.36}
$$

where x_i, y_i and z_i are the cartesian coordinates of the wave incidence angle Ω_i. check.

2.6 Reverberation

In the context of room acoustics, reverberation refers to *"the energy of a sound source that reaches a listener indirectly, by reflecting from surfaces within the surrounding space occupied by the sound source and the listener"* [Begault and Trejo, 2000]. Conversely, in anechoic or free-field conditions, where reverberation is not present, only the direct path of the sound source exists. Assuming linearity and time-invariance, room reverberation can be fully characterised by its impulse response (IR).

Reverberation models often consider two differentiated parts of the reverberant tail, based on both physical and perceptual characteristics: the *early reflections* and the *late reverberation*. Early reflections, as the name suggests, refers to the individual sound paths

arriving to the listener after a few reflections on the room surfaces, which cause some degree of attenuation. Early reflections typically arrive with a time difference between 1 and 80 ms after the direct path [Begault and Trejo, 2000]. The term late reverberation encompasses all sound paths arriving to the listener after many reflections. Since the temporal density of such reflections increases with time, late reverberation is often modelled in statistical terms. An schematic representation of a room impulse response (RIR) is shown in Figure 2.5.

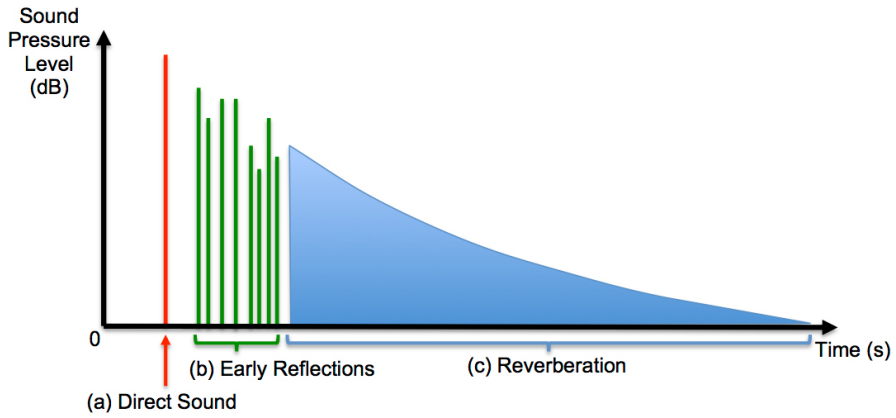


Figure 2.5: Room impulse response model, from [Murphy et al., 2017].

By following this model, a RIR $h(t)$ can be described as a sequential combination of responses:

$$h(t) = h_D(t) + h_R(t), \tag{2.37}$$

where $h_D(t)$ and $h_R(t)$ represent the *direct* (direct path plus early reflections) and *reverberant* (late reverberation) components of the

RIR. respectively.

The room impulse response is a function of both the source and the receiver locations. Different levels, delays and directions of direct path and early reflections can obtained from measurements in the same room. However, it is generally assumed that the late reverberation is fixed for a given room, regardless of source/receiver positions.

Room reverberation plays an important role in psychoacoustics. While early reflections are usually perceived together with the direct path as a single auditory event, due to the *precedence effect* [Haas, 1972], late reverberation has often an influence on the received signal. In the specific case of speech, late reverberation is associated with a loss of intelligibility [Braun, 2018]. In the context of spatial perception, it has been shown that early reflections help the localization and externalization of sources [Rudrich and Frank, 2019], while the late reverberation is associated with a spaciousness perception of the room [Begault and Trejo, 2000].

There are a number of measurable parameters which help to characterise room acoustics. Perhaps one of the most widespread is the *reverberation time* T_{60} [Kuttruff, 2016]. It represents the time required for the reverberant sound field power to decay by 60 dB. Reverberation time can be accurately computed from the room geometry [Sabine, 1927] or from the IR [Schroeder, 1965].

In the latter case, the T_{60} value is usually estimated from the *Energy Decay Curve* (EDC), which is defined as:

$$\mathrm{EDC}(t) = 10 \log_{10} \sum_{t'=t}^{\infty} h^2(t), \tag{2.38}$$

where $h(t)$ represents the room impulse response. The values are normalized such that the maximum peak of the curve corresponds to 0 dB.

Table 2.3: Reverberation time computation: usual reference levels

	EDT	T_{10}	T_{20}	T_{30}
$L_{max}(dB)$	0	-5	-5	-5
$L_{min}(dB)$	-10	-15	-25	-35

The EDC is usually modelled as a straight line in logarithmic scale. Therefore, the T_{60} estimation is performed by estimating the slope of a straight line between two reference levels on the EDC time series. Some of the most used reference levels receive specific names: *Early Decay Time* (EDT), T_{60}, and reverberation times T_{10}, T_{20} and T_{30}. Table2.3 shows their correspondent reference levels, where the maximum energy peak is normalized to 0 dB. An schematic representation of the reference levels is depicted in Figure 2.6.

An alternative parameter is the *decay rate* α_{60}, which is related to reverberation time T_{60} as:

$$\alpha_{60} = \frac{3\ln(10)}{T_{60}}(\text{dB/s}). \tag{2.39}$$

The decay rate is thus the slope of the EDC curve, in logarithmic scale, expressed in dB per second.

To conclude, it is important to notice that reverberation time is frequency-dependent. Accordingly, it is usual to report it for octave or third-octave bands, or alternatively to provide its value at a specific frequency.

The *Direct to Reverberant Ratio* (DRR) is another relevant acoustic parameter. DDR represents the ratio between direct and reverberant parts of the RIR, as defined in Eq. 2.37:

$$DRR = 10\log_{10}\frac{\sum_{t=1}^{L_D} h_D^2(t)}{\sum_{t=1}^{L_R} h_R^2(t)}, \tag{2.40}$$

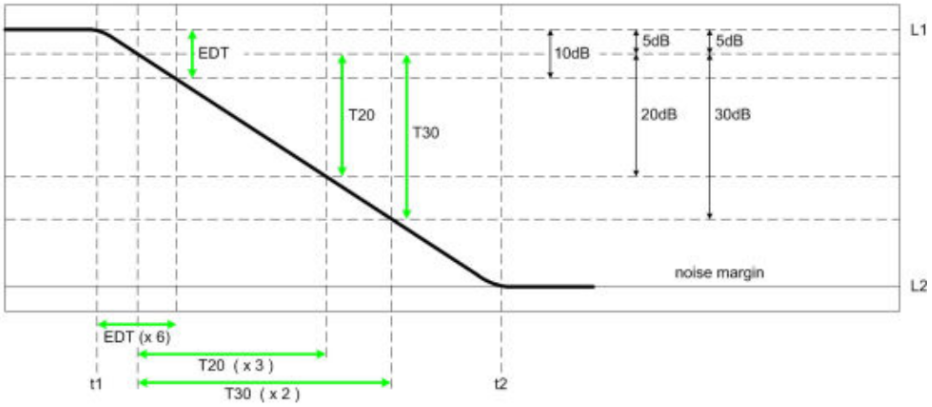


Figure 2.6: Room impulse response model, adapted from http://www.bnoack.com/.

with L_D and L_R as the length of the direct $h_D(t)$ and reverberant $h_R(t)$ filters, respectively. At a psychoacoustic level, the direct to reverberant ratio is one of the main cues for distance perception [Begault and Trejo, 2000].

Since the direct path and early reflections (but not the late reverberation) depend on the relative position between source and receiver , the filter $h_D(t)$ and therefore the DRR are as well location-dependent. For a given room, the source-receiver distance that produces a DRR of 0 dB is known as the *critical distance*.

2.6.1 SOFA Conventions

The availability of recorded room impulse responses is of great importance to many acoustic signal processing problems. As we have seen, many different RIRs can be obtained from the same room, by just varying the position of the source and the receiver; when the number of source and receiver positions increases, the total amount of measurements increases geometrically. Besides

that, the actual format and organisation of the produced data (not only the RIR themselves, but also the source/position annotations) can be arbitrarily different when produced by different groups of people.

In order to overcome potential interoperatibility and reusability issues, the *Spatially Oriented Format for Acoustics* (SOFA) convention [Majdak et al., 2013], also known as the *AES-69* standard [Majdak and Noisternig, 2015], proposes a unified file format for the storage of IR-related data. Despite that SOFA was initially created with a focus on *Head-Related Impulse Response* (HRIR) data, its structure is very convenient to any kind of multi-location impulse responses, including ambisonics.

2.7 Signal Models

Let us consider a sound source represented by the signal $s(t)$, located in a given acoustic enclosure characterised by its room impulse response $h(t)$. The resulting reverberant signal $x(t)$ can be therefore described as the *convolutive mixture* of the source and the RIR:

$$x(t) = s(t) * h(t). \tag{2.41}$$

When dealing with multichannel room impulse responses, as it is the case in ambisonics, the multichannel reverberant signal $x_m(t)$ is obtained by the convolutive mixture of each RIR channel independently:

$$x_m(t) = s(t) * h_m(t). \tag{2.42}$$

The time domain convolution operation, under certain assumptions, is equivalent to the multiplication in frequency domain. By doing so, Eq. 2.41 can be expressed as:

$$X(k,n) = S(k,n)H(k,n). \tag{2.43}$$

Eq. 2.43, also known as the *Multiplicative Transfer Function (MTF) model* is only valid when the length of the filter $h(t)$ is smaller than the length of the analysis window used in the STFT.

On the contrary, when the filter $h(t)$ spans across several analysis windows, the resulting model is referred to as the *Convolutive Transfer Function (CTF) model*:

$$X(k,n) = \sum_{l=0}^{L_h-1} H(k,l)S(k,n-l), \tag{2.44}$$

where L_h is the length of the filter $H(k,n)$ in time frames.

isotropic noise

Chapter 3

Blind reverberation time estimation

3.1 Introduction

Knowledge about the acoustic properties of an enclosure is a fundamental topic with many applications in the microphone array and acoustic signal processing field. Problems such as dereverberation [Braun et al., 2018] or source separation [Gannot et al., 2017] may benefit from this information, and may require prior estimation of the related parameters.

The 2016 Acoustic Characterisation of Environments (ACE) Challenge [Eaton et al., 2016] gathered dozens of methods designed for blind T_{60} and Direct-to-Reverberation Ratio (DRR) estimation; nowadays, it is still considered as a state-of-the-art source for performance evaluation and comparison among methods.

Most of the model-based T_{60} estimation algorithms consider the reverberant signal envelope as an exponential decay, so that the problem is reduced to finding a signal offset and estimate the decay rate. Moreover, in last years, data-driven models have outperformed the previous state-of-the-art results [Gamper and Tashev, 2018, Looney and Gaubitch, 2020, Bryan, 2020]. A compara-

tive review on single-channel blind T_{60} estimation algorithms was recently published [Löllmann et al., 2019].

However, most of the existing reverberation time estimation methods focus on the single-channel case. A representative example can be drawn from the ACE Challenge, where, despite the fact that one of the reverberant datasets was recorded with an *em32 Eigenmike* spherical microphone array, none of the methods use of it for the T_{60} estimation task.

On the other hand, recent years have witnessed a growing interest in immersive audio for virtual and augmented reality. This situation has consolidated Ambisonics [Zotter and Frank, 2019] as the *de facto* standard for spatial audio. Dedicated spherical microphone arrays have reached the market in last years; their multichannel nature makes possible spatial manipulations that complement traditional signal enhancement methods.

In this chapter, we present a novel approach to the problem of multichannel blind reverberation time estimation, specifically focusing on first order ambisonic (FOA) recordings. The method is based on a dereverberation stage followed by system identification. To the best of our knowledge, the proposed algorithm is the first reverberation time estimation method specifically designed for first order ambisonic audio.

The rest of the chapter is organized as follows. Section 3.2 introduces the nomenclature and the signal model. Sections 3.3 and 3.4 describe the baseline and the proposed methods, respectively. The experimental setup is described in Section 3.5, and the results are discussed in Section 5.4. Finally, a conclusion is presented in Section 5.5.

3.2 Signal Model

Let us consider a FOA signal $x_n^m(t)$, with $M = 4$ as the number of channels. Let us further assume the convolutive mixture signal

model described in Eq. 2.42, where the reverberant signal $x_n^m(t)$ represents the signal captured by an ideal spherical microphone array located in a reverberant enclosure. Let $s(t)$ denote the signal of the only sound source present in the scene, and $h_n^m(t)$ denote the ambisonic RIR modelling the acoustic enclosure:

$$x_n^m(t) = s(t) * h_n^m(t) \tag{3.1}$$

It is important to remark that T_{60} estimation here assumes no receiver directionality. In an ambisonic context, this corresponds to the zeroth order component. Therefore, in what follows, all methods estimating IR parameters will be applied to the zeroth order channel, $x_0(t)$.

3.3 Baseline method

The baseline algorithm, taken from [Prego et al., 2012], is based on the detection of abrupt event offsets in the time-frequency domain. The subband energy decay on the transitions can be then used to compute an estimate of the full-band decay. This method performed best in the ACE Challenge regarding the Pearson correlation coefficient between estimated and true T_{60} [Eaton et al., 2016].

Let us consider the zeroth order channel of the recorded signal, $x_0(t)$, and its Short-Time Frequency Transform (STFT) counterpart $X_0(k, n)$. The *subband energy* $\bar{E}(k, n)$ of the recorded signal can be expressed as:

$$\bar{E}(k, n) = |X_0(k, n)|^2. \tag{3.2}$$

A *Free Decay Region* (FDR) is defined as a group of consecutive bins within the same subband which exhibit a monotonically decreasing energy. A FDR search is performed on the subband energy spectrogram $\bar{E}(k, n)$: for each band, the algorithm tries to find at least one FDR, iterartively reducing the FDR length if no candidates are found.

The next step is the estimation of the reverberation time, which is performed using a subband equivalent of Schroeder's method [Schroeder, 1965]. The *Subband Energy Decay Function* (SEDF) associated with a given FDR is computed as:

$$\bar{c}(k,n) = 10\log_{10}\frac{\sum_{\nu=n}^{L_c-1}\bar{E}(k,\nu)}{\sum_{\nu=0}^{L_c-1}\bar{E}(k,\nu)}\text{dB}, \tag{3.3}$$

where $n = 0\ldots, L_c-1$ spans the length of the FDR. A linear regression is then performed on each SEDF curve: T_{60} is computed as the time required by the resulting line to reach the -60 dB reference.

This procedure yields a T_{60} estimate per FDR. In order to obtain a global estimate, the algorithm proposes a two-step statistical filtering. First, it obtains a narrowband estimate as the median of all estimates within each subband. Then, the resulting broadband value $\bar{T}_{60}$ is computed as the median of all subband estimates. The last step of the method is the expansion of the resulting dynamic range by a linear mapping. This procedure is required because of the compression introduced by the median operator. The final value T_{60} is thus a linear mapping of $\bar{T}_{60}$, where the parameters α and β might be obtained by linear regression on a training stage:

$$T_{60} = \alpha\bar{T}_{60} + \beta \tag{3.4}$$

3.4 Proposed method

We propose a novel method for reverberation time estimation, based on two steps: signal dereverberation, and system identification. The main idea consist in obtaining an estimate of the dereverberated signal, which is later used for estimating the multichannel IR given the recorded reverberant signal. The reverberation time can be thus computed by the decay slope of the estimated IR.

3.4.1 Dereverberation

Let us consider now the *CTF* model (Eq. 2.44) version of the proposed signal model:

$$X_m(k,n) = \sum_{l=0}^{L_h-1} H_m(k,l)S(k,n-l), \tag{3.5}$$

where the multichannel filter $H_m(k,l)$ of length L_h contains the *CTF coefficients* between the source and the microphones.

Considering the room impulse response model of Eq. 2.37, it is possible to sequentially split the former expression in the following way:

$$\begin{gathered} X_m(k,n) = D_m(k,n) + R_m(k,n) = \\ = \sum_{l=0}^{\tau-1} H_m(k,l)S(k,n-l) + \sum_{l=\tau}^{L_h-1} H_m(k,l)S(k,n-l), \end{gathered} \tag{3.6}$$

where the parameter τ represents the *mixing time*, which states the transition time between early reflections and late reverberation. In other words, the captured signal is divided between a *direct* part $D_m(k,n)$, containing the direct path and the early reflections, and a *reverberant* part $R_m(k,n)$, which mainly contains the diffuse part of the reverberation.

Assuming a Multichannel Auto-Regressive (MAR) model, $R_m(k,n)$ can be expressed as a multichannel Infinite Impulse Response (IIR) filter applied to the recorded signal:

$$R_m(k,n) = \sum_{i=1}^{M} \sum_{l=0}^{L_g-1} X_i(k,n-\tau-l)G_{mi}(k,l), \tag{3.7}$$

where the coefficients $G_{mi}(k,l) \in \mathbb{C}$ model the relation between channels m and i, and have a length of L_g frames.

By grouping all time frames $n = 1 \ldots, N-1$, it is possible to express Eq. 3.7 in vector notation:

$$\boldsymbol{R}_m(k) = \tilde{\boldsymbol{X}}_\tau(k)\boldsymbol{G}_m(k), \tag{3.8a}$$

$$\tilde{\boldsymbol{X}}_\tau(k) = [\tilde{\boldsymbol{X}}_{\tau,1}(k), \ldots, \tilde{\boldsymbol{X}}_{\tau,M}(k)], \tag{3.8b}$$

where $\tilde{\boldsymbol{X}}_{\tau,m}(k)$ is a $N \times L_g$ matrix, and $\boldsymbol{R}_m(k)$ and $\boldsymbol{G}_m(k)$ are column vectors with lengths N and $L_g M$, respectively.

Finally, the expression can be further simplified by omitting the frequency dependence, and by expressing the channels as columns in the vector notation. Substituting this expression in Eq. 3.6 leads to the MAR equation:

$$\boldsymbol{D} = \boldsymbol{X} - \tilde{\boldsymbol{X}}_\tau \boldsymbol{G}. \tag{3.9}$$

Here, the dereverberation problem consists in the estimation of the MIMO filter $\boldsymbol{G}$, so that the *clean* signal $\boldsymbol{D}$ (containing both direct path and early reflections) can be computed.

The algorithm proposed here is based on the method described in [Jukić et al., 2015]. In this case, the dereverberation problem is tackled as an optimization problem, considering that the spectrograms of the reverberant signal are less sparse than those of the corresponding *clean*, and ensuring that the inter-channel signal properties are mantained. Although the presented method is applied on the whole signal in *batch* mode, alternative *online* methods could be also used, e.g. [Braun and Habets, 2016].

By using *iteratively reweighted least squares* (IRSL) [Chartrand and Yin, 2008], it can be shown that an iterative solution for the estimation of $\boldsymbol{G}$ at the iteration (i) is given by the following expression:

$$\boldsymbol{G}^{(i)} = (\tilde{\boldsymbol{X}}_\tau^H \boldsymbol{W}^{(i)} \tilde{\boldsymbol{X}}_\tau)^{-1} \tilde{\boldsymbol{X}}_\tau^H \boldsymbol{W}^{(i)} \boldsymbol{X}, \tag{3.10}$$

where $\boldsymbol{W}^{(i)}$ is a $N \times N$ diagonal matrix whose diagonal values, $w_n^{(i)}$, can be updated as:

$$w_n^{(i)} = (\boldsymbol{d}_n^{H(i-1)} \boldsymbol{\Phi}^{-1(i-1)} \boldsymbol{d}_n^{(i-1)})^{\frac{p-2}{2}} + \epsilon. \tag{3.11}$$

In turn, $\boldsymbol{d}_n$ represents the rows of $\boldsymbol{D}$ arranged as column vectors of length M, $\boldsymbol{\Phi}$ is the $M \times M$ Spatial Covariance Matrix (SCM) of $\boldsymbol{D}$, ϵ is an arbitrary small positive value, and $p \leq 1$. The computation and update of the SCM matrix is given by:

$$\boldsymbol{\Phi}^{(i)} = \frac{1}{N}\boldsymbol{D}^{T(i)}\boldsymbol{W}^{(i)}\boldsymbol{D}^{*(i)}. \tag{3.12}$$

To conclude the dereverberation method, Eqs. 3.9, 3.10, 3.11 and 3.12 can be applied iteratively, starting by updating Eq. 3.11, until convergence is reached:

$$\|\boldsymbol{D}^{(i)} - \boldsymbol{D}^{(i-1)}\|_F / \|\boldsymbol{D}^{(i)}\|_F < \eta, \tag{3.13}$$

where η is an arbitrary small positive value, or alternatively until the maximum number of iterations i_{max} is exceeded. For the initialization, the following values are proposed: $\boldsymbol{D} = \boldsymbol{X}$ and $\boldsymbol{\Phi} = \boldsymbol{I}_M$ (the identity matrix of size $M \times M$).

3.4.2 System Identification

The output of the dereverberation step is the multichannel signal D_m, which ideally contains the direct plus early reflection components of the source. Therefore, given the reverberant signal X_m and the dereverberated signal D_m, an estimate of the late room impulse response might be derived by identifying the filter connecting the two. As stated in Section 3.2, we are primarily interested on the response of the omnidirectional channel; for that reason, the filter estimation is performed with the zeroth order components of both recorded and dereverberated signals. We perform system identification directly in the STFT through a linear fit between input and output independently for every frequency bin:

$$\hat{H}_0(k) = \frac{\mathbf{d}_0^{\mathrm{H}}(k)\mathbf{x}_0(k)}{\mathbf{d}_0^{\mathrm{H}}(k)\mathbf{d}_0(k)}, \tag{3.14}$$

where $\mathbf{d}_0, \mathbf{x}_0$ are $N \times 1$ length vectors. To avoid complex cross-band modeling of the system response, we use a long STFT window, assumed longer than the twice the length of the IR so that a reduction of the CTF to a Multiplicative Transfer Function (MTF) holds [Avargel and Cohen, 2007].

As a last sep, the estimated time-frequency filter $\hat{H}_0(k, n)$ is transformed into the time domain filter $\hat{h}(t)$. The T_{60} is then computed by linear fitting of the Schroeder integral in the $[-5, -15]$ dB range (T_{10} estimation method), after filtering $\hat{h}(t)$ with an octave-band filter centered at 1 kHz.

3.5 Experimental setup

3.5.1 Dataset

The proposed method is evaluated using two different reverberant datasets, containing recordings of *speech* and *drums* respectively. In order to have full control over the reverberation conditions in the experimental setup, the audio clips under consideration have been rendered by the convolutive mixture of clean monophonic recordings with FOA IRs.

The *speech* dataset is composed of the LibriSpeech [Panayotov et al., 2015] *test-clean* audio samples longer than 25 s, making a total of 30 audio clips. It contains English language sentences by male and female speakers, often with a small level of background noise. We have used only a 20 s long excerpt of each clip, preceded by an initial offset of 5 s. The *drums* dataset is the *test* subset of the isolated drum recordings from the DSD100 dataset [Liutkus et al., 2017]. It contains 50 different audio clips, covering a wide range of music and mixing styles. The same audio lengths and offsets as in the previous case are applied.

The IRs are FOA room impulse responses simulated by the image method with the *Multichannel Acoustic Signal Processing* library [ref to chapter?]. There are 9 different IRs of 1 s, with random T_{60}

values in the range between 0.4 s and 1.1 s approximately, estimated by the T_{10} method at the 1 kHz band. The angular position of the sources is randomized for each IR, while the receiver position is fixed at the room center, which has a size of $10.2 \times 7.1 \times 3.2$ m. The source distance is set to half the *critical distance*, thus providing positive DRRs.

The combination of the dry audio clips with the IRs yields a total of 270 and 450 audio clips for the *speech* and *drums* datasets, respectively, after removing the audio clips which mostly contain silence. Those datasets will be referred in the following as the *evaluation* datasets.

Finally, the baseline method requires a previous *fitting* step for the computation of the mapping parameters α and β from Eq. 3.4. The procedure has been performed as follows. For the *speech* dataset, we selected again the subset of audio clips longer than 25 s, but in this case on the *dev-clean* dataset, which yields a total of 20 audio clips. For the *drums* dataset, we used the 50 clips of the *development* subset. The generation of the convolutive mixes has followed the same procedure as in the previous case. We will refer to the resulting datasets as the *development* datasets.

3.5.2 Setup

The sampling frequency for all methods is 8 kHz. For the baseline system, the window size is 1024 samples long, with an overlap of 256 samples. The FDR length is set to 500 ms, which has been reported as the ideal theoretical minimum [Prego et al., 2012]; it corresponds to a FDR length of $L_c = 15$ samples. At any frequency band, the value of L_c is iteratively decreased if no FDR is found, until a minimum value of 3 samples (96 ms). If still no FDR is found, the sound clip is discarded.

In order to compute α and β, we run the baseline method on both *development* datasets. For each IR, the mean and standard deviation of the results are computed across all sound clips. Then,

Table 3.1: Baseline system: linear regression parameters

Dataset	α	β	σ
Speech	6.6619	-1.4517	0.2131
Drums	8.2421	-2.1939	1.0055

these values are used for a *weighted least squares* linear regression against the true T_{60} values. The results are shown in Table 3.1, where σ represents the joint standard deviation of α and β after the linear regression; the resulting values are in the same range as the values reported in [Prego et al., 2012].

In the dereverberation stage, the STFT uses a small window size of 128 samples, with 64 samples overlap. The value of p is set to 0.25, given the good results reported in [Jukić et al., 2015]. Other parameter values are $\tau = 2$, $i_{max} = 10$, $\eta = 10^{-4}$ and $\epsilon = 10^{-4}$. After an exploratory search, the length of the IIR filter $L_g = 20$ has been chosen as a compromise between method performance and computation time. We have observed a tendency towards poor dereverberation and non-convergence of the IRSL when using small values of L_g and short audios.

For the SID, the recorded and dereverberated signals are reshaped into much larger STFTs, with a window size of 8 s and a hop size of 0.5 s. The predicted filter size is 1 s.

For both *evaluation* datasets, the two presented methods are employed; we will refer to them as *Baseline* and *MAR+SID*. Furthermore, with the aim of evaluating the performance of the SID method in an isolated manner, we have included a third method, *Oracle SID*. As its name suggests, it performs the System Identification step using the true anechoic signal.

Table 3.2: Experiment results

	speech		drums	
Metric	*Baseline*	*MAR+SID*	*Baseline*	*MAR+SID*
Bias	-0.0599	**0.0305**	**0.1521**	0.2568
MSE	0.6366	**0.0594**	**13.9376**	16.5261
ρ	0.8212	**0.9848**	0.3705	**0.7552**

3.5.3 Evaluation metrics

We have considered the three metrics from the ACE Challenge [Eaton et al., 2016], all of them based on the differece between estimated and true values: the *bias*, or mean error; the Mean Squared Error (*MSE*); and the Pearson correlation coefficient. The evaluation has been performed after discarding the outliers, defined as the reverberation time estimates greater than 1.5 s.

3.6 Results

Figure 3.1 shows the experiment result specified for all audio clips individually. Each boxplot represents the statistics of the mean estimation error (*bias*) for a single audio clip subject to all 9 different IRs. The results are organized by method (rows) and dataset (columns). Figure 3.2 aggregates all experiment results into the same plot, showing the statistical distribution of the *bias* per method and dataset. In this case, the *Oracle SID* results are omitted for clarity. The evaluation metrics for all methods are shown in Table 3.2.

According to the results, the proposed method clearly outperforms the baseline in the *speech* dataset by a tenfold MSE improvement. For the *drums* dataset, our method only outperforms the

baseline regarding correlation. Nevertheless, an inspection of the statistical distribution of mean estimation errors in Figure 3.2 brings in an interesting observation: the variability of the results given by our method is substantially smaller than the results of the baseline system. This behaviour is consistent across datasets: the mean error distributions with the *speech* dataset are approximately five times narrower than with the *drums* dataset, regardless of the method.

Moreover, all methods behave significantly better on the *speech* dataset. The main reason might be the heterogeneity of the *drums* dataset dataset with respect to dynamic range or timbre, and the potential application of audio effects of any kind. Furthermore, some audio clips of the *drums* dataset contain sounds with a high degree of self-similarity, such as cymbal rolls or exaggerated reverbs; these characteristics would explain the outliers on the proposed method results. It is also interesting to notice the robustness of the proposed method against noise, present in the *speech* dataset. Such robustness is consistent with the behavior reported in [Jukić et al., 2015].

The performance of the *ORACLE SID* method is close to ideal. The *bias* is in all cases under 0.05 s (excepting a *drums* clip containing mostly silence). This result validates the system identification, and allows, in practical terms, a direct evaluation of the proposed method against the groundtruth values.

The results obtained in our analysis are very similar to the results reported in recent deep-learning state-of-the-art proposals, e.g. [Gamper and Tashev, 2018]. Since all those methods perform single-channel estimation, and our method requieres FOA recordings, the results are not directly comparable. However, given the similar results obtained with the same evaluation metrics, it might be anticipated that out method may perform as well as other recent data-driven algorithms.

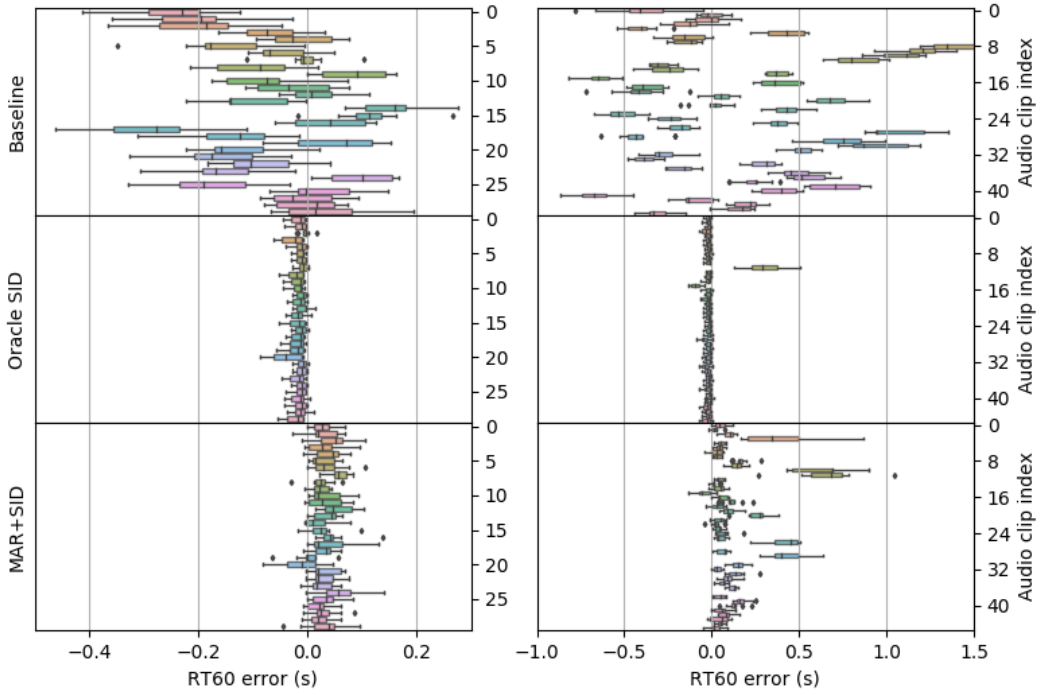


Figure 3.1: Experiment results for *speech* (left column) and *drums* (right column) datasets. Estimation error computed for each audio clip.

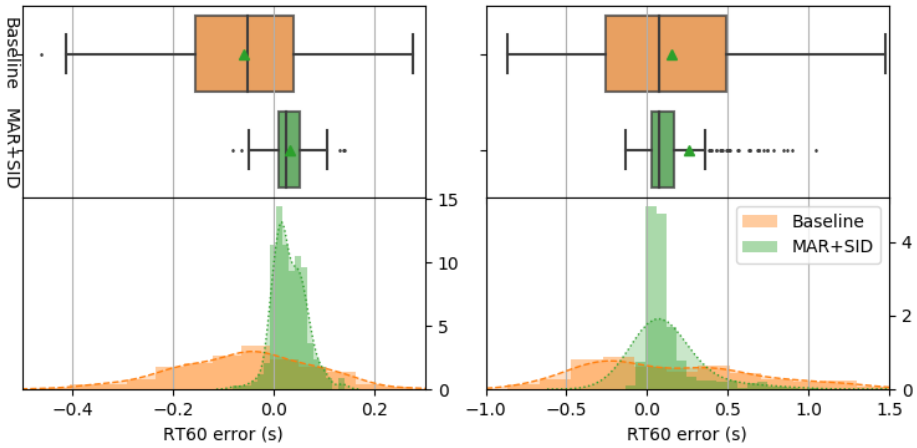


Figure 3.2: Experiment results for *speech* (left column) and *drums* (right column) datasets. Total estimation error across audio clips and acoustic conditions. Top: boxplot. Bottom: histogram and density plot

3.7 Conclusion

We have presented in this work a novel method for blind reverberation time estimation for multichannel audio, with the aim of applying it to the context of ambisonic recordings. Our method is based on a first dereverberation step, performed by a multichannel autoregressive model of the late reverberation. The resulting dry signal is then used to estimate the impulse response decay by means of system identification. The performance of the method is evaluated in a simulated experimental environment with two different reverberant datasets, and compared against a state-of-the-art method. Results show that our method outperforms the baseline method in a majority of evaluation metrics and conditions, and consistently provides results with less variability than the baseline method. In future work, we plan to extend the experimental setup by using recorded IRs. Furthermore, the proposed method could be extended to the case of moving sources by using an *online* autoregressive model. Finally, an extension of the method for higher ambisonic orders remains to be done.

Chapter 4

Coherence Estimation

4.1 Introduction

A number of practical applications benefit of the knowledge about the diffuseness of a sound field, including speech enhancement and dereverberation [P. Habets et al., 2006], noise suppression [Ito et al., 2010], source separation [Duong et al., 2009] or background estimation [Stefanakis and Mouchtaris, 2015]. In the field of spatial audio, diffuseness estimation is often used for parametrization [Pulkki, 2006, Politis et al., 2018a], Direction-of-Arrival estimation [Thiergart et al., 2009] or source separation [Motlicek et al., 2013].

In this chapter, we study diffuseness estimation by subjecting a tetrahedral microphone array to spherically isotropic noise fields. The motivation for this work is, first, that tetrahedral arrays are a well known type of microphone arrays, which have today become popular for applications related to Virtual and Augmented Reality. Second, the spherical isotropic sound field is known to be a good approximation to the reverberant part of the sound field in a room [Elko, 2001, McCowan and Bourlard, 2003], and therefore it would be interesting to investigate how different microphone arrays behave under such conditions.

4.1.1 Problem definition

Under spherical isotropic noise, the theoretical coherence between any pair of zeroth- and first-order ambisonic virtual microphones is equal to 0 for all frequencies, due to the spherical harmonic orthogonality (Eq. 2.7) [Elko, 2001]. This result can also be assessed by Eq. (2.36).

However, there are several practical factors that might corrupt the coherence estimation, such as the approximation of the temporal expectation by time averaging [Thiergart et al., 2011] in Eq. (2.31), or the non-ideal implementation of the radial filters $\Gamma_n(kR)$ (Eq. 2.9) for the *A-B conversion* [Schörkhuber and Höldrich, 2017].

In the following sections, we present several experiments that illustrate the behavior of different coherence estimators applied on the signals captured with a tetrahedral microphone subjected to spherical isotropic noise, using both simulated and real sound recordings.

4.2 Methods

4.2.1 Simulation

Spherical isotropic noise has been generated following the *geometrical method* [Habets and Gannot, 2007, Habets and Gannot, 2010], using $I = 1024$ plane waves. The resulting *A-Format* signals correspond to a virtual tetrahedral microphone array mimicking the Ambeo[1] characteristics ($R = 0.015$ meter, $\alpha = 0.5$). The generated audio has a duration of 60 seconds.

[1]Sennheiser Ambeo VR Mic. https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic

4.2.2 Recording

Spherical isotropic noise has been rendered to a spherical loudspeaker layout with 25 *Genelec 8040*. The loudspeakers are arranged into three azimuth-equidistant 8-speaker rings at inclinations $\vartheta = [\pi/4, \pi/2, 3\pi/4]$, plus one speaker at the zenith. The different speaker distances to the center are delay- and gain-corrected, and the signal feeds are equalized to compensate for speaker coloration. The room has an approximate T_{60} of 300 ms measured at the 1 kHz third-band octave.

The spherical isotropic noise has been also created by the *geometrical method*, encoding a number of uncorrelated noise plane waves in ambisonics with varying orders $N \in [1, 5]$. Due to practical limitations related with the software, the minimum number of sources $I = 256$ for an accurate sound field reconstruction [Habets and Gannot, 2010] could not be reached - instead, the analysis has been performed parametrically with $I = [8, 16, 32, 64]$. For each value of N and I, approximately 15 seconds of audio have been recorded with an Ambeo microphone located at the center of the speaker array.

Ambisonics decoding is performed with an AllRAD decoder, passing through a spherical 64-point 10-design virtual speaker layout, and includes an imaginary speaker at the nadir. The decoding matrix uses *in-phase* weights.

4.2.3 Data processing and metrics

The sampling rate of all signals is 48 kHz. All frequency-domain results have been obtained by averaging their time-frequency representations over time. *A-B conversion* has been computed using *Ambeo A-B converter* AU plugin, version 1.2.1.

Two error metrics are considered: the frequency-dependent squared error $\varepsilon(k)$:

$$\varepsilon(k) = |X_1(k) - X_2(k)|^2, \tag{4.1}$$

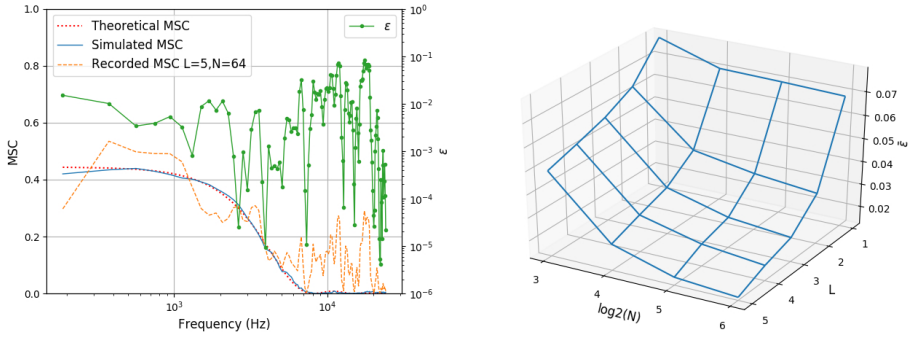


Figure 4.1: *A-Format* coherence between microphone signals. Left: MSC as a function of the frequency of theoretical, simulated and recorded *((BLD,BRU)*, $N = 5, I = 64$) signals. Right: mean error $\bar{\varepsilon}$ of the recorded signals' MSC *(BLD,BRU)* compared to the simulated values, for all values of N and I.REDO FIGURE WITH N AND I

and the mean squared error $\bar{\varepsilon}$:

$$\bar{\varepsilon} = \frac{1}{K}\sum_{k=1}^{K} |X_1(k) - X_2(k)|^2 \tag{4.2}$$

4.3 Results and discussion

4.3.1 A-Format

The coherence of the generated *A-Format* signals is exemplified in Fig. 4.1 (left), which shows the MSC between the capsule pair (*BLD,BRU*) for the theoretical, simulated and recorded cases. The theoretical coherence is derived from Eq. (2.36), while simulated and recorded MSC have been computed by Welch's method, using a *hanning* window of 256 samples and 1/2 overlap. The difference between theoretical and simulated coherence is negligible for practical applications. However, there is a noticeable difference when

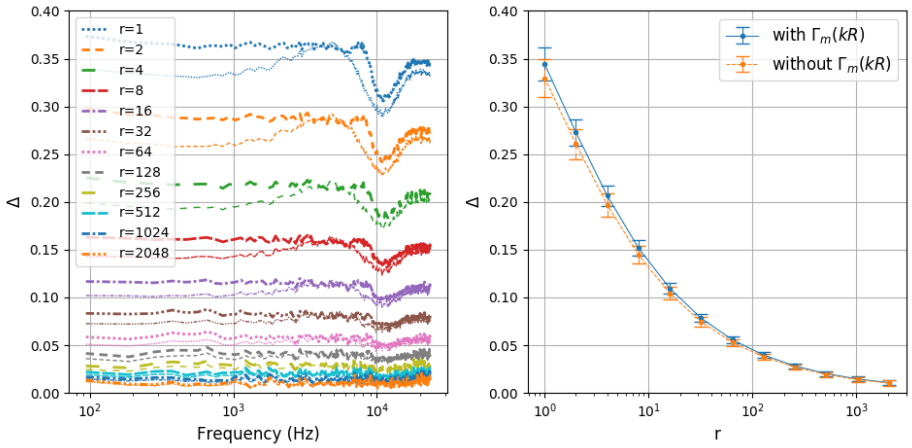


Figure 4.2: Estimated *B-Format* coherence (Δ) of a simulated diffuse sound field, as a function of the temporal averaging vicinity radius r. Left: $\Delta(k)$ for different values of r, with (coarse) and without (fine) application of radial filters. Right: mean and standard deviation of $\Delta(k)$ as a function of r.

compared to the recorded coherence. In general, the recorded MSC follows the tendency of the simulated curve up to around 5 kHz. Above this frequency, the recorded MSC presents several spectral peaks, which might be partially explained by the interference of the microphone itself in the recorded sound field, and by the non-ideal directivity of the capsules. The squared error $\varepsilon(k)$ with respect to the simulated curve is shown in Fig. 4.1 (left), while Fig. 4.1 (right) represents the same error averaged over frequency $\bar{\varepsilon}$ for different spatial resolution values of the diffuse field reproduction algorithm. As expected, $\bar{\varepsilon}$ decreases with increasing values of N and I.

4.3.2 B-Format

In order to evaluate the dependency of the *B-Format coherence* Δ on the number of time frames used for averaging, the following

procedure is presented. The simulated *A-Format* sound field has been transformed into the spherical harmonic domain, with and without the application of radial filters $\Gamma_n(kR)$ (Eq. 2.9). Then, Δ has been computed with Eq. (2.32) for exponentially growing values of r between 1 (8 ms) and 2048 (10.92 s), where r is the vicinity radius used for time averaging, and the number of time windows is given by $T = 2r + 1$. The time-frequency representation is derived by applying the STFT with the same window parameters as in Subsection 4.3.1.

Figure 4.2 (left) shows the great dependence of Δ on r. The estimated coherence tends to the theoretical values with increasing values of r. This tendency is better appreciated in Fig. 4.2 (right): the curve asymptotically decreases to a value $\Delta_{min} \approx 0$.

Another interesting observation comes from the frequency response of the curves. For all values of r, the coherence of the compensated *B-Format* signal (with $\Gamma_m(kR)$) is roughly flat up to around 7 kHz, which approximately corresponds to the operational spatial frequency range of the microphone [Gerzon, 1975]. Above this value, the coherence response looses the flatness due to spatial aliasing (Eq. 2.11). The response above the maximum frequency could be stabilized, if needed, by alternative diffuseness estimation methods [Politis et al., 2015].

The coherence level differences along frequency are inversely proportional to r — the effect is better depicted by the standard deviation values (right). The effect of the radial filters in the coherence measurement is also shown: for a given r, the shape of the coherence is always less flat if no filters are applied. Conversely, in this case, coherence values are always smaller for the same r. This effect might be explained taking into account the inter-channel coherence introduced by microphone and encoder imperfections in real scenarios [Schörkhuber and Höldrich, 2017].

As a remark, the comparison between Figs. 4.1 and 4.2 provides evidence that the application of the spherical harmonic transform

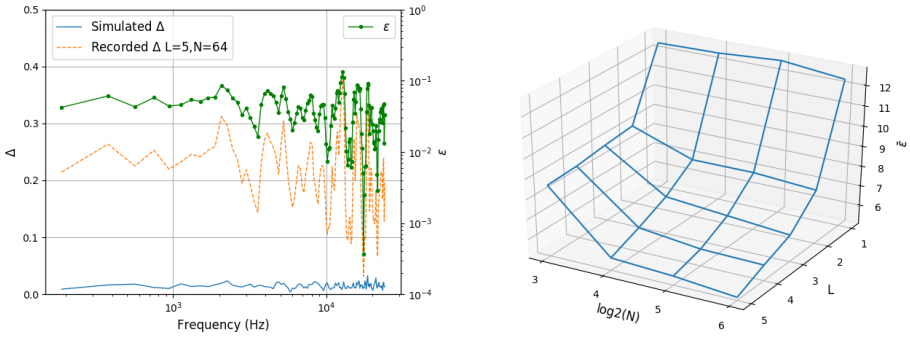


Figure 4.3: *B-Format coherence* between microphone signals. Left: Δ of simulated and recorded ($N = 5, I = 64$) signals. Right: $\bar{\varepsilon}$ of the recorded signals coherence across all values of N and I.REDO FIGURE

might be able to yield more accurate diffuseness estimations, due to a better signal conditioning [**?**].

Figure 4.3 (left) shows the estimated coherence for the recorded sound field with $N = 5$ and $I = 64$, using a vicinity radius of $r = 1024$ (≈ 5 s). The curve is centred around $\Delta = 0.25$ and presents several spectral peaks, as in the *A-Format* case. It is important to notice here that the deviations between the coherence of the simulated and the recorded sound fields are much stronger compared to those of Fig. 4.1.

This effect can be also appreciated in Fig. 4.3 (right): the mean squared error is around two orders of magnitude higher in *B-Format*. Nevertheless, similar as in Fig. 4.1 (right), $\bar{\varepsilon}$ decreases with increasing values of N and I. This behavior suggests that the deviations between the recorded and the simulated coherence can be to a large degree explained by the low spatial resolution of the reproduction system; given a higher number of loudspeakers, we expect that the reproduced diffuseness will tend to the theoretical expression.

4.4 Conclusions

The diffuseness of a sound field is an important parameter for several applications. In this work, two different metrics of diffuseness have been defined and measured with a tetrahedral microphone subjected to spherical isotropic noise.

The analysis shows, first, the impact of the time-averaging window length on the *B-Format* diffuseness estimator. This result might be useful for designing coherence estimators that are parametrized with respect to the length of the analysis window [Thiergart et al., 2011].

Second, the feasibility of diffuse sound field reproduction by a spherical loudspeaker array using ambisonics plane-wave encoding and the *geometrical method* is studied. Results suggest that this approach is viable, given a sufficient spatial resolution; a quantification of the impact of the number of loudspeakers remains for future work.

Chapter 5

Sound Event Localization and Detection

5.1 Introduction

Sound Event Localization and Detection (SELD) refers to the problem of identifying, for each individual event present in a sound field, the temporal activity, spatial location, and sound class to which it belongs. SELD is a current research topic which deals with microphone array processing and sound classification, with potential applications in the fields of signal enhancement, autonomous navigation, acoustic scene description or surveillance, among others.

SELD arises from the combination of two different problems: Sound Event Detection (SED) and Direction of Arrival (DOA) estimation. The number of works in the literature which jointly address SED and DOA problems is relatively small. It is possible to classify them by the type of microphone arrays used: distributed [Grobler et al., 2017, Butko et al., 2011, Chakraborty and Nadeu, 2014] or near-coincident [Hirvonen, 2015, Lopatka et al., 2016, Adavanne et al., 2018]. As mentioned in [Adavanne et al., 2018], the usage of near-coincident circular/spherical arrays enables the represen-

tation of the sound field in the spatial domain, using the spherical harmonic decomposition, also known as Ambisonics [Gerzon, 1973, Daniel, 2000]. Such spatial representation allows a flexible, device-independent comparison between methods. Furthermore, the number of commercially available ambisonic microphones has increased in recent years due to their suitability for immersive multimedia applications. Taking advantage of the compact spatial representation provided by the spherical harmonic decomposition, several methods for parametric analysis of the sound field in the ambisonic domain have been proposed [Pulkki, 2006, Berge and Barrett, 2010, Politis et al., 2018b, Pulkki et al., 2018]. These methods ease sound field segmentation into direct and diffuse components, and further localization of the direct sounds. The advent of deep learning techniques for DOA estimation has also improved the results of traditional methods [Adavanne et al., 2018]. However, none of the deep learning-based DOA estimation methods explicitly exploits the spatial parametric analysis. This situation is further extended to the SELD problem, with the exception of [Lopatka et al., 2016], where DOAs are estimated from the *active intensity vector* [Pulkki, 2006].

The motivation for the proposed methodology is two-fold. First, we would like to check whether the usage of spatial parametric analysis in the ambisonic domain can improve the performance of SELD algorithms. Second, temporal information derived by the parametric analysis could be further exploited to estimate event onsets and offsets, thus lightening the event classifier complexity; such reduction might positively impact algorithm's performance.

In what follows, we present the methodology and the architecture of the proposed system (Section 5.2). Then, we describe the design choices and the experimental setup (Section 5.3), and discuss the results in the context of DCASE2019 Challenge - Task 3 (Section 5.4). A summary is presented in Section 5.5.

5.2 Method

The proposed method presents a solution for the SELD problem splitting the task into four different problems: *DOA estimation*, *association*, *beamforming* and *classification*, which will be described in the following subsections. The former three systems follow a heuristic approach—in what follows, they will be jointly referred to as the *parametric front-end*. Conversely, the *classification* system is data-driven, and will be referred to as the *deep learning back-end*. The method architecture is depicted in Figure 5.1.

redo figures in this section

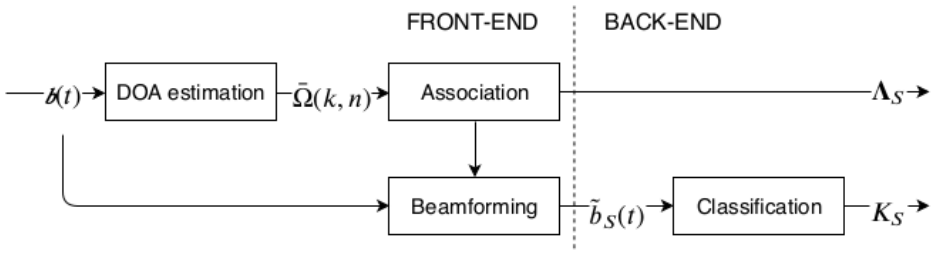


Figure 5.1: System architecture.

5.2.1 DOA estimation

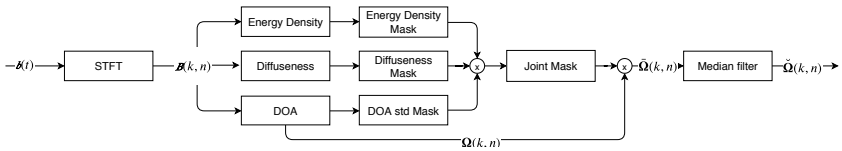


Figure 5.2: DOA estimation architecture.

The *DOA estimation* system (Figure 5.2) is based on parametric time-frequency spatial audio analysis.

Let us consider a *N3D*-normalized first-order ambisonic signal $x_n^m(t)$. Using the DirAC parametric model, it is possible to obtain

instantaneous time-frequency estimates of the Direction of Arrival $\Omega(k,n)$ (Eq. 2.29), energy density $E(k,n)$ (Eq. 2.30) and the diffuseness $\Psi(k,n)$ (Eq. 2.31).

It is desirable to identify the TF regions of $\Omega(k,n)$ which carry information from the sound events, and discard the rest. Three binary masks are computed with that aim.

The first mask is the *energy density mask*, which is used as an activity detector. A gaussian adaptive thresholding algorithm is applied to $E(k,n)$. This procedure selects TF bins with local maximum energy density, as expected from the direct path of the sources.

The *diffuseness mask* selects the TF bins with high energy propagation. Bins with a low diffuseness value represent a TF region where one sound source is predominant. A fixed threshold value Ψ_{max} is used for the masking.

The third mask is the *DOA variance mask*. It tries to select TF regions with small standard deviation with respect to their neighbor bins—a characteristic of sound fields with low diffuseness [Pulkki et al., 2018].

The three masks are then applied to the DOA estimation, obtaining the TF-filtered instanteneous DOAs $\bar{\Omega}(k,n)$. Finally, a median filter is applied, with the aim of improving DOA estimation consistency and removing spurious TF bins. The median filter is applied in a TF bin belonging to $\bar{\Omega}(k,n)$ only if the number of TF bins belonging to $\bar{Omega}(k,n)$ in its vicinity is greater than a given threshold B_{min}. The resulting filtered DOA estimation is referred to as $\breve{\Omega}(k,n)$.

5.2.2 Association

The association step (Figure 5.3) tackles the problem of assigning the time-frequency-space observation $\breve{\Omega}(k,n)$ to a set of events, each one having a specific onset, offset and location.

Figure 5.3: Association architecture.

First, DOA estimates are resampled into *frames* with the task's required length (0.02 s). In what follows, frames will be represented by index m. An additional constraint is applied: for a given window n_0, the DOA estimates $\breve{\Omega}(k, n_0)$ are assigned to the corresponding frame m_0 only if the number of estimates is higher than a given threshold K_{min}.

Next, the standard deviation in azimuth σ_φ and elevation σ_θ of the frame-based DOA estimates $\breve{\Omega}(k, m)$ are compared to a threshold value σ_{max}, and the result is used to estimate the frame-based event overlapping amount $o(m)$:

$$o(m) = \begin{cases} 1, & \text{if } \sigma_\varphi/2 + \sigma_\theta < \sigma_{max}, \\ 2, & \text{otherwise.} \end{cases} \tag{5.1}$$

The clustered values $\Omega_{\text{cluster}}(m)$ are then computed as the $K = o(m)$ centroids of $\breve{\Omega}(k, m)$, using a modified version of K-Means which minimizes the central angle distance. Notice that, for $o(m) = 1$, the operation is equivalent to the median.

The following step is the grouping of clustered DOA values into events. Let us define $\Omega_S(m)$ as the frame-wise DOA estimations belonging to the event S. A given clustered DOA estimation $\Omega_{\text{cluster}}(m)$ belongs to the event S if the following criteria are met:

- The central angle between $\Omega_{\text{cluster}}(m)$ and the median of $\Omega_S(m)$ is smaller than a given threshold d_{max}^{ANGLE}, and
- The frame distance between m and the closest frame of $\Omega_S(m)$ is smaller than a given threshold d_{max}^{FRAME}.

The resulting DOAs $\Omega_S(m)$ are subject to a postprocessing step with the purpose of delaying event onsets in frames where $o(m) > 2$, and discarding events shorter than a given minimum length.

Finally, the frame-based event estimations are converted into *metadata annotations* in the form $\Lambda_S = (\Omega_S, \text{onset}_S, \text{offset}_S)$.

5.2.3 Beamforming

The last step performed in the front-end is the input signal segmentation. The spatial and temporal information provided by the annotations Λ_S are used to produce monophonic signal estimations of the events, $\tilde{x}_S(t)$, as the signals captured by a virtual first-order cardioid (Eq. 2.16):

$$\tilde{x}_S(t) = x_n^m(t) Y_n^m(\Omega_S) \alpha_n, \tag{5.2}$$

with $\alpha_n = [1, 1, 1, 1]$ using the *basic* decoding weights.

5.2.4 Deep learning classification back-end

The parametric front-end performs DOA estimation, temporal activity detection and time/space segmentation, and produces monophonic estimations of the events, $\tilde{s}_S(t)$. Then, the back-end classifies the resulting signals as belonging to one of a target set of 11 classes. Therefore, the multi-task nature of the front-end allows us to define the back-end classification task as a simple multi-class problem, even though the original SELD task is multi-label. It must be noted, however, that due to the limited directivity of the first-order beamformer, the resulting monophonic signals can present a certain leakage from additional sound sources when two events overlap, even when the annotations Λ_S are perfectly estimated.

The classification method is divided into two stages. First, the incoming signal is transformed into the log-mel spectrogram and split into TF patches. Then, the TF patches are fed into a single-mode based on a Convolutional Recurrent Neural Network

(CRNN), which outputs probabilities for event classes $c \in \{1...C\}$, with $C = 11$. Predictions are done at the event-level (not at the frame level), since the temporal activities have been already determined by the front-end. Only one label is predicted for each incoming event, such that there is no binarization stage needed as in a standard SED task.

The proposed CRNN is depicted in Figure 5.4. It presents three convolutional *blocks* to extract local features from the input representation. Each convolutional block consists of one convolutional layer, after which the resulting feature maps are passed through a ReLU non-linearity [Nair and Hinton, 2010]. This is followed by a max-pooling operation to downsample the feature maps and add invariance along the frequency dimension. The target classes vary to a large extent in terms of their temporal dynamics, with some of them being rather impulsive (e.g., *Door slam*), while others being more stationary (e.g., *Phone ringing*). Therefore, after stacking the feature maps resulting from the convolutional blocks, this representation is fed into one bidirectional recurrent layer in order to model discriminative temporal structures. Specifically, 64 nodes of gated recurrent units (GRU) are used with *tanh* activations. The recurrent layer is followed by a Fully Connected (FC) layer, and finally a 11-way softmax classifier layer produces the event-level probabilities. Dropout is applied extensively. The loss function used is categorical cross-entropy. The model has ∼175k weights.

5.3 Experiments

5.3.1 Dataset, evaluation metrics and baseline system

We use the TAU Spatial Sound Events 2019 - Ambisonic, which provides first-order ambisonic recordings. Details about the recording format and dataset specifications can be found in [Adavanne et al.,

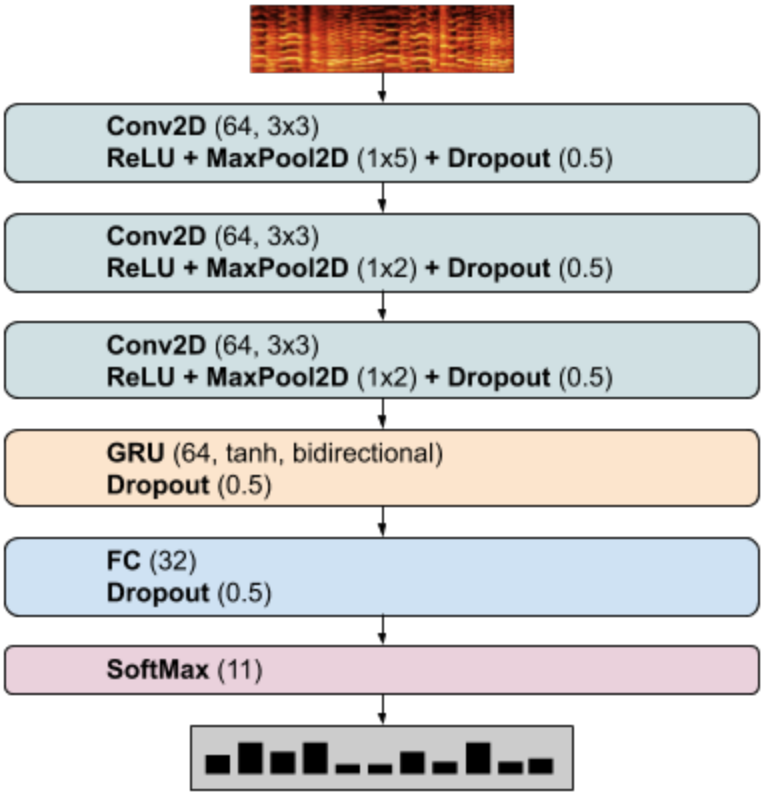


Figure 5.4: Back-end architecture.

2019]. The dataset features a vocabulary of $C = 11$ classes encompassing human sounds and sound events typically found in indoor office environments. The dataset is split into a development and evaluation sets. The development set consists of a four fold cross-validation setup.

The SELD task is evaluated with individual metrics for the SED and DOA problems:

- SED: F-score (F) and error rate (ER) calculated in one-second segments.

- DOA: DOA error (*DOA*) and frame recall (*FR*) calculated frame-wise.

The *SELD score* is an averaged summary of the system performance. A more detailed description of the evaluation metrics can be found in [Adavanne et al., 2018].

The baseline system features a CRNN that jointly performs DOA and SED through multi-task learning [Adavanne et al., 2018]. Baseline results are shown in Table 5.2.

5.3.2 Parametric front-end

Based on the method's exploratory analysis, we propose the following set of parameter values, which are shown in Table 5.1.

5.3.3 Deep learning classification back-end

We use the provided four fold cross-validation setup. Training and validation stages use the outcome of an *ideal* front-end, where the groundtruth DOA estimation and activation times are used to feed the beamformer for time-space segmentation. Conversely, we test the trained models with the signals coming from the *complete* front-end described in Section 5.2.

We conducted a set of preliminary experiments with different types of networks including a VGG-like net, a less deep CNN [Fonseca et al., 2019b], a Mobilenetv1 [Howard et al., 2017] and a CRNN [Cakır et al., 2017]. The latter was found to stand out, and we explore certain facets of the CRNN architecture and the learning pipeline.

Sound events in the dataset last from $\sim$ 0.2 to 3.3 s. First, clips shorter than 2s are replicated to meet this length. Then, we compute TF patches of log-mel spectrograms of $T = 50$ frames (1 s) and $F = 64$ bands. The values come from the exploration of

Table 5.1: Parameter values for the selected configuration. Top: *DOA analysis* parameters. Bottom: *Association* parameters.

Parameter	Unit	Value
sampling rate	Hz	48000
STFT window size	sample	256
STFT window overlap	sample	128
STFT window type	-	Hann
analysis frequency range	Hz	[0,8000]
time average vicinity radius r	bin	10
diffuseness mask threshold Ψ_{max}	-	0.5
energy density filter length	bin	11
std mask vicinity radius	bin	2
std mask normalized threshold	-	0.15
median filter minimum ratio B_{min}	-	0.5
median filter vicinity radius (k,n)	bin	(20, 20)
frame size h	s	0.02
resampling minimum valid bins K_{min}	bin	1
overlapping std threshold σ_{max}	degree	10
grouping maximum angle d_{max}^{ANGLE}	degree	20
grouping maximum distance d_{max}^{FRAME}	frame	20
event minimum length	frame	8

$T \in \{25, 50, 75, 100\}$ and $F \in \{40, 64, 96, 128\}$. $T = 50$ is the top performing value, roughly coinciding with the median event duration. In turn, more than 64 bands provide inconsistent improvements, at the cost of increasing the number of network weights.

Regarding the network structure, several variants of the CRNN architecture were explored until reaching the network of Figure 5.4. This included a small grid search over number of CNN filters, CNN filter size and shape, number of GRU units, number of FC units, dropout [Srivastava et al., 2014], learning rate, and the usage of Batch Normalization (BN) [Ioffe and Szegedy, 2015]. Network extensions (involving more weights) were considered only if providing major improvements, as a measure against overfitting. The main takeaways are: *i)* squared 3x3 filters provide better results than larger filters, *ii)* dropout of 0.5 is critical for overfitting mitigation, *iii)* more than one recurrent layer does not yield improvements, while slowing down training, and *iv)* surprisingly, slightly better performance is attained without BN nor pre-activation [Fonseca et al., 2018].

For all experiments, the batch size was 100 and Adam optimizer was used [Kingma and Ba, 2014] with initial learning rate of 0.001, halved each time the validation accuracy plateaus for 5 epochs. Earlystopping was adopted with a patience of 15 epochs, monitoring validation accuracy. Prediction for every event was obtained by computing predictions at the patch level, and aggregating them with the geometric mean to produce a clip-level prediction.

Finally, we apply *mixup* [Zhang et al., 2017] as data augmentation technique. Mixup consists in creating virtual training examples through linear interpolations in the feature space, assuming that they correspond to linear interpolations in the label space. Essentially, virtual TF patches are created on the fly as convex combinations of the input training patches, with a hyper-parameter α

controlling the interpolation strength. Mixup has been proven successful for sound event classification, even in adverse conditions of corrupted labels [Fonseca et al., 2019a]. It seems appropriate for this task since the front-end outcome can present leakage due to overlapping sources, effectively mixing two sources while only one training label is available, which can be understood as a form of label noise [Fonseca et al., 2019b]. Experiments revealed that mixup with $\alpha = 0.1$ boosted testing accuracy in $\sim 1.5\%$.

5.4 Results and Discussion

Table 5.2: Results for development (top) and evaluation (bottom) sets.

Method	*ER*	*F*	*DOA*	*FR*	***SELD***
Baseline	0.34	79.9%	28.5°	85.4%	0.2113
Proposed	0.32	79.7%	9.1°	76.4%	**0.2026**
Ideal front-end	0.08	93.2%	$\sim 0°$	$\sim 100\%$	0.0379
Baseline	0.28	85.4%	24.6°	85.7%	0.1764
Proposed	0.29	82.1%	9.3°	75.8%	**0.1907**

Table 5.2 shows the results of the proposed method for both development and evaluation sets, compared to the baseline. Focusing on evaluation results, our method and the baseline obtain similar performance in SED (*ER* and *F*). However, there is a clear difference in the DOA metrics: in our method, *DOA* error is reduced by a factor of 2.6, but *FR* is ~ 10 points worst. In terms of *SELD score*, our method performs slightly worse than the baseline in evaluation mode, while marginally outperforming it in development mode.
m,

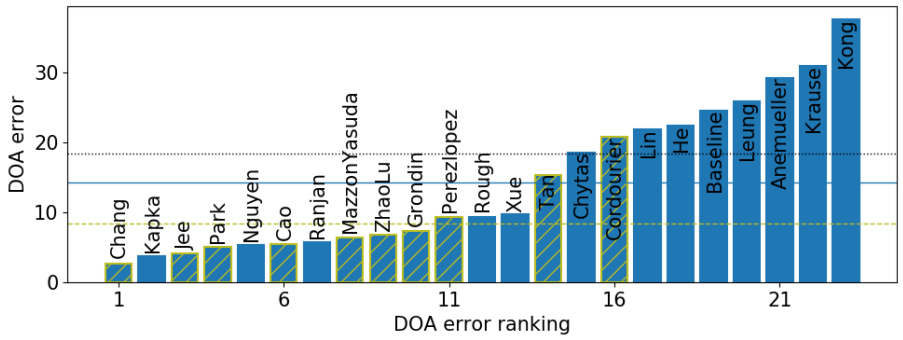
The most relevant observation is the great improvement in *DOA* error. Results suggest that using spatial audio parametric analysis

as a preprocessing step can help to substantially improve localization. Figure 5.5a provides further evidence for this argument: Challenge methods using some kind of parametric preprocessing (*GCC-PHAT* with the microphone dataset, and *Intensity Vector-Based* in ambisonics) obtained in average better *DOA* error results.

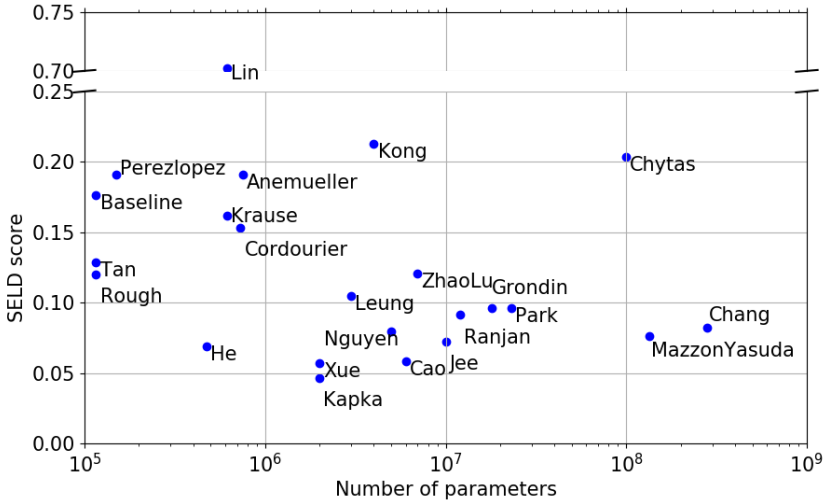
Conversely, the front-end fails regarding *FR*. This is probably due to the complexity added by the association step [Adavanne et al., 2018], and its lack of robustness under highly reverberant scenarios. Including spectral information at the grouping stage might help to improve *FR* — such information could be provided by the classification back-end, in a similar approach to the baseline system. Another option would be the usage of more sophisticated source counting methods [He et al., 2010, Stefanakis et al., 2017].

In order to gain a better insight of the classification back-end performance, Table 5.2 shows the method results when the testing clips are obtained by feeding the beamformer with groundtruth annotations (*ideal* front-end). In this ideal scenario of DOA performance, the SED metrics show a significant boost. This result suggests that the low *FR* given by the front-end has a severe impact on the back-end performance. Yet, the proposed system reaches similar performance to the baseline system in terms of SED metrics.

Finally, we would like to discuss algorithm complexity among Challenge methods. As depicted in Figure 5.5b, there is a general trend towards architectures with very high number of weights, as a consequence of the usage of ensembles and large capacity networks. Specifically, 66% of submitted methods employ 1M weights or more, 30% employ 10M or more, and 15% employ 100M or more. Such complexities are several orders of magnitude greater than the baseline (150k weights) or the proposed method ($\sim$175k weights). In this context, our method represents a low-complexity solution to the SELD problem, featuring a number of parameters and a



(a) *DOA error* across submissions. Hatched bars denote methods using parametric preprocessing. Horizontal lines depict average DOA error across different subsets: all methods (solid), parametric methods (dashed), non-parametric methods (dotted).



(b) *SELD score* versus complexity.

Figure 5.5: DCASE2019 Challenge Task 3 results, evaluation set.

performance comparable to the baseline method.

5.5 Conclusion

We present a novel approach for the SELD task. Our method relies on spatial parametric analysis for the computation of event DOAs, onsets and offsets. This information is used to filter the input signals in time and space, and the resulting event estimations are fed into a CRNN which predicts the class to which the events belong; the classification problem is thereby handled from a simple multi-class perspective. The proposed method is able to obtain an overall performance comparable to the baseline system. The localization accuracy achieved by our method greatly improves the baseline performance, suggesting that spatial parametric analysis might enhance performance of SELD algorithms. Moreover, detection and classification performance in our method suffers from a low Frame Recall; improving this metric could lead to promising SELD scores.

Chapter 6

Data generation and storage

6.1 Ambiscaper

6.2 Ambisonic SOFA convention, pysofaconventions

6.3 masp library

Chapter 7

Conclusions

7.1 Summary of Contributions

is that needed here?

- Academic Contributions

  1. Blind reverberation time estimation from ambisonic recordings ref
     - **Parameter estimation**: Novel technique for blind RT60 estimation of ambisonic recordings from autoregressive models
  2. Analysis of spherical isotropic noise fields with an A-Format tetrahedral microphone ref
     - **Parameter estimation**: Contribution to the characterization of coherence with tetrahedral microphones (the most common spherical arrangement)
  3. A hybrid parametric-deep learning approach for sound event localization and detection ref
     - **Scene Description**: Novel state-of-the-Art methodology for Sound Event Localization and Detection

- Software Contributions

    1. Ambiscaper: A Tool for Automatic Generation and Annotation of Reverberant Ambisonics Sound Scenes ref
        - **Data Generation**: Novel tool for reverberant ambisonic dataset generation
    2. Ambisonics Directional Room Impulse Response as a New Convention of the Spatially Oriented Format for Acoustics ref
        - **Recorded IRs**: File standard/convention proposal for storage of recorded ambisonic IRs
    3. A Python library for Multichannel Acoustic Signal Processing ref
        - **Simulated IRs**: Library for acoustic simulation (IR generation, microphone array simulation, etc)
    4. pysofaconventions, a Python API for SOFA ref
        - **Recorded IRs**: implementation of SOFA for python

7.2 Conclusion

7.3 Future work

Bibliography

[Adavanne et al., 2018] Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2018). Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–1.

[Adavanne et al., 2019] Adavanne, S., Politis, A., and Virtanen, T. (2019). A multi-room reverberant dataset for sound event localization and uetection. In *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*.

[Ahonen and Pulkki, 2009] Ahonen, J. and Pulkki, V. (2009). Diffuseness estimation using temporal variation of intensity vectors. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 285–288. IEEE.

[Avargel and Cohen, 2007] Avargel, Y. and Cohen, I. (2007). On multiplicative transfer function approximation in the short-time fourier transform domain. *IEEE Signal Processing Letters*, 14(5):337–340.

[Begault and Trejo, 2000] Begault, D. R. and Trejo, L. J. (2000). 3-d sound for virtual reality and multimedia.

[Berge and Barrett, 2010] Berge, S. and Barrett, N. (2010). High angular resolution planewave expansion. In *Proc. of the 2nd In-*

ternational Symposium on Ambisonics and Spherical Acoustics May, pages 6–7.

[Braun, 2018] Braun, S. (2018). *Speech dereverberation in noisy environments using time-frequency domain signal models*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

[Braun and Habets, 2016] Braun, S. and Habets, E. A. (2016). Online dereverberation for dynamic scenarios using a kalman filter with an autoregressive model. *IEEE Signal Processing Letters*, 23(12):1741–1745.

[Braun et al., 2018] Braun, S., Kuklasiński, A., Schwartz, O., Thiergart, O., Habets, E. A., Gannot, S., Doclo, S., and Jensen, J. (2018). Evaluation and comparison of late reverberation power spectral density estimators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(6):1056–1071.

[Bryan, 2020] Bryan, N. J. (2020). Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Butko et al., 2011] Butko, T., Pla, F. G., Segura, C., Nadeu, C., and Hernando, J. (2011). Two-source acoustic event detection and localization: Online implementation in a smart-room. In *2011 19th European Signal Processing Conference*, pages 1317–1321. IEEE.

[Cakır et al., 2017] Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1291–1303.

[Carpentier, 2017] Carpentier, T. (2017). Ambisonic spatial blur. In *142nd Audio Engineering Societiy Convention*. AES.

[Chakraborty and Nadeu, 2014] Chakraborty, R. and Nadeu, C. (2014). Sound-model-based acoustic source localization using distributed microphone arrays. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 619–623. IEEE.

[Chartrand and Yin, 2008] Chartrand, R. and Yin, W. (2008). Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872. IEEE.

[Daniel, 2000] Daniel, J. (2000). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris VI.

[Duong et al., 2009] Duong, N., Vincent, E., and Gribonval, R. (2009). Under-determined reverberant audio source separation using a full-rank spatial covariance model. *arXiv:0912.0171 [stat]*. arXiv: 0912.0171.

[Eaton et al., 2016] Eaton, J., Gaubitch, N. D., Moore, A. H., and Naylor, P. A. (2016). Estimation of room acoustic parameters: The ace challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693.

[Elko, 2001] Elko, G. W. (2001). Spatial coherence functions for differential microphones in isotropic noise fields. In *Microphone Arrays*, pages 61–85. Springer, New York.

[Epain and Jin, 2016] Epain, N. and Jin, C. T. (2016). Spherical harmonic signal covariance and sound field diffuseness. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1796–1807.

[Fahy and Salmon, 1990] Fahy, F. J. and Salmon, V. (1990). *Sound intensity*. Acoustical Society of America.

[Fonseca et al., 2019a] Fonseca, E., Font, F., and Serra, X. (2019a). Model-agnostic approaches to handling noisy labels when training sound event classifiers. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, US.

[Fonseca et al., 2018] Fonseca, E., Gong, R., and Serra, X. (2018). A simple fusion of deep and shallow learning for acoustic scene classification. In *Proceedings of the 15th Sound & Music Computing Conference (SMC 2018)*, Limassol, Cyprus.

[Fonseca et al., 2019b] Fonseca, E., Plakal, M., Ellis, D. P. W., Font, F., Favory, X., and Serra, X. (2019b). Learning sound event classifiers from web audio with noisy labels. In *Proc. IEEE ICASSP 2019*, Brighton, UK.

[Gamper and Tashev, 2018] Gamper, H. and Tashev, I. J. (2018). Blind reverberation time estimation using a convolutional neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140. IEEE.

[Gannot et al., 2017] Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4):692–730.

[Gerzon, 1973] Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10.

[Gerzon, 1975] Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio Engineering Society Convention 50*. Audio Engineering Society.

[Gerzon, 1985] Gerzon, M. A. (1985). Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871.

[Grobler et al., 2017] Grobler, C., Kruger, C. P., Silva, B. J., and Hancke, G. P. (2017). Sound based localization and identification in industrial environments. In *IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society*, pages 6119–6124. IEEE.

[Haas, 1972] Haas, H. (1972). The influence of a single echo on the audibility of speech. *Journal of the Audio Engineering Society*, 20(2):146–159.

[Habets and Gannot, 2007] Habets, E. A. P. and Gannot, S. (2007). Generating sensor signals in isotropic noise fields. *The Journal of the Acoustical Society of America*, 122(6):3464–3470.

[Habets and Gannot, 2010] Habets, E. A. P. and Gannot, S. (2010). Comments on generating sensor signals in isotropic noise fields.

[He et al., 2010] He, Z., Cichocki, A., Xie, S., and Choi, K. (2010). Detecting the number of clusters in n-way probabilistic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2006–2021.

[Hirvonen, 2015] Hirvonen, T. (2015). Classification of spatial audio location and content using convolutional neural networks. In *Audio Engineering Society Convention 138*. Audio Engineering Society.

[Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456.

[Ito et al., 2010] Ito, N., Ono, N., Vincent, E., and Sagayama, S. (2010). Designing the Wiener post-filter for diffuse noise suppression using imaginary parts of inter-channel cross-spectra. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2818–2821, Dallas, TX, USA. IEEE.

[Jarrett et al., 2017] Jarrett, D. P., Habets, E. A., and Naylor, P. A. (2017). *Theory and applications of spherical microphone array processing*, volume 9. Springer.

[Jukić et al., 2015] Jukić, A., van Waterschoot, T., Gerkmann, T., and Doclo, S. (2015). Group sparsity for mimo speech dereverberation. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *ICLR 2015*.

[Kuttruff, 2016] Kuttruff, H. (2016). *Room acoustics*. Crc Press.

[Liutkus et al., 2017] Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., and Fontecave, J. (2017). The 2016 signal separation evaluation campaign. In Tichavský, P., Babaie-Zadeh, M., Michel, O. J., and Thirion-Moreau, N., editors, *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pages 323–332, Cham. Springer International Publishing.

[Löllmann et al., 2019] Löllmann, H. W., Brendel, A., and Kellermann, W. (2019). Comparative study for single-channel algo-

rithms for blind reverberation time estimation. In *Proc. Intl. Congress on Acoustics (ICA)*.

[Looney and Gaubitch, 2020] Looney, D. and Gaubitch, N. D. (2020). Joint estimation of acoustic parameters from single-microphone speech observations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 431–435. IEEE.

[Lopatka et al., 2016] Lopatka, K., Kotus, J., and Czyzewski, A. (2016). Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations. *Multimedia Tools and Applications*, 75(17):10407–10439.

[Majdak et al., 2013] Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H., et al. (2013). Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society.

[Majdak and Noisternig, 2015] Majdak, P. and Noisternig, M. (2015). Aes69-2015: Aes standard for file exchange-spatial acoustic data file format. *Audio Engineering Society.*

[Malham, 2003] Malham, D. (2003). Higher order ambisonic systems. *Abstracted from "Space in Music-Music in Space", an Mphil thesis by Dave Malham, submitted to the University of York in April*.

[McCowan and Bourlard, 2003] McCowan, I. and Bourlard, H. (2003). Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing*, 11(6):709–716.

[Merimaa and Pulkki, 2005] Merimaa, J. and Pulkki, V. (2005). Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127.

[Moreau et al., 2006] Moreau, S., Daniel, J., and Bertet, S. (2006). 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.

[Motlicek et al., 2013] Motlicek, P., Duffner, S., Korchagin, D., Bourlard, H., Scheffler, C., Odobez, J.-M., Del Galdo, G., Kallinger, M., and Thiergart, O. (2013). Real-Time Audio-Visual Analysis for Multiperson Videoconferencing. *Advances in Multimedia*, 2013:1–21.

[Murphy et al., 2017] Murphy, D., Shelley, S., Foteinou, A., Brereton, J., and Daffern, H. (2017). Acoustic heritage and audio creativity: the creative application of sound in the representation, understanding and experience of past environments.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[P. Habets et al., 2006] P. Habets, E., Gannot, S., and Cohen, I. (2006). Dual-Microphone Speech Dereverberation in a Noisy Environment. In *2006 IEEE International Symposium on Signal Processing and Information Technology*, pages 651–655, Vancouver, BC, Canada. IEEE.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

[Politis et al., 2015] Politis, A., Delikaris-Manias, S., and Pulkki, V. (2015). Direction-of-arrival and diffuseness estimation above spatial aliasing for symmetrical directional microphone arrays. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6–10, South Brisbane, Queensland, Australia. IEEE.

[Politis et al., 2018a] Politis, A., Tervo, S., and Pulkki, V. (2018a). COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6802–6806, Calgary, AB. IEEE.

[Politis et al., 2018b] Politis, A., Tervo, S., and Pulkki, V. (2018b). COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (May):6802–6806.

[Prego et al., 2012] Prego, T. d. M., de Lima, A. A., Netto, S. L., Lee, B., Said, A., Schafer, R. W., and Kalker, T. (2012). A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. *The Journal of the Acoustical Society of America*, 131(4):2811–2816.

[Pulkki, 2006] Pulkki, V. (2006). 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.

[Pulkki, 2007] Pulkki, V. (2007). Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516.

[Pulkki et al., 2018] Pulkki, V., Delikaris-Manias, S., and Politis, A. (2018). *Parametric time-frequency domain spatial audio*. Wiley Online Library.

[Rafaely, 2004] Rafaely, B. (2004). Analysis and design of spherical microphone arrays. *IEEE Transactions on speech and audio processing*, 13(1):135–143.

[Rudrich and Frank, 2019] Rudrich, D. and Frank, M. (2019). Improving externalization in ambisonic binaural decoding. In *DAGA 2019 Fortschritte der Akustik*.

[Sabine, 1927] Sabine, W. C. (1927). *Collected papers on acoustics*. Harvard University Press Cambridge, MA.

[Schörkhuber and Höldrich, 2017] Schörkhuber, C. and Höldrich, R. (2017). Ambisonic microphone encoding with covariance constraint. In *Proceedings of the International Conference on Spatial Audio*, pages 7–10.

[Schroeder, 1965] Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6):1187–1188.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[Stanzial et al., 1996] Stanzial, D., Prodi, N., and Schiffrer, G. (1996). Reactive acoustic intensity for general fields and energy polarization. *The Journal of the Acoustical Society of America*, 99(4):1868–1876.

[Stefanakis and Mouchtaris, 2015] Stefanakis, N. and Mouchtaris, A. (2015). Foreground suppression for capturing and reproduction of crowded acoustic environments. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 51–55, South Brisbane, Queensland, Australia. IEEE.

[Stefanakis et al., 2017] Stefanakis, N., Pavlidi, D., and Mouchtaris, A. (2017). Perpendicular Cross-Spectra Fusion for Sound Source Localization with a Planar Microphone Array. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(9):1517–1531.

[Thiergart et al., 2011] Thiergart, O., Galdo, G. D., and Habets, E. A. P. (2011). Diffuseness estimation with high temporal resolution via spatial coherence between virtual first-order microphones. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 217–220, New Paltz, NY, USA. IEEE.

[Thiergart et al., 2009] Thiergart, O., Schultz-Amling, R., Del Galdo, G., Mahne, D., and Kuech, F. (2009). Localization of sound sources in reverberant environments based on directional audio coding parameters. In *Audio Engineering Society Convention 127*. Audio Engineering Society.

[Zhang et al., 2017] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[Zotter and Frank, 2019] Zotter, F. and Frank, M. (2019). *Ambisonics*. Springer.