Parametric analysis of ambisonic audio

Contributions to methods, applications and data generation

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

Universitat Pompeu Fabra Barcelona

To my cat

Thanks

First of all, I must thank my supervisors, Emilia and Adan, who have made possible the realization of this journey.

Thanks to all my friends from the Audio Team at Eurecat: Umut, Julien, Tim, Niklas, Toni, Gerard, Andrés, John, Stefan, Marc. And also to the people from Image and the Visualization teams. I spent last four years working with the finest of the companies.

Thanks also to all the great people I met at the MTG, including Pritish, Xavier, and specially Edu for the logistic and intellectual help. I would like to give public acknowledgment to Xavier Serra for his successful task of leading and gathering such unique international group of talented people.

Special thanks to Nikos and all the Signal Processing Lab crew for warmly welcoming me in Heraklion, and helping me in many different ways in your lovely place. I hope to come back there soon.

Many thanks to Archontis and Tuomas for the visit opportunity, and for all the facilities they gave me in my stay in this amazing place called Tampere. I take the opportunity to say hello and thanks to Kostas, Natalya, Annamaria, Toni, and all the nice people there.

Thanks to Max, who has been always very receptive and supportive on numerous occasions, and has indirectly contributed to the success of this project. Thanks also to Rafa, who decided to get on board for some unknown reason.

Thanks to all the friends I made and preserved in Barcelona in the last four years. Special thanks to my musician family from Balkumbia, Fanfarrai and Telewawachi Kilili; I learned a lot from you. Special acknowledgements for the good times with the people in Can Collet (best farm-school in town) and the Valencian oldies.

Thanks a lot to Chedar for his company in the final steps. And, of course, million thanks to my family and to Marta for their unconditional love and support, powering all the way up here.

To all of you: thanks.

Abstract

Due to the recent advances in virtual and augmented reality, ambisonics has emerged as the *de facto* standard for immersive audio. Ambisonic audio can be captured using spherical microphone arrays, which are becoming increasingly popular. Yet, many methods for acoustic and microphone array signal processing are not specifically tailored for spherical geometries. Therefore, there is still room for improvement in the field of automatic analysis and description of ambisonic recordings. In the present thesis, we tackle this problem using methods based on the parametric analysis of the sound field. Specifically, we present novel contributions in the scope of blind reverberation time estimation, diffuseness estimation, and sound event localization and detection. Furthermore, several software tools developed for ambisonic dataset generation and management are also presented.

Resum

Ambisonics ha esdevingut l'estàndard d'àudio immersiu als últims anys, afavorit pels avançaments en realitat virtual i augmentada. L'àudio *ambisonic* es pot obtindre mitjançant *arrays* de micròfons esfèrics, que són cada vegada més populars. Tot i això, la majoria de mètodes acústics i de processament de senyal basats en *arrays* de micròfons no estan adaptats al cas específic de geometries esfèriques. Per tant, encara hi ha moltes possibilitats de millora en l'àmbit d'anàlisi automàtica i descripció de gravacions *ambisonic*. En la present tesi plantegem aquest problema basant-nos en l'anàlisi paramètrica del camp acústic. Més concretament, presentem contribucions originals en les àrees d'estimació de reverberació, estimació de difusió acústica, i detecció i localització d'esdeveniments sonors. Així mateix, presentem diverses eines de programari desenvolupades per la generació i manteniment de bases de dades d'àudio *ambisonic*.

Preface

The present thesis has been carried out within the context of the Industrial Doctorate program from the Catalan government, as a collaboration between the Music Technology Group of the Pompeu Fabra University, and the Multimedia Unit of Eurecat, the Catalan Technology Center. It has been jointly supervised by Dr. Emilia Gómez and Dr. Adan Garriga.

The industrial nature of the doctorate program implies a tendency towards the innovative application of the findings and conclusions generated during the development of the thesis. Therefore, one of the targets of the research performed in the context of this thesis is the stress on real-world problems, scenarios and technologies, with the aim of developing methods that could be easily adapted into applications with a high level of technology readiness. Moreover, there has been also an stress on research reproducibility. This has not been motivated only by an academic perspective; but also by the open research and open innovation paradigms present in our today's society.

The work presented here would not have been possible without the contributions from an excellent network of researchers. Apart from internal collaborations (UPF and Eurecat), two of the main contributions in the thesis followed two research visits to internationally recognized institutions: the Audio Research Group from Tampere University, Finland (Chapter 3), and the Signal Processing Group at the FORTH - Foundation for Research and Technology Hellas, Heraklion, Greece, (Chapter 4).

The alphabetical list of academic collaborators is: Eduardo Fonseca, Emilia Gómez, Rafael Ibáñez-Usach, Julien de Muynke, Archontis Politis, Xavier Serra, and Nikolaos Stefanakis.

Contents

List of figures **xvii**

List of tables **xix**

1 Introduction **1**
1.1 Motivation 1
1.2 Problem Description 5
1.3 Scientific Objectives 8
1.4 Outline 8

2 Scientific Background **11**
2.1 Conventions 11
2.1.1 Reference system 11
2.1.2 Nomenclature 13
2.2 Spherical Harmonics 14
2.2.1 Definition 14
2.2.2 Spherical array processing 15
2.3 Ambisonics 18
2.3.1 Ambisonics Theory 18
2.3.2 Practical considerations 23
2.4 Parametric Spatial Audio Analysis 27
2.5 Spatial Coherence Analysis 31
2.6 Reverberation 32
2.7 Signal Models 36

3 Blind reverberation time estimation **39**
3.1 Introduction 39
3.2 Signal Model 41
3.3 Baseline method 41
3.4 Proposed method 42
3.4.1 Dereverberation 43
3.4.2 System Identification 45
3.5 Experimental setup 46
3.5.1 Dataset 46
3.5.2 Setup 48
3.5.3 Evaluation metrics 49
3.6 Results 50
3.7 Conclusion 51

4 Coherence Estimation **55**
4.1 Introduction 55
4.1.1 Problem definition 56
4.2 Methods 56
4.2.1 Simulation 56
4.2.2 Recording 57
4.2.3 Data processing and metrics 57
4.3 Results and discussion 58
4.3.1 A-Format 58
4.3.2 B-Format 59
4.4 Conclusions 62

5 Sound Event Localization and Detection **65**
5.1 Introduction 65
5.2 System description 67
5.2.1 Single-source estimation 67
5.2.2 Particle tracking 68
5.2.3 Signal filter 70
5.2.4 Event classification 71
5.3 Experiments 72

5.3.1 Dataset and baseline system 72
5.3.2 Experimental setup 73
5.3.3 Evaluation metrics 75
5.4 Results 77
5.5 Conclusion 83

6 Data generation and storage **85**
6.1 Introduction 85
6.2 MASP: a Python library for multichannel acoustic signal processing 87
6.2.1 Description 87
6.2.2 Related software 89
6.3 SOFA 92
6.3.1 Problem statement 92
6.3.2 Ambisonics Directional Room Impulse Response as a SOFA convention 93
6.3.3 Pysofaconventions 94
6.4 Ambiscaper 95
6.4.1 Motivation 95
6.4.2 Implementation 97
6.4.3 Experiment reproducibility 99
6.5 Conclusion 100

7 Conclusions **103**
7.1 Summary of Contributions 103
7.2 Future Work 105
7.3 List of Contributions 108
7.4 List of Software Resources 110

List of Figures

1.1 Number of ambisonic microphones released in last years (from [Wikipedia, 2020]). From left to right, the vertical lines correspond to (1) Oculus acquisition, (2) *Time* cover page on VR, (3) M. Zuckerberg's speech in MWC, and (4) Jaunt announcement of shift towards AR. 3
1.2 Venture investments in VR in the period 2014-2018. Adapted from [Fortune, 2019]. 5
1.3 General scheme of the B-Format audio generation and analysis framework. Solid lines represent audio signals, while outlined arrows refer to non-audio information. 6
1.4 General scheme of the B-Format audio generation and analysis framework, including the thesis contributions in form of Chapter numbers. 9

2.1 Spherical coordinate system used. 12
2.2 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. 16
2.3 Magnitude of $\Gamma_n(kR)$ for different ambisonic orders, in the case of (a) rigid sphere, and (b) open sphere configurations. Adapted from [Rafaely, 2004]. . . . 17
2.4 Directive patterns of first-order ambisonic decoding. 21

2.5 Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from [Carpentier, 2017]. 25
2.6 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 $\Psi(k, n)$. . 30
2.7 Room impulse response model, from [Murphy et al., 2017]. 33
2.8 Room impulse response model, adapted from [AV_INFO, 1995]. 36

3.1 Experiment results for *speech* (left column) and *drums* (right column) datasets. Estimation error computed for each audio clip. 52
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 53

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. 59
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. 60

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. 62

5.1 Architecture of the proposed methodology. 67
5.2 Estimation of localization and temporal activation. Top: azimuth spectrogram after diffuseness mask; color indicates estimated position of a TF bin passing the single-source test. Bottom: input/output of the particle tracking; the crosses represent the measurement space, and the continuous lines are the resulting events. 69
5.3 Gradient boosting machine learning process. Adding weak estimators allows reducing overall error in the predictions. 72
5.4 Number of occurrences of each event class in the training set, for both proposed methods. 74
5.5 Most representative features in event classifier. . . . 79
5.6 Event durations of all elements in the *PAPAFIL2* training set. 80
5.7 DCASE 2020 Task 3 submissions: complexity versus SELD score. 83

6.1 2018 projections of future internet traffic for major programming languages. Adapted from [Robinson, 2017]. 86
6.2 Frequency response of an arbitrary spherical array. . 89
6.3 Evaluation of radial filters for an arbitrary spherical microphone array. 90
6.4 *AmbiScaper* architecture. 98

List of Tables

1.1 List of ambisonic microphones released in recent years (from [Wikipedia, 2020]). 4

2.1 Cartesian and spherical representation of characteristic points along the unit sphere. 12
2.2 Ambisonic decoding: standard values of $alpha_n$ weightings. Adapted from [Daniel, 2000]. 22
2.3 Reverberation time computation: usual reference levels 35

3.1 Baseline system: linear regression parameters 48
3.2 Experiment results 49

5.1 Acoustic features used for classification, grouped by type. 72
5.2 (Hyper-)parameter values. 76
5.3 System evaluation. Top: results on the cross-validation development set. Bottom: results on the evaluation set. 77
5.4 DCASE 2020 Challenge Task 3 evaluation results . . 81

6.1 Features of *MASP* compared to *pyroomacoustics*. . . . 91
6.2 Summary of audio data used across Ambisonics-based Source Localization (above) and Source Separation (below) methods. 96

Chapter 1

Introduction

1.1 Motivation

Ambisonics is a spatial audio theory based on the directional decomposition of the sound field. Conceived in its primal form during the 70s [Gerzon, 1973], it was not until the 21st century, with a modern mathematical formulation [Daniel, 2000] and much more computational power available, that it definitely drew the attention of the research community.

Nevertheless, the greatest contributor to the current interest in ambisonics has been the rise of Virtual Reality (VR) in recent years. Although VR focuses primarily on visual cues, the immersive experience can be greatly enhanced by spatial audio [Begault and Trejo, 2000]. In this context, Ambisonics has been rapidly adopted as *de facto* standard for spatial audio transmission, supposedly due to a variety of factors:

Layout independence As opposed to other audio spatialization techniques that rely on specific playback layouts, ambisonics makes use of an intermediate sound field representation, known as *B-Format* (or just ambisonic audio). This representation, often referred to as *scene-based*, can be then further processed to match any playback configuration.

Recording device independence Regardless of the specific characteristics of an ambisonic microphone, the recorded signal is usually converted into B-Format, which is effectively the standard exchange format.

Ease of manipulations Signal-independent transformations of the ambisonic stream, and specifically rotations, are computationally inexpensive.

Binaural transformation Spatial audio in VR is mostly consumed as binaural; methods for ambisonic to binaural conversion have been known for a long time [Noisternig et al., 2003]. Furthermore, VR headsets can easily provide head rotation information, which can be used in combination with scene rotations to provide head-locked audio, which greatly improves localization accuracy and inmersiveness [Begault and Trejo, 2000]. This is a key feature of ambisonics when compared to static binaural recordings.

Coming back to the issue of the popularity of Virtual Reality, we have selected three events that might epitomize the growth undergone in the second half of the 2010s:

1. The billionaire acquisition by Facebook of the VR headset manufacturer Oculus, in March 2014 [Facebook, 2014]

2. *Time* maagazine cover page devoted to VR: *"The surprising joy of Virtual Reality. And why it's about to change the world"* (August 2015) [Time, 2015];

3. M. Zuckerberg's invited talk at the *Samsung Unpacked* event within the World Mobile Congress 2016 in Barcelona: *"VR is the next platform where anyone can experience anything they want"* [BBCNews, 2016].

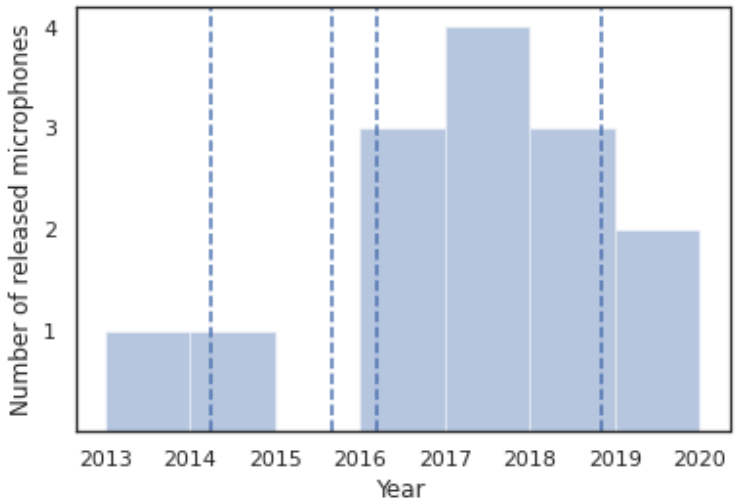

Figure 1.1: Number of ambisonic microphones released in last years (from [Wikipedia, 2020]). From left to right, the vertical lines correspond to (1) Oculus acquisition, (2) *Time* cover page on VR, (3) M. Zuckerberg's speech in MWC, and (4) Jaunt announcement of shift towards AR.

Given this context, many microphone manufacturers and audio-related companies have followed the industry trend in the search for new markets. Only in the interval 2016-2019, 12 different ambisonic microphones have reached the market (Figure 1.1) — a greater amount than all previous existing ambisonic microphones together. A comprehensive list of recent ambisonic microphone releases is shown in 1.1.

At the present time, however, the high expectations put into VR have significantly lowered, as shown in Figure 1.2. This is due to a variety of reasons, including lack of interesting content and the high production cost of the headsets [Fortune, 2019]. The change of focus of Jaunt (formerly one of the biggest VR film production

Table 1.1: List of ambisonic microphones released in recent years (from [Wikipedia, 2020]).

Manufacturer	Model	Year	Order
MH Acoustics	EigenMike	2013	4
Brahma	(Brahma)	2014	1
Sennheiser	Ambeo	2016	1
Twirling	720 VR	2016	1
Zoom	H2n	2016	1
Zylia	ZM-1	2017	3
Twirling	720 Lite	2017	1
Ricoh	TA-1	2017	1
Nevaton	Nevaton VR	2017	1
Rode	Rode NT-SF1	2018	1
CoreSound	OctoMic	2018	2
Zoom	H3-VR	2018	1
Brahma	Brahma 8	2019	2
Voyage Audio	Spatial Mic	2019	2

companies) towards Augmented Reality (AR), as of October 2018, might be a paradigmatic example of this tendency [TheVerge, 2018].

In any case, the current high availability and affordability of ambisonic microphones brings new challenges from the signal processing perspective. More specifically, ambisonic microphones conform a subset of near-coincident spherical microphone arrays, a category which possesses some specific characteristics.

Although the VR momentum has also reached spherical microphone array processing, many challenges remain still open, and the number of research works specifically focusing on such geometry are low yet. Besides that, the growing interest in AR posses new problems related to acoustical signal processing. But since ambisonics is still the standard choice for immersive audio, existing solutions might be successfully adapted.

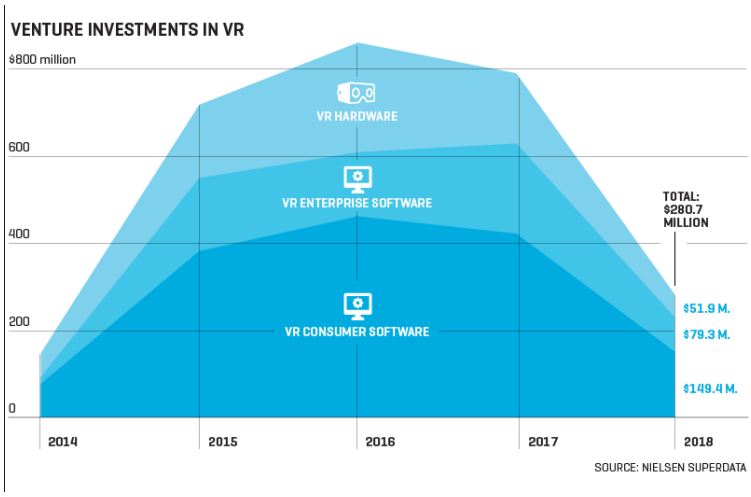

Figure 1.2: Venture investments in VR in the period 2014-2018. Adapted from [Fortune, 2019].

Lastly, the advance in signal processing methods for ambisonics can give rise to applications that enhance the work of immersive audio producers, providing meaningful information about the recorded scenes and automating some of the repetitive tasks, thus allowing a more flexible and creative workflow.

1.2 Problem Description

The scientific context of the work developed in this thesis is shown in Figure 1.3, which has been inspired by [Jarrett et al., 2017]. As we can observe, there are two main topics related with B-Format audio: *generation* and *analysis*, with the signal flow going from the former to the latter. Although the conceptual approach of the scheme might be very similar for any type of audio, the spatial information conveyed by the ambisonic signal places an emphasis on the informed analysis and description of the sound scene.

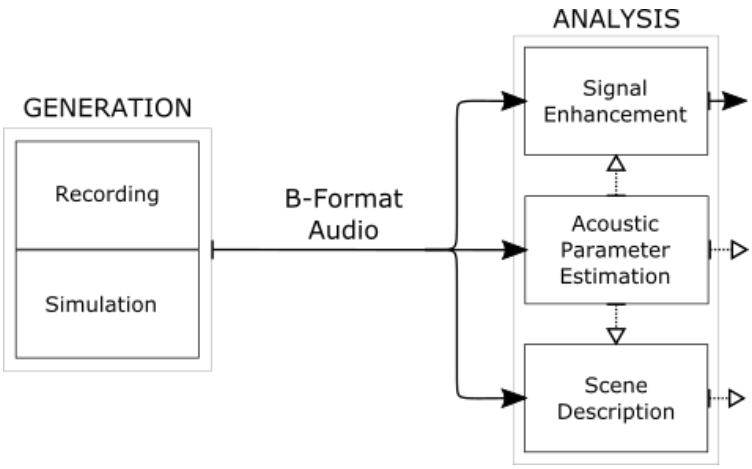

Figure 1.3: General scheme of the B-Format audio generation and analysis framework. Solid lines represent audio signals, while outlined arrows refer to non-audio information.

The problem of B-Format generation is mostly related with dataset generation, which is a common issue for many audio signal processing problems. In our case, we consider two different approaches to the data generation problem:

Recording Using spherical microphone arrays for the recording of ambisonic material.

Simulation Using numerical methods for the simulation of acoustic scenes.

While recordings are by definition more similar to real scenarios, they are expensive to perform, and can only provide a limited set of parameter possibilities. Simulations, on the other hand, have the potential to cover any desired condition. Therefore, it is of our interest to consider the strengths of both signal generation paths.

Ambisonic signal analysis has been divided in three categories:

Signal Enhancement Modification of the input signal in order to obtain one or more output signals with desired attributes.

There are a variety of well known signal enhancement problems, including dereverberation [Braun, 2018], source separation [Gannot et al., 2017], or foreground-background segmentation [Stefanakis and Mouchtaris, 2015]. As it has been shown, many of them benefit (or even depend) from the knowledge derived by acoustic parameter estimation methods.

Acoustic Parameter Estimation Low-level analysis of the sound field, which yields quantitative information about different acoustic parameters used to model the acoustic scene.

The knowledge about the acoustic parameters of a sound field can be considered either as a goal by itself, or alternatively as a preprocessing step which complements the other analysis categories. Examples of typical estimated acoustic parameters are the *Direction-of-Arrival* (DOA), the sound field diffuseness, or the reverberation time of the enclosure [Jarrett et al., 2017].

Scene Description Textual representation of different high-level characteristics of the sound field.

Under the scene description typology we can find a set of applications that provide abstract representations of the sound scene under analysis. Most of the recent research performed in this scope is grouped around the Detection and Classification of Acoustic Scenes and Events (DCASE) community [DCASE, 2013]. Examples of the problems under consideration are acoustic scene classification [Mesaros et al., 2018], sound event localization and detection [Adavanne et al., 2018] or audio captioning [Drossos et al., 2017].

1.3 Scientific Objectives

The list that follows concentrates the main scientific objectives to be developed on this thesis:

1. To develop methods for the characterisation of acoustic parameters from recordings originated from ambisonic microphones.

2. To propose methodologies for sound event localization and detection in ambisonic domain which are grounded on spatial parametric analysis.

3. To contribute to the generation and storage of ambisonic sound scenes, for their usage in controlled experimental environments.

1.4 Outline

The present dissertation is organised as follows.

Chapter 2 introduces the basic concepts that will be developed throughout this 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 this 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 enables computation of the reverberation time from an estimation of the room impulse response. The evaluation metrics show a

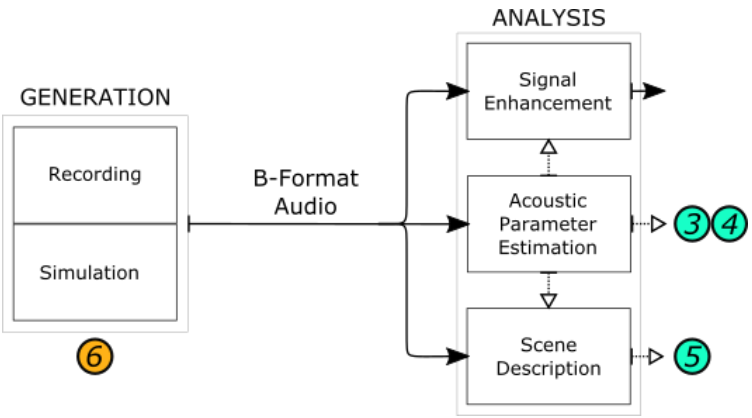

Figure 1.4: General scheme of the B-Format audio generation and analysis framework, including the thesis contributions in form of Chapter numbers.

method performance similar to other state-of-the-art methods. The

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 this 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 industrial nature of this thesis, we have considered relevant to include them in the present dissertation.

In order to provide a schematic representation of the thesis structure and scope, Figure 1.4 features the problem structure stated in Figure 1.3, with the addition of the Chapter numbers with the contributions of this thesis.

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. The spherical coordinate system used in this thesis is depicted in Figure 2.1.

Table 2.1 shows the spherical coordinate values for some reference points on the unit sphere. Notice that the poles ($\vartheta = 0, \pi$) are a special case for the spherical coordinate system – in that case, the azimuth angle is not defined.

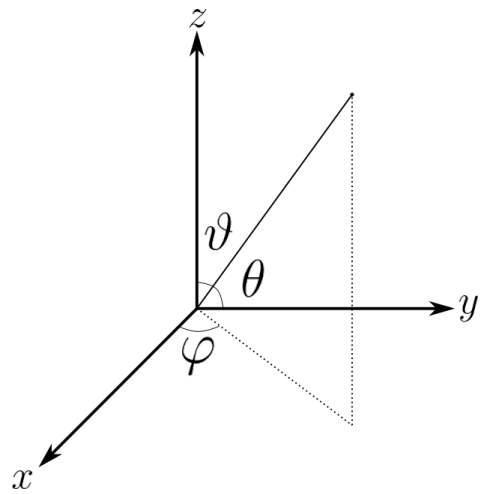

Figure 2.1: Spherical coordinate system used.

The transformation between spherical and cartesian coordinate systems is given by the following relationship:

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

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	*

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 φ will refer to the 2π-*periodic* version or the operator; this situation does not affect the inclination/elevation coordinate.

2.1.2 Nomenclature

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 $k \in [0, K-1]$ is the frequency bin index, and $n \in [0, N-1]$ 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 continuously differentiable 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 2.2 depicts all spherical harmonics from orders 0 to 3.

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]:

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

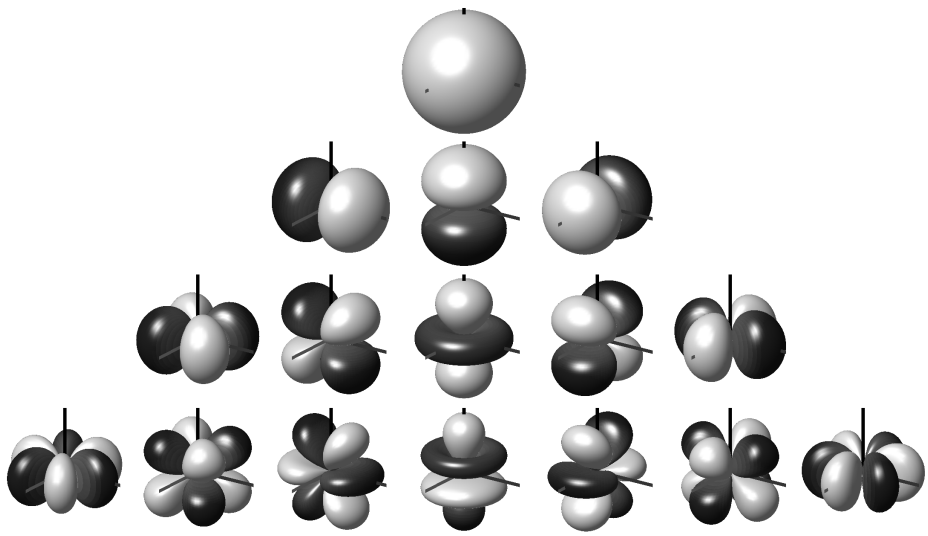

Figure 2.2: Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values.

where the term $\Gamma_n(kR)$ models the radial transfer function, and depends on a number of factors, being the *sphere configuration* one of the most prominent considerations. Sphere configuration, in its basic form, refers to the physical properties of the baffle where the capsules are mounted, and it can be either *open* or *rigid*. While open configuration is the simplest solution, it might present numerical problems in the form of zeros in its frequency response. Conversely, a rigid baffle interferes with the sound field and might create undesired interferences, but it improves the numerical condition from the open case. Fig. 2.3 shows the simulated magnitude response of $\Gamma_n(kR)$ for a spherical array considering both configurations. The reader is referred to [Moreau et al., 2006] and [Rafaely, 2004] for a deeper insight into the topic of spherical microphone array design.

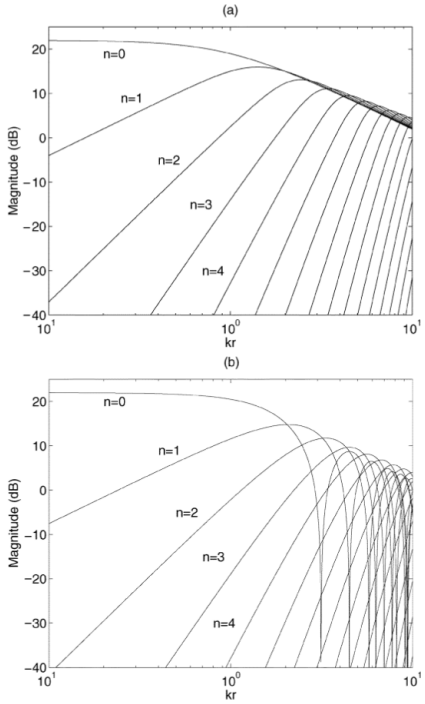

Figure 2.3: Magnitude of $\Gamma_n(kR)$ for different ambisonic orders, in the case of (a) rigid sphere, and (b) open sphere configurations. Adapted from [Rafaely, 2004].

By using the model from Eq. 2.9, 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. According to [Moreau et al., 2006], the maximum aliasing-free operational frequency of a spherical microphone array is given by:

$$f_{max} = \frac{c}{2R\gamma}, \tag{2.11}$$

with c being the sound speed, and γ the maximum aperture angle between two capsules. It is important to notice the existence of a practical minimum frequency of the spherical microphone array, given by the low magnitude in low frequencies of high ambisonic order components, as shown in Fig. 2.3.

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:

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

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:

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

Note that, because of the superposition principle, a sound field composed of several different point sources can be broken down to the addition of the individual contributions.

Although the term *B-Format* was initially introduced as an alternative name for first-order ambisonic signals [Daniel, 2000], 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, 1975a]. 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:

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

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 2.4.

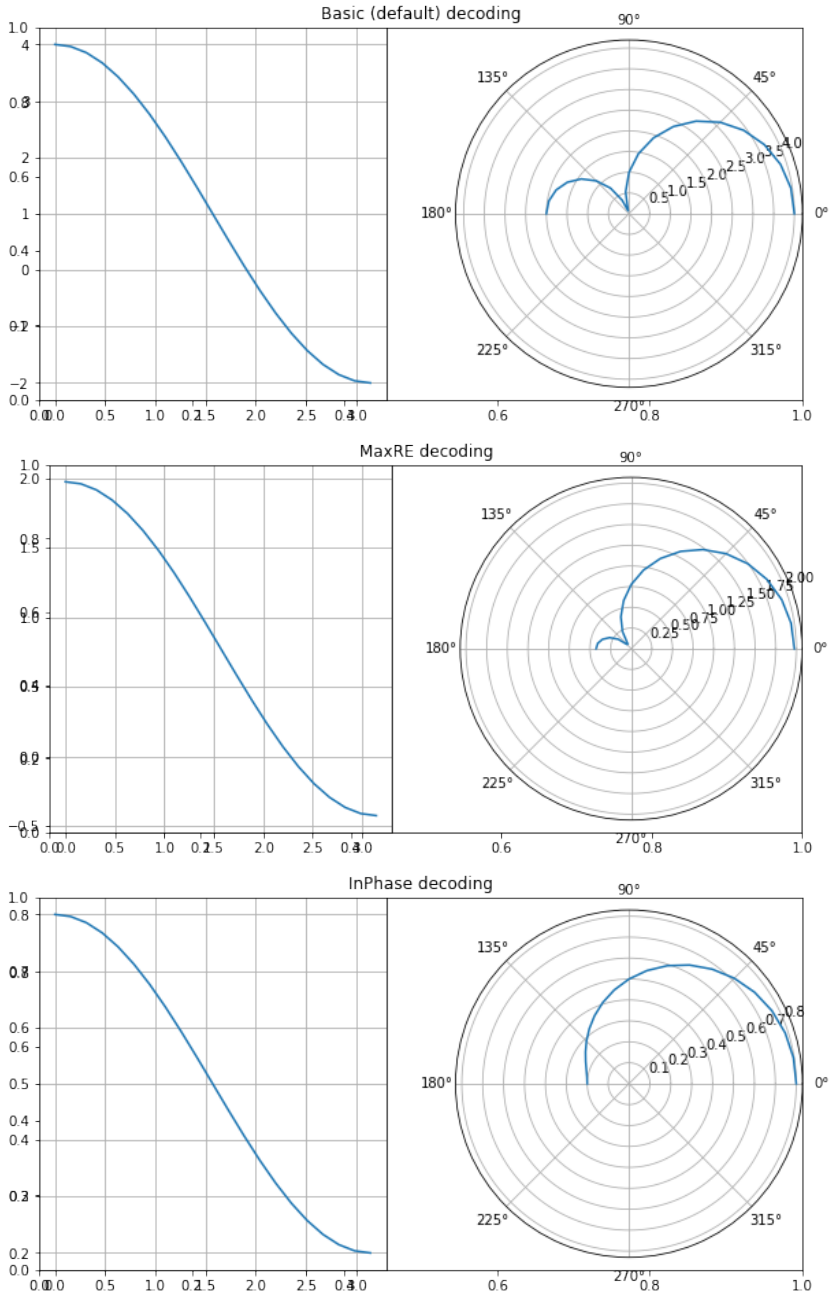

Figure 2.4: Directive patterns of first-order ambisonic decoding.

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

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 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 diag(α_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 [Zotter and Frank, 2012]. 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 [Pulkki, 1997].

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].

[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.

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.

Alternatively, the *Schmidt 3d semi-normalized* or *SN3D* [Daniel, 2000] 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 as *SN3D*. Besides this, *FuMa* has been historically the default normalization [Gerzon, 1985], while the more modern *N3D* and *SN3D* were popularized after J. Daniel's work [Daniel, 2000].

As a summary, Figure 2.5 displays the different normalization schemes. The reader is referred to [Carpentier, 2017] for an extensive review on the topic.

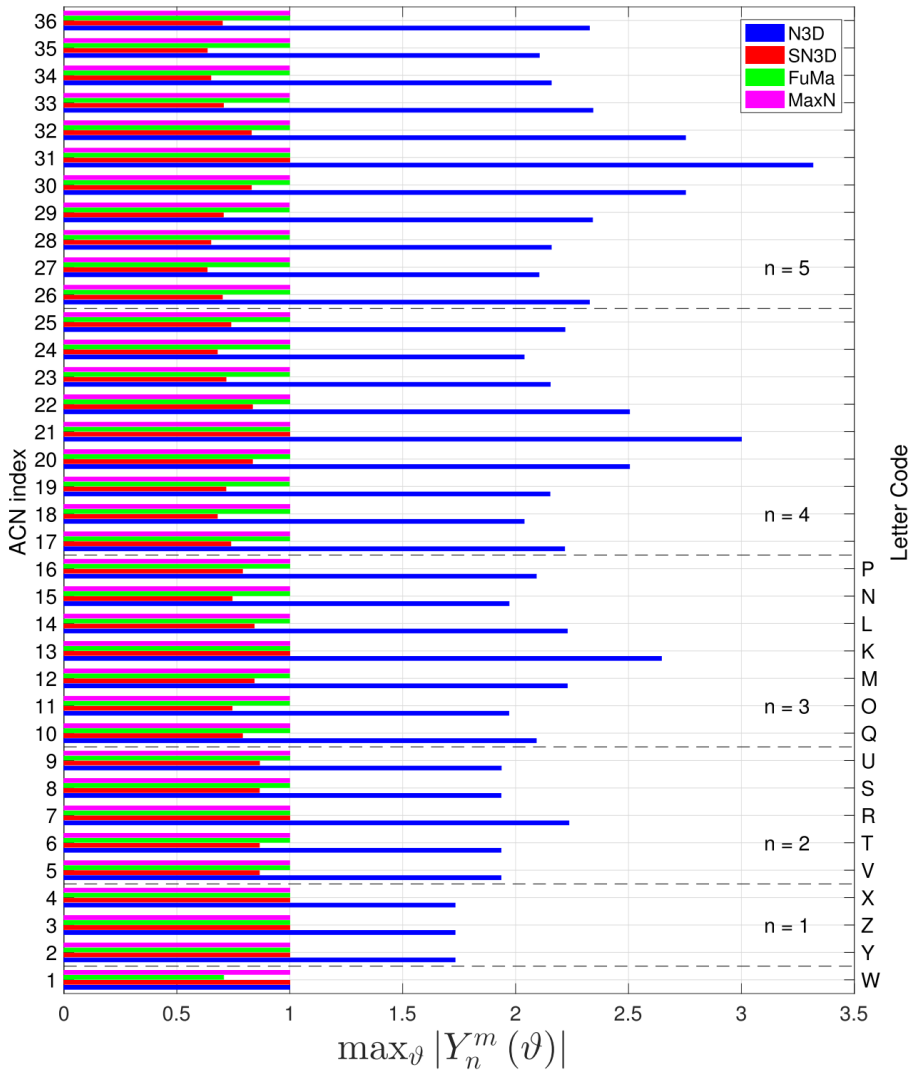

Figure 2.5: Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from [Carpentier, 2017].

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 2.5 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 medium density, and c is the sound speed.

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 2.6 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.

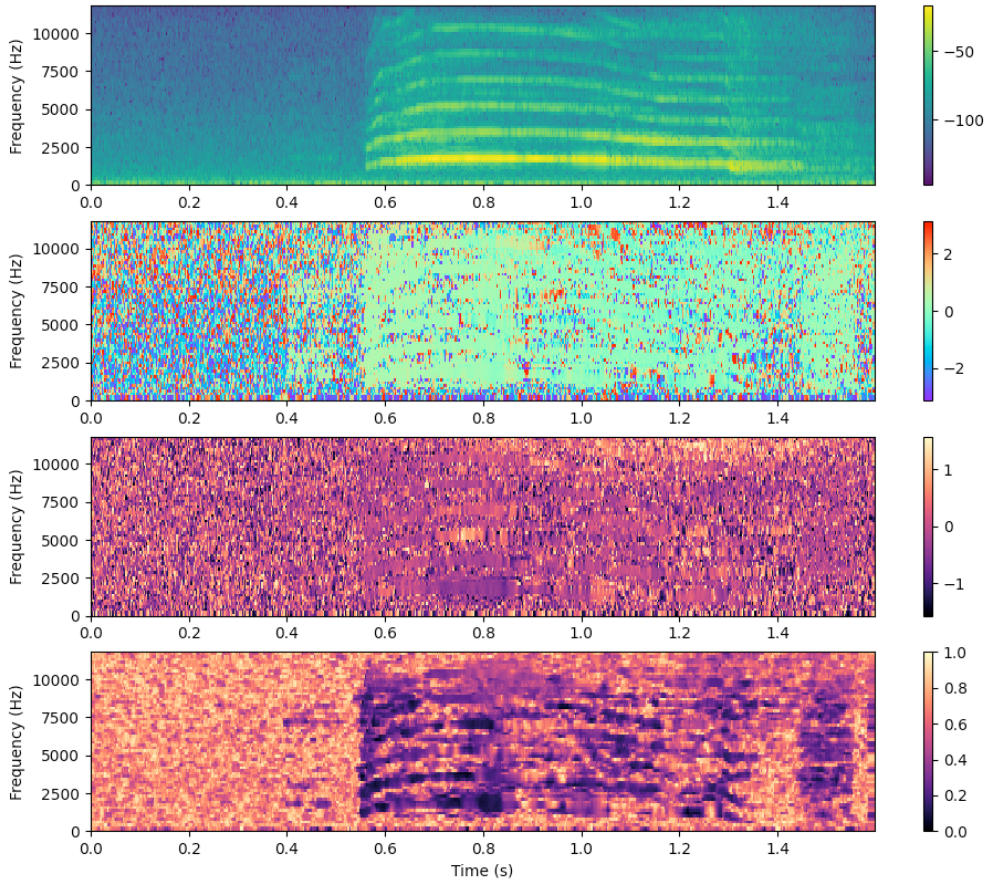

Figure 2.6: Parametric time-frequency spatial audio analysis of a first order ambisonic recording. From top to bottom: 1.) Magnitude spectrogram of the omnidirectional channel. 2.) and 3.) Azimuth and elevation of the estimated instantaneous narrowband DOAs $\Omega(k, n)$. 4.) Instantaneous narrowband diffuseness $\Psi(k, n)$.

2.5 Spatial Coherence Analysis

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 capsule 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]:

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

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):

$$
\begin{aligned}
N_{12}(kr) &= \frac{\alpha_1\alpha_2 sin(kr)}{kr} \\
&+ \frac{(1-\alpha_2)(1-\alpha_2)(x_1x_2+y_1y_2)}{(kr)^3}(sin(kr)-krcos(kr)) \\
&+ \frac{z_1z_2}{kr^3}[((kr)^2sin(kr)+2krcos(kr))(1-\alpha_1)(1-\alpha_2)+2sin(kr)(1-\alpha_1)(1-\alpha_2)] \\
&+ \frac{z_1}{(kr)^3}[j(kr)^2\alpha_2cos(kr)(\alpha_1-1)+jkr\alpha_2sin(kr)(1+\alpha_1)] \\
&+ \frac{z_2}{(kr)^3}[j(kr)^2\alpha_1cos(kr)(\alpha_2-1)+jkr\alpha_1sin(kr)(1+\alpha_2)], \\
D_{12}(kr) &= \frac{\sqrt{3\alpha_1^2+(1-\alpha_1)^2}\sqrt{3\alpha_2^2+(1-\alpha_2)^2}}{3},
\end{aligned}
\tag{2.36}
$$

where x_i, y_i and z_i are the cartesian coordinates of the wave incidence angle $\Omega_i = (\varphi_i, \vartheta_i)$.

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

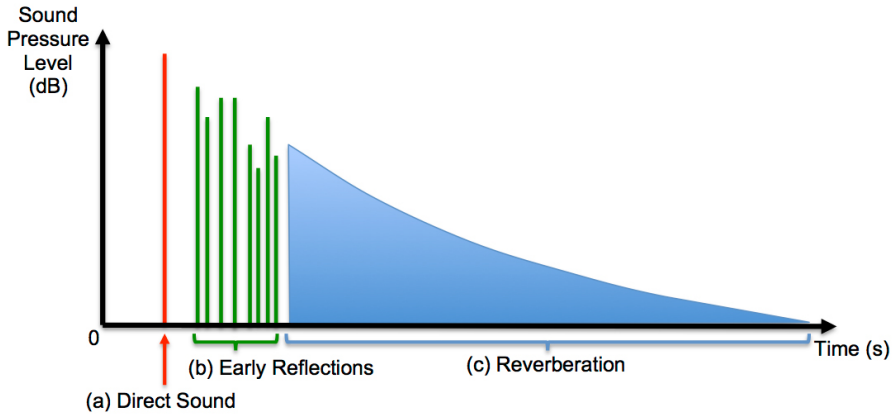

Figure 2.7: Room impulse response model, from [Murphy et al., 2017].

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 2.7.

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.

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

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

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 2.8.

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:

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

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

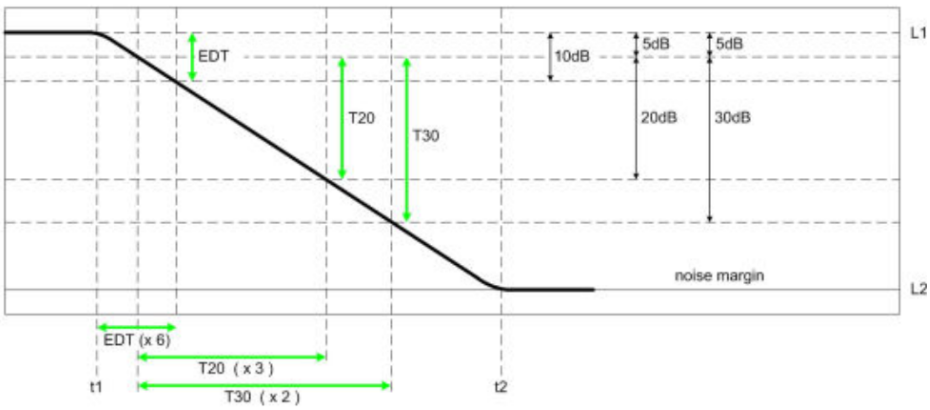

Figure 2.8: Room impulse response model, adapted from [AV_INFO, 1995].

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*.

2.7 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.

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 comparative 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 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 5.5. Finally, a conclusion is presented in Section 3.7.

The research presented in this Chapter has been submitted for publication at the IEEE 22nd International Workshop on Multimedia Signal Processing [Pérez-López et al., 2020].

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:

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

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 (Section 6.2). 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.

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

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, 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 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 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.

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**

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.

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 5.3.

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 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 our 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.

3.7 Conclusion

In this Chapter, we have presented 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.

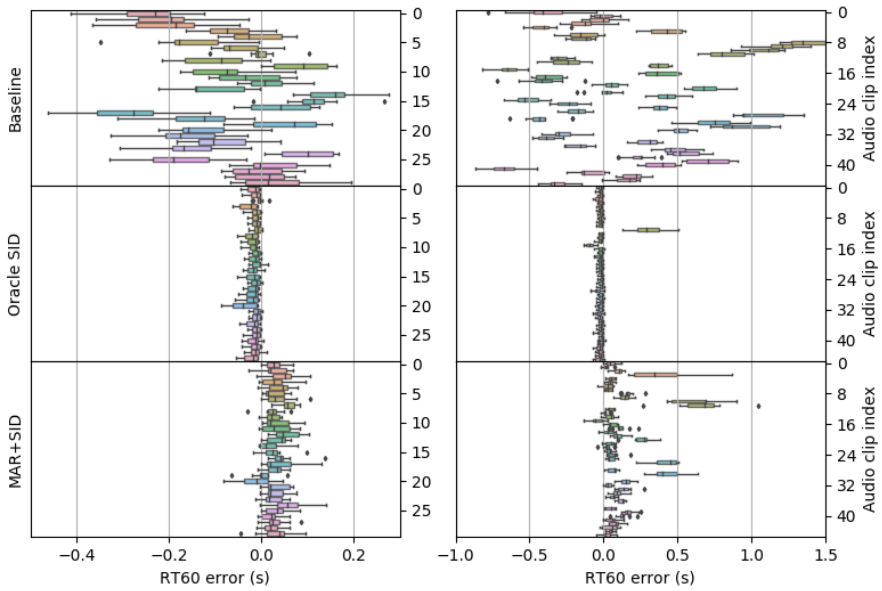

Figure 3.1: Experiment results for *speech* (left column) and *drums* (right column) datasets. Estimation error computed for each audio clip.

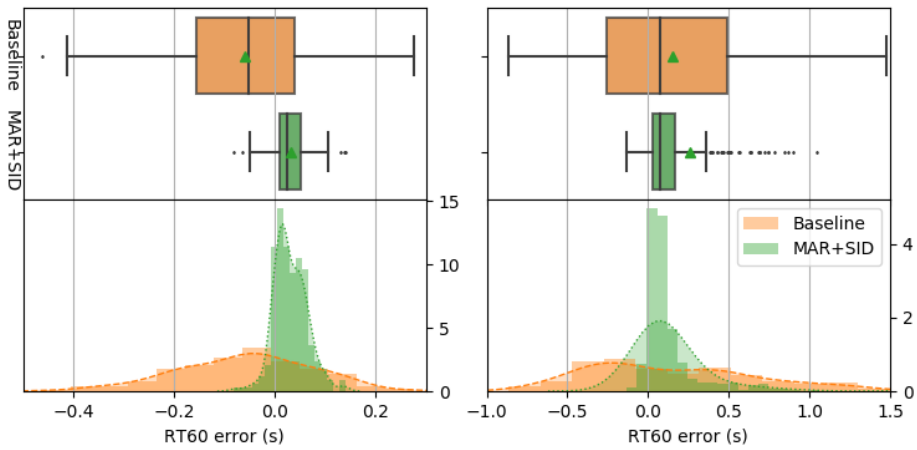

Figure 3.2: Experiment results for *speech* (left column) and *drums* (right column) datasets. Total estimation error across audio clips and acoustic conditions. Top: boxplot. Bottom: histogram and density plot

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 [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., 2018], Direction-of-Arrival estimation [Thiergart et al., 2009] or source separation [Motlicek et al., 2013].

In this Chapter, adapted from a previous academic publication [Pérez-López and Stefanakis, 2019], 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 [Sennheiser, 2020].

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 $\varepsilon(k)$:

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

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 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.

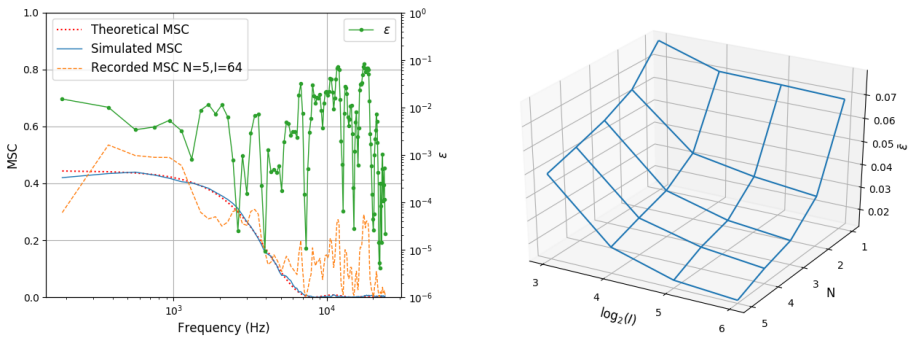

Figure 4.1: *A-Format* coherence between microphone signals. Left: MSC as a function of the frequency of theoretical, simulated and recorded *((BLD,BRU)*, $N = 5, I = 64$) signals. Right: mean error $\bar{\varepsilon}$ of the recorded signals' MSC *(BLD,BRU)* compared to the simulated values, for all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

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 $\Delta_{min} \approx 0$.

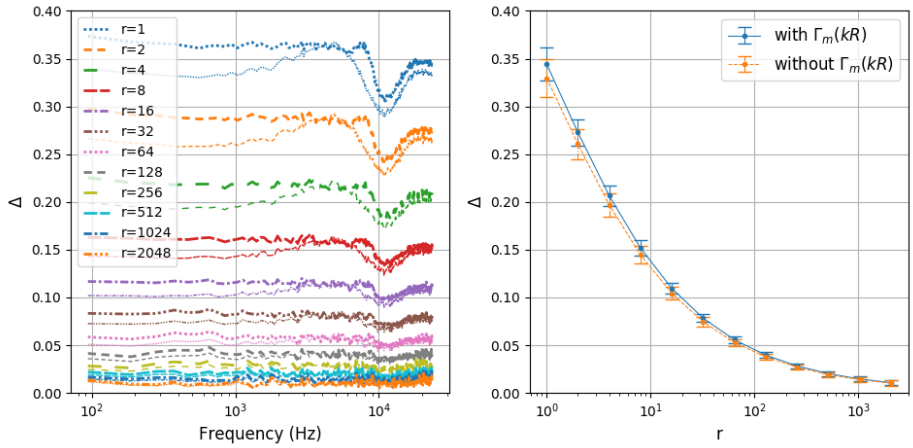

Figure 4.2: Estimated *B-Format* coherence (Δ) of a simulated diffuse sound field, as a function of the temporal averaging vicinity radius r. Left: $\Delta(k)$ for different values of r, with (coarse) and without (fine) application of radial filters. Right: mean and standard deviation of $\Delta(k)$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

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, 1975a]. 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 might be able to yield more accurate diffuseness estimations, due to a better signal conditioning [Epain and Jin, 2016].

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.

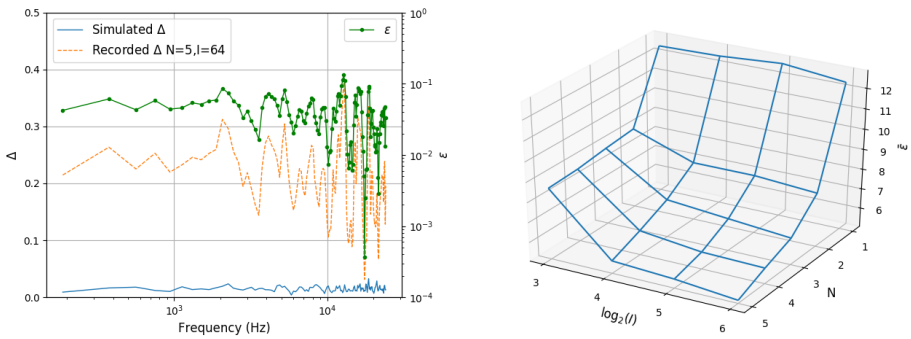

Figure 4.3: *B-Format coherence* between microphone signals. Left: Δ of simulated and recorded ($N = 5, I = 64$) signals. Right: $\bar{\varepsilon}$ of the recorded signals coherence across all values of N and I.

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 **spatial location** Ω, **temporal activity** Υ, and **sound class** κ to which it belongs.

The organization of a dedicated SELD task within the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 can be considered as a milestone for the development of the SELD research problem. Indeed, a large number of novel methodologies were developed for the Challenge, most of them based on Convolutional Recurrent Neural Networks (CRNN). The performance of the baseline method, a CRNN that performed jointly the localization and classification tasks [Adavanne et al., 2018], was vastly exceeded by a variety of deep-learning based algorithms [Kapka and Lewandowski, 2019, Cao et al., 2019, Grondin et al., 2019]. Some of these improvements have been included in the baseline system for the SELD Challenge of DCASE 2020.

Despite the predominant trend towards high-complexity deep-learning architectures, some recent works have been able to match or even improve CRNN-based methods with regard to localization, by using parametric analysis of the ambisonic sound field [Pérez-López et al., 2019, Nguyen et al., 2020b]. Apart from the benefit derived by their simplicity, these approaches are able to resolve the case of overlapping events of the same class, a situation difficult to disambiguate for CRNN-based methods [Politis et al., 2020].

The present work continues the exploration of possibilities of parametric SELD methods, focusing on a low-complexity architecture that makes use of traditional, feature-based machine learning techniques. The method has been developed in the context of the SELD task within DCASE 2020 Challenge, and therefore utilizes the proposed dataset, baseline system and evaluation metrics.

Finally, it is important to remark that the method described here has been reported in a conference paper submitted to the DCASE Workshop 2020 [Pérez-López and Ibanez-Usach, 2020], which in turn describes the improvement of an algorithm presented at the DCASE 2019 Challenge [Pérez-López et al., 2019].

While both algorithms share a common structure and a similar approach to the SELD problem, the current proposal tries to solve some of the problems identified on our early approach, mainly related with a low frame recall derived from a naive approach to event segmentation. Moreover, the current method also presents a complexity reduction, regarding the single-source DOA estimation and the classifier.

5.2 System description

The proposed method, referred to as *PAPAFIL*, can be summed up in four steps:

1. Estimate single-source time-frequency bins.
2. Use a particle tracking system to estimate event trajectories and activation times from single-source bins.
3. Perform spatio-temporal filtering on the input signal.
4. Assign a class label to the estimated event.

A scheme of the method is shown in Fig. 5.1.

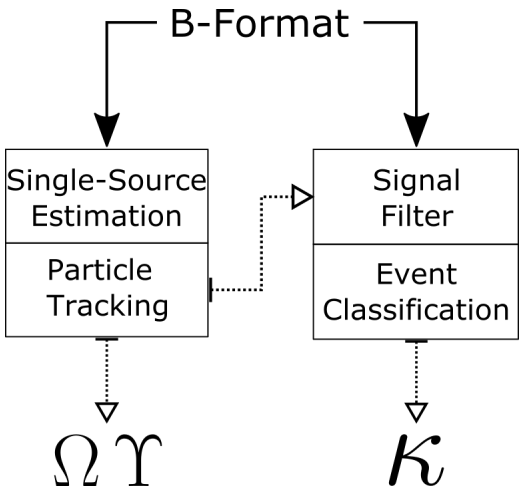

Figure 5.1: Architecture of the proposed methodology.

5.2.1 Single-source estimation

The first step is the transformation of the B-Format input signal $x_n^m(t)$ using the Short-Time Fourier Transform (STFT) into the time-frequency (TF) signal $X_n^m(k,n)$.

In the resulting spectrogram, the frequencies above a given limit f_{max} are discarded; this procedure speeds up the method while maintaining the directional information, given that the microphone geometry produces spatial aliasing above approx. 5 kHz [Bertet et al., 2006].

Assuming that the sources are sparse in time-frequency, it could be possible to identify TF bins which contain a significant energetic contribution from only one source. These bins could be then used to produce accurate DOA estimates. The effectiveness of this approach has already been demonstrated [Tho et al., 2014, Nguyen et al., 2020b].

Single-source TF bins are computed from the DirAC parametric analysis. A variety of alternative subspace methods are known [Epain and Jin, 2016, Madmoni and Rafaely, 2018]; however, those methods require local estimation of eigenvalues through the Spatial Covariance Matrix (SCM), which is a computationally expensive procedure; this is the main reason for the choice of DirAC-based analysis in this work.

A TF bin is counted as single-source if its diffuseness $\Psi(k, n)$ is lower than a threshold Ψ_{max}. Diffuseness is computed here using Eq. 2.31. Finally, the DOA $\Omega(k, n)$ of the TF bins passing the aforementioned single-source test is computed as the angle of the active intensity vector by Eq. 2.29. To illustrate the process, an example of the method output is plotted in Fig. 5.2 (top).

5.2.2 Particle tracking

Once a set of reliable TF DOA estimates is obtained, the next step is the generalization of the individual measurements into trajectories and temporal activations. In our case, we opted for the Rao-Blackwellized Monte-Carlo Data Association (RBMCDA) algorithm [Särkkä et al., 2004], which decomposes the multiple target tracking problem in two: it solves first the data association problem, and then performs the single target tracking

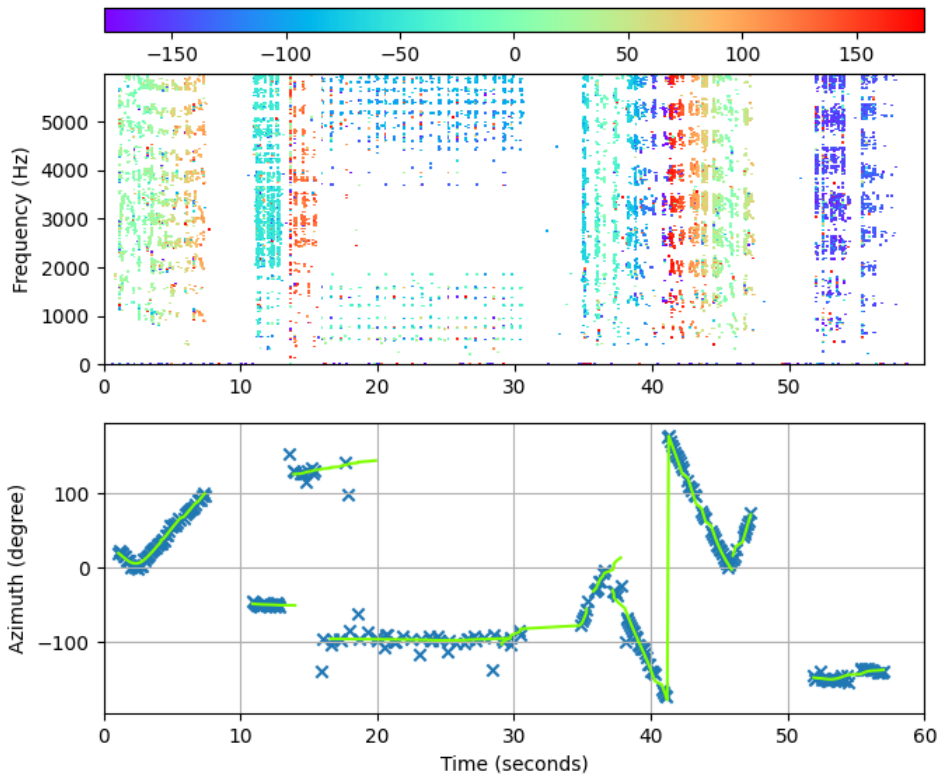

Figure 5.2: Estimation of localization and temporal activation. Top: azimuth spectrogram after diffuseness mask; color indicates estimated position of a TF bin passing the single-source test. Bottom: input/output of the particle tracking; the crosses represent the measurement space, and the continuous lines are the resulting events.

individually. This method has been recently used in the context of sound event localization and tracking with successful results [Adavanne et al., 2018, Adavanne et al., 2019b]; the code used for our implementation has been adapted from the same authors[1].

[1] https://github.com/sharathadavanne/multiple-target-tracking

The system takes as the input the set of TF DOA values passing the single source test, and produces spatio-temporal event trajectories, considering an event as an entity with contiguous temporal activation and continuous spatial position. More specifically, for each time frame, the median[2] of all narrowband masked DOA estimates is computed. The resulting value is added to the measurement space of the tracker if the number of single-source frequency bins for that frame exceeds a minimum K_{min}.

The performance of the RBMCDA algorithm is controlled by several parameters. Some of the most relevant ones are the angular velocity prior v, the standard deviation σ_ν and the spectral density s_ν of the measurement noise, the prior probabilities of birth p_{birth} and noise percentage p_ν, and the number of Monte-Carlo particles N. Position-related parameters are adjusted with respect to their ranges, so that azimuth-related magnitudes double elevation values.

The procedure is followed by a numerical post-processing step, which includes data interpolation, resampling (if needed), and removal of elements shorter than T_{min}. Finally, the system provides a list of J events, each one having an instantaneous position $\Omega_j(t)$ and a temporal activation Υ_j. An example of the system inputs and outputs is depicted in Fig. 5.2 (bottom).

5.2.3 Signal filter

The information provided by the particle tracking system is used to spatially filter the input signal. This can provide an enhanced monophonic estimate of an event $\tilde{s}_j(t)$ with reduced influence of simultaneous events. The process is performed by steering a virtual first-order cardioid in the direction of interest, using Eq. 2.15. The result of this process is a monophonic estimate for each event, $\tilde{s}_j(t)$, temporally delimited by Υ_j. As a last step, each

[2]Circular median in the case of azimuth.

estimate is amplitude peak-normalized, in order to minimize potential amplitude variability due to arbitrary configurations of the scene.

5.2.4 Event classification

As a final step, a class label is assigned to each estimated event $\tilde{s}_j(t)$ using a single-class classifier. Since the objective is to keep complexity low and make results interpretable, a machine learning algorithm is used instead of deep learning frameworks. The main advantages of this choice are: (i) low number of parameters; (ii) low train and predict computational time, easing reproducibility; and (iii) relative importance of the features in the output can be interpreted, which is not possible with deep learning approaches.

Gradient Boosting Machine (GBM, Fig. 5.3) has been selected as the classification algorithm since it is a powerful yet simple technique for predictive modeling. In essence, the algorithm is aimed to minimize the loss of the objective function by adding many weak learners. These learners are typically simple decision trees and their parameters are tuned using gradient descent techniques. GBM implementation makes use of the *scikit-learn* library [Pedregosa et al., 2011].

Sound features are obtained using extractors from *Essentia*, an open-source library for audio analysis [Bogdanov et al., 2013]. Given the heterogeneous nature of the sound classes, a mixture of spectral, temporal and harmonic features are used, as shown in Table 5.1. Features are be computed either frame-based or on the whole event; in the former case, the classifier is fed with their temporal first-order statistics.

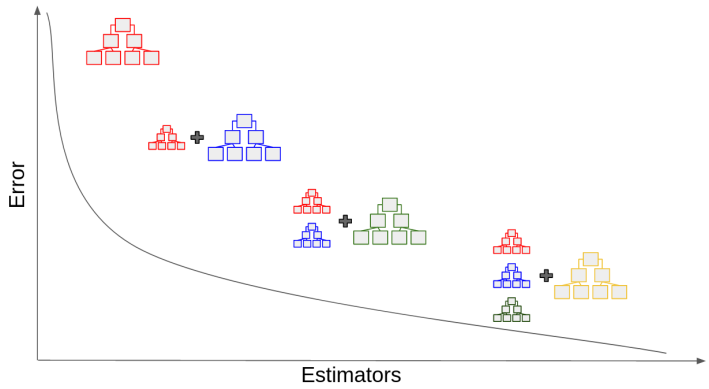

Figure 5.3: Gradient boosting machine learning process. Adding weak estimators allows reducing overall error in the predictions.

Table 5.1: Acoustic features used for classification, grouped by type.

Type	Features	Number
Low-level	Mel bands	24
	MFCC	13
	Spectral Features	26
SFX	Duration	2
	Harmonic	4
	Sound envelope	11
	Pitch envelope	4

5.3 Experiments

5.3.1 Dataset and baseline system

The dataset used is the FOA subset of the development set of the *TAU-NIGENS Spatial Sound Events 2020* [Politis et al., 2020], which features 600 different B-Format clips of 60 seconds long each. Each clip contains multiple sound events, which belong to one of the

fourteen sound classes from the NIGENS database [Trowitzsch et al., 2019]. Events are also located at a potentially time-varying positions, and the maximum instantaneous overlapping of sources allowed is limited to two. Fifteen different Room Impulse Responses (RIR) are used for scene reverberation, covering a vast range of acoustic conditions. Furthermore, the audio clips contain a moderate amount of recorded background sounds.

The baseline method is based on the recently proposed SELDnet architecture [Adavanne et al., 2018], which features a Convolutional Recurrent Neural Network (CRNN) that solves both localization and classification problems jointly. Additionally, the baseline implementation has been improved with several changes inspired by one of the best performing methods in DCASE 2019 Task 3 Challenge [Cao et al., 2019].

5.3.2 Experimental setup

In order to explore the performance of the system, two different approaches have been undertaken regarding the creation of the training dataset for the monophonic single-class classifier. The first approach, referred to as *PAPAFIL1*, collects all event localization, temporal activation and class information by parsing the annotation files. Conversely, the second approach, called *PAPAFIL2*, uses the proposed parametric particle filter to estimate localizations and activations, and the class label is assigned to each event by a custom association algorithm based on spatio-temporal distance. In both cases, the input signal is filtered with the obtained information in order to conform the monophonic event estimates.

Therefore, the difference between training datasets is noticeable: while the training events in *PAPAFIL1* are more accurately determined than in *PAPAFIL2*, the differences with respect to the prediction scenario are much bigger in the former case. The number of individual events for each of the approaches is plotted in Fig. 5.4. Approximately half of the classes have

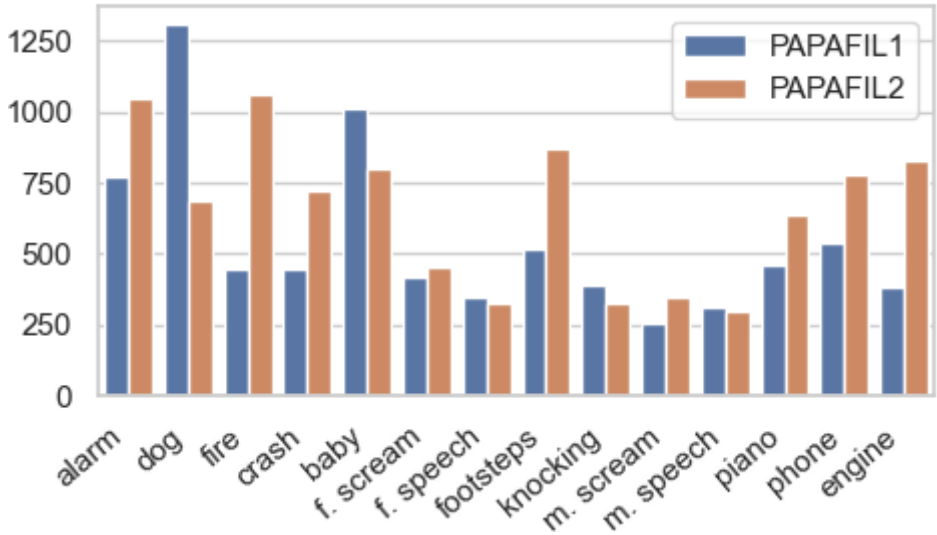

Figure 5.4: Number of occurrences of each event class in the training set, for both proposed methods.

similar number of instances in both datasets. However, the other half presents noticeable differences, which might be explained by the different criteria applied for the consideration of event temporal activations: the groundtruth seems to follow a frame-based activity detection approach, while the output of the proposed method tends to consider events as time-continuous manifestations, influenced by the particle filter.

This situation leads to two different *oracle* systems (referred to by appending -*O* in the method name), which represent the best performance theoretically achievable for the corresponding method.

The accurate information of the *PAPAFIL1* training set suggests a need for data augmentation; in contrast, the training material used in *PAPAFIL2* is already provided by a certain extent of variability. This situation motivates the implementation of data augmentation methods in the *PAPAFIL1* training set. Specifically,

several standard data augmentation techniques are implemented: pitch shifting, time shifting, time stretching and white noise addition. Furthermore, given the observed high influence of reverberation in the system performance, a reverberant data augmentation technique based on synthetic RIRs has been considered. Ten different single-channel RIRs, with reverberation times between 0.3 and 1.1 seconds, have been synthetically created using the *masp* library [Pérez-López and Politis, 2020]. During training, each event estimate is convolved with one of the RIRs, randomly chosen. RIR augmentation has recently been shown very effective for blind reverberation time estimation [Bryan, 2020] but, to the best of the authors' knowledge, this is the first application in SELD.

Table 5.2 shows a comprehensive list of the parameters used throughout the different steps of the proposed method. All values are equal for both presented approaches, except for the number of Monte-Carlo particles N. The values for Single-Source Estimation and Particle Filtering parameters have been iteratively refined by manual tuning and inspection, departing from standard values. The beamforming weights α_m correspond to the *maximum directivity beamformer*, which minimizes the energy contributions from directions other than the lookup direction [Rafaely, 2015]. In the spatial audio field, such property is also known as the *max-rE* decoder [Daniel, 2000]. Regarding event classification, a cross-validation scheme has been implemented for tuning GBM hyperparameters.

5.3.3 Evaluation metrics

The system is evaluated according to the joint metrics proposed in the Challenge [Mesaros et al., 2019]. The metrics evaluate jointly the localization and the classification, and are divided into two types: location-aware classification, and classification-aware localization. There are two classification metrics: Error Rate

Table 5.2: (Hyper-)parameter values.

Step	Parameter	Value	Unit
Single-Source	sample rate	24	kHz
	window size	2400	samples
	window overlap	50	%
	f_{max}	6	kHz
	N_Ψ	2	frames
	Ψ_{max}	0.1	
Particle Filtering	v	2	°/frame
	σ_ν	5	
	s_ν	20	
	p_{birth}	0.25	
	p_ν	0.25	
	N	100 / 30	
	K_{min}	10	bins/frame
	T_{min}	10	frames
Signal Filter	α_0	0.775	
	α_1	3 * 0.4	
Event Classification	number of estimators	1300	trees
	loss	*mlogloss*	
	learning rate	0.05	
	max depth	4	
	min samples leaf	10	samples

(ER_{20}) and F-Score (F_{20}). As the name suggests, the metrics are conditioned to a minimum localization performance, which is set to 20° in this case. Localization metrics are also two-fold: Localization Error (LE_{CD}) and Localization Recall (LR_{CD}); as their name suggests, the metrics are class-dependent, and thus are conditioned to a correct classification. Finally, the SELD score is an average of the four other metrics, used to conveniently sum up the results.

Table 5.3: System evaluation. Top: results on the cross-validation development set. Bottom: results on the evaluation set.

Method	ER_{20}	F_{20}	LE_{CD}	LR_{CD}	SELD
BASELINE	0.70	39.5 %	23.2°	**62.1** %	0.45
PAPAFIL1	0.60	49.8 %	**13.4**°	54.4 %	0.41
PAPAFIL2	**0.57**	**54.0** %	13.8°	59.7 %	**0.38**
PAPAFIL1-O	0.37	67.0 %	2.0°	68.6 %	0.26
PAPAFIL2-O	0.32	79.6 %	8.5°	82.4%	0.19
BASELINE	0.72	37.4 %	22.8°	60.7 %	0.47
PAPAFIL1	0.55	56 %	12.8°	61.1 %	0.36
PAPAFIL2	**0.51**	**60.1** %	**12.4**°	**65.1** %	**0.33**

5.4 Results

Table 5.3 summarizes the results of the experiments for development and evaluation datasets. Results on the development set have been computed using the provided cross-validation scheme: split 1 for testing, split 2 for validation, and splits 3 to 6 for training.

Results are reported for three different systems: the baseline and the two proposed methods *PAPAFIL1* and *PAPAFIL2*. The results of their respective oracle results, *PAPAFIL1-O* and *PAPAFIL2-O*, are also provided for the development set.

Regarding the development set, both proposed approaches outperform the baseline system in three out of the four evaluation metrics (ER_{20}, F_{20} and LE_{CD}). Although the results obtained by both of them are similar, *PAPAFIL2* obtains better classification scores (ER_{20} and F_{20}), and *PAPAFIL1* performs subtly better regarding localization error (LE_{CD}). However, the localization recall results (LR_{CD}) are slightly worst than the baseline in both cases. This fact does not prevent the proposed methods to have a

SELD score better than the baseline: 0.41 (*PAPAFIL1*) and 0.38 (*PAPAFIL2*), against 0.47 (*BASELINE*).

The results obtained by the oracle methods are within the expected ranges. *PAPAFIL1-O* performs almost perfectly regarding LE_{CD}, but the classification errors influence the LR_{CD} result. In turn, *PAPAFIL2-O* performs better than *PAPAFIL1-O* regarding all metrics, excepting LE_{CD}; this improvement is specially noticeable in LR_{CD}, with a performance difference of about 15%.

The good results obtained by *PAPAFIL2-O* validate the proposed particle filtering approach, and leave space for improvements that might be given by a better understanding and fine tuning of the model.

The overall tendency is maintained in the evaluation set results. Both proposed methods outperform again the baseline, improving the development set SELD score in five points; since the baseline slightly decreases the performance for the evaluation set, the score difference with respect to the proposed methods diverges significantly. *PAPAFIL2* is the method that clearly performs better on the evaluation set, with better results than *PAPAFIL1* in all evaluation metrics.

The performance of the proposed methods deteriorates noticeably with overlapping sounds. A closer inspection to the development set results reveals that, in many occasions, the TF bins passing the single-source test mostly belong to one out of two simultaneous sources. It is a known issue that performance of DirAC diffuseness is reduced when two sources are present [Epain and Jin, 2016]; similar problems have been reported in [Adavanne et al., 2019b], where an instantaneous source number estimator is used in combination with the particle filter. As in that case, the results suggest the need for more sophisticated source detection and counting methods.

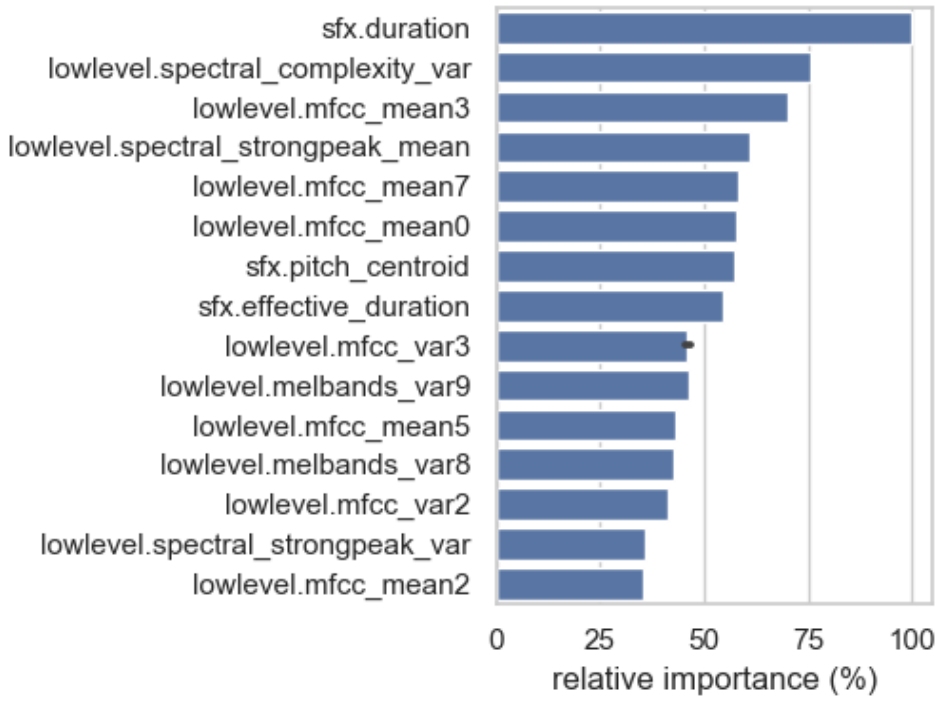

Figure 5.5: Most representative features in event classifier.

Fig. 5.7 shows the relative importance of the fifteen most relevant acoustic features for the *PAPAFIL2* classifier model. Event duration is clearly the feature with the highest importance, and effective duration (duration of the signal discarding silence) also appears in the eighth position. This fact can help to explain the better performance of *PAPAFIL2* over *PAPAFIL1*: the temporal activities of the events in training and prediction are much more similar to each other in the former method, as a consequence of the training set generation approach. In order to provide a deeper insight on the temporal characteristics of the data under consideration, the durations of all sound events in the *PAPAFIL2* training set are summarized in Figure 5.6.

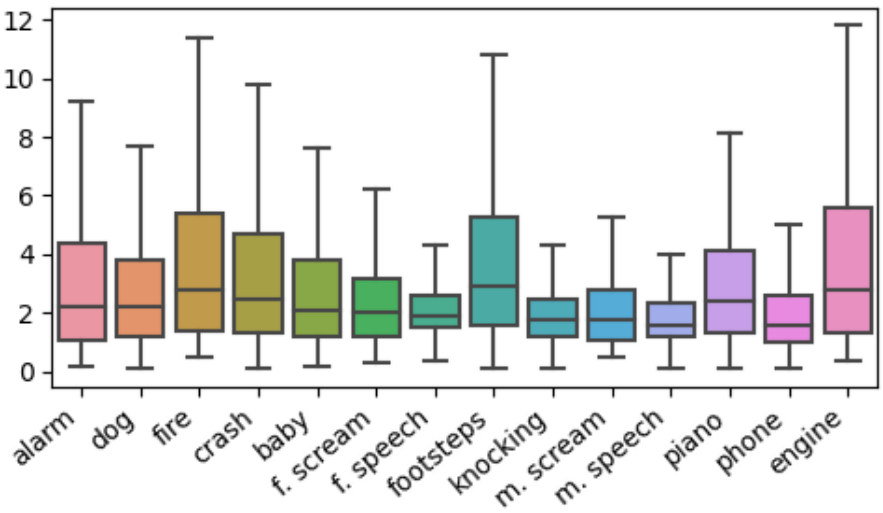

Figure 5.6: Event durations of all elements in the *PAPAFIL2* training set.

Furthermore, it is interesting to notice the high relevance of low-level features; specifically, several MFCC combinations (eight of the fifteen reported features) and various extractors related to the spectral structure. The absence of pitch, harmonic and envelope features in the list represents also a significant finding.

The results of the DCASE 2020 Challenge for the SELD task are shown in Table 5.4. The method entries corresponds to the best scoring method for each team submission, respecting the naming conventions established for the Challenge. Accordingly, *PerezLopez* refers there to the *PAPAFIL2* method. The final rank is computed as the cumulative rank across all evaluation metrics, sorted by ascending order. The data has been taken from the Task results website [DCASE, 2020], which provides complete information regarding the challenge.

Table 5.4: DCASE 2020 Challenge Task 3 evaluation results

Rank	Method	ER_{20}	F_{20}	LE_{CD}	LR_{CD}	SELD
1	Du	**0.20**	**84.9** %	**6.0**°	88.5 %	**0.12**
2	Nguyen	0.23	82.0 %	9.3°	**90.0** %	0.14
3	Shimada	0.25	83.2 %	7.0°	86.2 %	0.15
4	Cao	0.36	71.2 %	13.3°	81.1 %	0.22
5	Park	0.43	65.2 %	16.8°	81.9 %	0.26
6	Phan	0.49	61.7 %	15.2°	72.4 %	0.30
7	PerezLopez	0.51	60.1 %	12.4°	65.1 %	0.33
8	Sampathkumar	0.53	56.6 %	14.8°	66.5 %	0.35
9	Patel	0.55	55.5 %	14.4°	65.5 %	0.38
10	Ronchini	0.58	50.8 %	16.9°	65.5 %	0.39
11	Naranjo-Alcazar	0.61	49.1 %	19.5°	67.1 %	0.38
12	Song	0.57	50.4 %	20.0°	64.3 %	0.38
13	Tian	0.64	47.6 %	24.5°	67.5 %	0.40
14	Singla	0.88	18.0 %	53.4°	66.2 %	0.58
15	Baseline	0.69	41.3 %	23.1°	62.4 %	0.45

It is important to remark that, apart from our proposed method, all other systems rely completely on deep learning algorithms, specially making use of different configurations and combinations of CRNNs.

The only exception to that tendency is the method presented by Nguyen and colleagues, which is an adaptation of their recently proposed methodology [Nguyen et al., 2020a, Nguyen et al., 2020b]. In this method, localization is also based on the identification of the single-source TF bins; instead of DirAC diffuseness, they use a subspace analysis to find the TF bins associated with nearly rank-1 spatial covariance matrices. Furthermore, the instantaneous DOA estimates are obtained by finding the azimuth and elevation histogram peaks, with the steering directions obtained from the covariance matrix diagonals.

While this method might contribute to provide a smaller localization error compared with our approach (9.3° versus 14.2° in LE_{CD}), it is more expense computationally, since it involves building a SCM for each TF bin, plus peak picking in two 1D histograms (one for each spherical coordinate). However, the LE_{CD} results are also influenced by the classification accuracy; a deeper analysis would be required to perform a direct comparison of both approaches.

Table 5.4 also highlights the good result obtained by our method regarding localization error: it is the fourth best method with respect to LE_{CD}. Moreover, the two model-based localization methods (our system and Nguyen) rank fourth and second in the LE_{CD} classification, respectively. This result suggests that parametric spatial audio analysis, combined with a robust tracking system, is able to produce state-of-the-art localization results in the context of SELD.

To conclude the analysis, Figure 5.7 shows the relationship between the system complexity, measured in number of parameters of the model, against SELD score, for all submitted systems, denoted by their ranking position. The overall tendency is clear: the higher the number of parameters, the better (smaller) the SELD score; the regression curve, computed from all observations, confirms such tendency. It must be noticed that the presented method (7 in the Figure) lies outside the 95% confidence interval of the regression curve (shaded area). In other words, its SELD score is significantly better than the tendency observed from all submissions, considering the system complexity. Figure 5.7 also highlights the great difference in complexity among submissions. Between our method (the less complex, with 20k parameters) and the first scoring method (the most complex, with 123M parameters) there are four magnitude orders of different, constituting a remarkable difference. The logarithmic mean of all submitted systems' complexities is around 2.3M parameters.

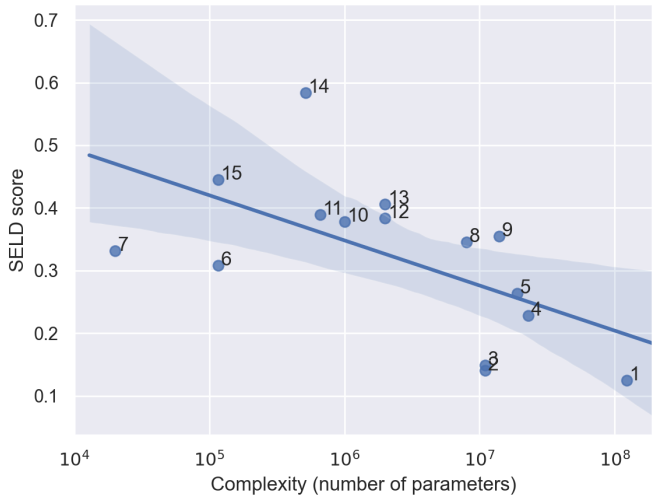

Figure 5.7: DCASE 2020 Task 3 submissions: complexity versus SELD score.

5.5 Conclusion

We present a novel low-complexity method for Sound Event Localization and Detection of First Order Ambisonic signals, based on four steps: estimation of single-source spectrogram regions by parametric analysis; computation of event trajectories and activations by means of a particle tracker; spatio-temporal filtering of the input signal; and single-class monophonic event classification by Gradient Boosting. Results show that the proposed method outperforms the baseline method, a state-of-the-art Convolutional Recurrent Neural Network. Specifically, our method is able to improve the baseline SELD score by almost ten points, while increasing the scores in three out of the four metrics under consideration.

Chapter 6

Data generation and storage

6.1 Introduction

This chapter gathers several proposals related with the creation, storage and transmission of ambisonic data for research purposes. The main objective of the contributions described here is the support for the generation of parametrizable ambisonic datasets, using both synthetic and recorded materials, and specifically emphasizing the usage of Room Impulse Responses.

Most of the contributions listed here (and also most of the code developed for this thesis) have been implemented in Python. Indeed, Python has recently become one of the most used programming languages worldwide [Robinson, 2017, PYPL, 2020, TIOBE, 2020]; as shown also in Figure 6.1.

One of the reasons behind this tendency shift is the popularity of the language among machine learning and data science communities, fields where Python holds the first place by usage [Elliott, 2019]. Since data-driven paradigms currently conform the state-of-the-art of many applied sciences, including audio signal processing, the availability of convenience Python packages and libraries is therefore of the highest interest to the research community.

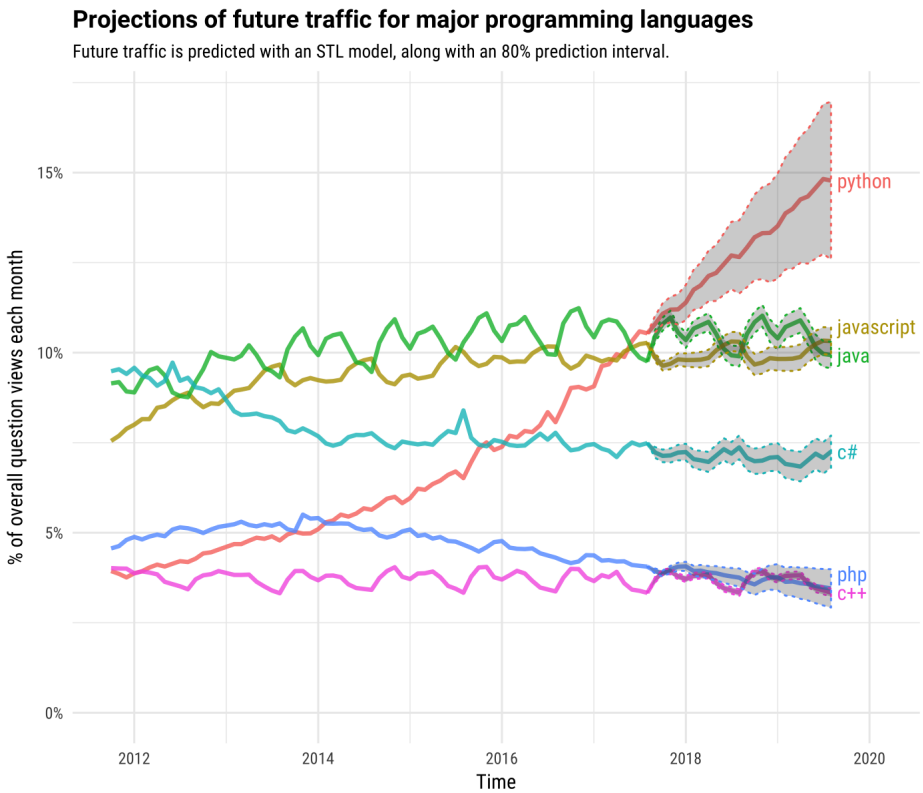

Figure 6.1: 2018 projections of future internet traffic for major programming languages. Adapted from [Robinson, 2017].

It is important to remark the predominant position that Matlab has always had regarding scientific computing. Indeed, it is still the tool of choice for many researchers, and the availability of libraries is accordingly very high. But the aforementioned tendency shift towards Python causes, as a side effect, the lack of many tools developed in Matlab by the research community.

Although Matlab code can be called and executed from Python, in practice this approach is suboptimal under several criteria. A better solution in the long run is the effective port of the code towards native Python code. Some of the libraries presented in this Chapter are partially or totally motivated by this scenario.

6.2 MASP: a Python library for multichannel acoustic signal processing

6.2.1 Description

The Multichannel Acoustic Signal Processing (*MASP*) is a Python library consisting of a collection of methods related with acoustics and microphone array processing [Pérez-López and Politis, 2020]. The library is mostly a transcoding from several Matlab libraries by A. Politis [Politis, 2016,Politis, 2020]. It can be conveniently installed using *pip*.

MASP implements a variety of methods for the simulation and analysis of reverberant acoustic scenes, with emphasis on microphone arrays with spherical geometries. More specifically, MASP is structured in submodules, with the following structure :

Array Response Simulator Simulation of spherical microphones:

- Rigid/open configurations.
- Scattering simulation.

- Arbitrary capsule distances, positions and directivities.

Shoebox Room Model Fast implementation of the Image Source Method [Allen and Berkley, 1979]:

- Convex 3D rooms.
- Arbitrary number of sources and receivers, with arbitrary positions, orientations and directivities.
- ISM expansion limited by order or time.
- Frequency-dependent wall absorption.
- RIR with spherical harmonic expansion.

Spherical Array Processing Transformation and analysis of signals measured with a spherical microphone array:

- A2B conversion with theoretical or measured filters.
- Signal-independent beamforming.
- Signal-dependent and adaptive beamforming.
- Direction of Arrival estimation.
- Diffuseness estimation.

Spherical Harmonic Transform Mathematical convenience tools.

The library implements a Unit Testing system, which numerically assesses the validity of the methods. More specifically, each function test calls the equivalent Matlab code under the hood. The numeric result is then sent back to Python, where it is evaluated against the own result.

In Figure 6.2, obtained with the Array Response Simulator package, the frequency response of a spherical microphone array to a plane-wave with varying incidence angle is shown. The array consists of a 2nd order supercardioid and a 3rd order hypercardioid, located at opposite directions of an open sphere,

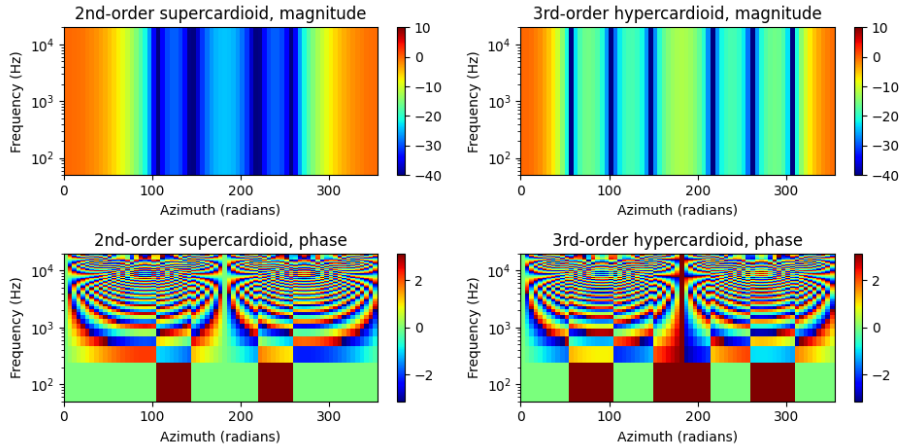

Figure 6.2: Frequency response of an arbitrary spherical array.

both of them facing front.

One of the features of the Spherical Array Processing package is shown in Figure 6.3. The plot shows the evaluation of radial filters $\Gamma_n(kR)$ for an arbitrary spherical array, generated by inverting the theoretical response of the array [Bertet et al., 2006]. The evaluation is performed following the metrics presented in the same paper, which compare spatial correlation, level difference and maximum amplification with respect to the ideal case.

6.2.2 Related software

There exists another recent Python library which covers a similar scope: *pyroomacoustics* [Scheibler et al., 2018]. This framework provides an object-oriented interface with two main application scopes: allow RIR simulation of complex rooms based on the image source method, and provide a reference implementation of

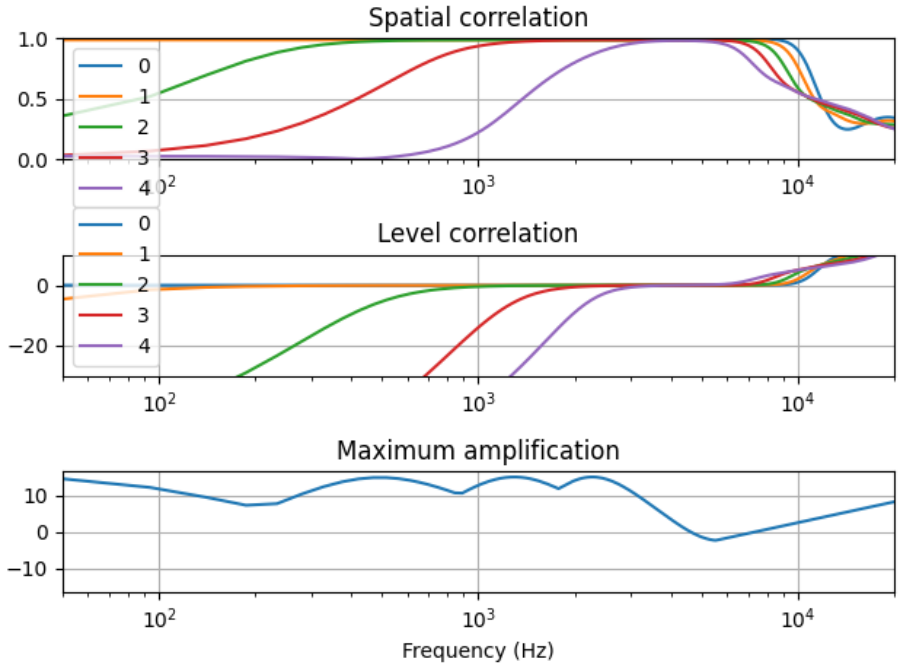

Figure 6.3: Evaluation of radial filters for an arbitrary spherical microphone array.

standard microphone array processing algorithms.

Although some of the features are common to both libraries, there is a significant difference regarding their target usage. While *MASP* primarily focuses on spherical geometries, *pyroomacoustics* is more concerned about arbitrary room geometries and computational performance. Therefore, both libraries might be considered as complementary to some extent. A comparative list of their features is shown in Table 6.1.

Table 6.1: Features of *MASP* compared to *pyroomacoustics*.

Package	Feature	MASP	PRA
Shoebox Room Model	Convex 3D room	✓	✓
	Non-convex 3D room	-	✓
	Arbitrary #sources	✓	✓
	Arbitrary #receivers, arrays	✓	✓
	ISM by max_order	✓	✓
	ISM by max_time	✓	-
	Wall absorption	✓	✓
	Frequency-dependent absorption	✓	-
	Plot methods	-	✓
	RIR rendering	✓	✓
	Audio simulation	✓	✓
	Acoustic descriptor estimation	✓	-
	Microphone orientation	✓	-
	Custom microphone directivity	✓	-
	RIR Spherical Harmonic Expansion	✓	-
Array Simulator	Rigid spherical arrays	✓	-
	Arbitrary capsule geometries	✓	-
	Recorded array IRs	✓	-
Spherical Processing	A2B conversion	✓	-
	Beamforming	✓	✓
	Plane-wave decomposition	✓	-
	Nullformer	✓	-
	Adaptive Beamforming	✓	-
	Adaptive Filtering	-	✓
	DoA Estimation	✓	✓
	Diffuseness Estimation	✓	-
	Diffuse-field coherence	✓	-
	Blind Source Separation	-	✓

6.3 SOFA

6.3.1 Problem statement

The availability of recorded room impulse responses is of great importance to many acoustic signal processing problems. Many different RIRs can be obtained from the same room, just by 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 final 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 an emphasis on *Head-Related Impulse Response* (HRIR) data, the framework that SOFA provides can be potentially applied to a variety of recording procedures and audio-related data. Such variety is associated with the concept of *conventions*: a specific data structure designed to hold a concrete type of data or measurement. Some examples of widespread conventions might be *SimpleFreeFieldHRIR* (for anechoic binaural measurements), *SimpleHeadphoneIR* (intended for storing headphone impulse responses), or *MultiSpeakerBRIR* (for binaural RIRs measured from loudspeaker arrays), to name a few of them.

6.3.2 Ambisonics Directional Room Impulse Response as a SOFA convention

Given the intrinsic spatial characterization capabilities of ambisonics, Gerzon proposed the technique as a potentially successful candidate format for acoustical heritage preservation, as early as 1975 [Gerzon, 1975b].

The increase in popularity of ambisonics since the beginning of the present century has turned this idea into reality; OpenAIRlib, a freely accesible dataset that gathers dozens of RIRs, might be a good example of it [Murphy and Shelley, 2010, OpenAIRlib, 2010].

In any case, the usage of recorded ambisonic RIRs is not limited to the field of acoustic heritage. Among others, the availability of such recordings has powered works in a variety of works, auralization [Postma et al., 2016], room acoustics analysis [Embrechts, 2015, Clapp et al., 2011] and modelling [Romblom, 2017], spatial audio synthesis [Coleman et al., 2017] or source separation [Baqué et al., 2016].

In general, all publicly available ambisonic RIR measurements share some common approaches for describing and organizing the recorded data. For instance, recordings from different rooms are usually stored as separated folders. Each combination of emitter and receiver positions is often saved as an individual file, and the different spherical harmonics match the audio channels. Moreover, it is also usual to provide a *metadata* file, describing the different emitter and receiver positions, and potentially some information about the measurement setup, methodology, etc. Such files might be formatted as plain text or delimiter-separated files.

Despite the common approach, it can be easily foreseen that each database generated by a different individual or institution might potentially have a different naming convention, folder structure, file format, and so on. This is exactly the same situation that motivated the development of the SOFA conventions.

On the other hand, the SOFA specification defines some criteria that must be fulfilled in order to propose a new convention [SOFA, 2018]. These criteria are:

1. Data must exist.

2. Data can not be described by existing SOFA conventions.

3. Relevant information about the data must be available.

Given that the described situation meets all requirements, the *Ambisonics Directional Room Impulse Response* (AmbisonicsDRIR) convention has been therefore proposed as SOFA convention [Pérez-López, 2018c].

The technical specifications of the proposed convention in its current state (version 0.2) are available online [Pérez-López and de Muynke, 2018].

6.3.3 Pysofaconventions

The situation described in Section , regarding the availability of acoustic signal processing libraries in the Python programming language, can be easily extended to the case of SOFA APIs.

The library *pysofaconventions* has been created with the aim to provide an alternative to the existing Matlab/Octave and C/C++ implementations [Pérez-López, 2020b]. For ease of installation, it is integrated in the standard python package manager, *pip*.

The current software version is 0.1.5. The library structure is inspired by the C++ implementation [Carpentier, 2018]. It features all functionalities described by SOFA version 1.0, plus the proposed AmbisonicsDRIR convention. The implementation is based on extensive error checking, to ensure code consistency.

6.4 Ambiscaper

6.4.1 Motivation

The availability of data is a fundamental requirement for research. More specifically, in the scope of the B-Format data analysis in which this thesis focuses, audio is frequently generated on an IR-based manner, employing both acoustic simulations and recordings. In that way, the analysis can focus on different signal types, such as speech or music, usually employing dedicated external datasets.

An alternative procedure for ambisonic data generation is the actual recording of sound scenes with spherical microphone arrays. Although this procedure might yield the most realistic sound field representations, the high cost, lack of scene control, and technical difficulty to obtain reliable groundtruth annotations lead to a limited usage of the technique, often reserved for real-life algorithm validation.

It is of interest to have an insight of the diverse methods used in the literature for ambisonic sound generation. Table 6.2 summarizes the information gathered from works on the scope of sound source localization and separation, which have been published until 2018[1]. The table displays, for each article, how the evaluation data was generated, and which was the type of audio content considered.

The statistics of the global usage of each generation method show that there is not a clear tendency towards any method. Nevertheless, it is also noticeable that the methods with

[1]The original research to which this section refers was conducted in 2018, for the publication of [Pérez-López, 2018a]. Since then, the outbreak of localization-related challenges, such as LOCATA 2018 and specially DCASE 2019 and 2020, and the consolidation of data-based methods, has largely contributed to the homogenization of datasets [Evers et al., 2020, Adavanne et al., 2019a, Politis et al., 2020]. Still, most assumptions and results of our analysis continue to hold at present.

Article	Data Generation Method	Audio Content
[Thiergart et al., 2009]	Audio recording	Speech
[Tervo, 2009]	Audio recording	Noise, music
[Jarrett et al., 2010]	IR simulation, audio recording	Noise
[Nadiri and Rafaely, 2014]	IR simulation, audio recording	Speech
[Moore et al., 2015]	IR simulation	Speech
[Pavlidi et al., 2015]	IR simulation	Noise, speech
[He and Chen, 2017]	IR simulation, audio recording	Speech
[Gunel et al., 2008]	IR recording	Speech, music
[Shujau et al., 2011]	Audio recording	Speech
[Riaz, 2015]	IR recording	Speech, music
[Chen et al., 2015]	IR simulation, IR recording	Speech

Table 6.2: Summary of audio data used across Ambisonics-based Source Localization (above) and Source Separation (below) methods.

evaluations performed on real recordings make use of only one one sound scene in each case. In contrast, when using IR-based scenes, the number of audios evaluated are usually one or two magnitude orders greater.

In addition, it is important to notice the lack of availability of the generated data. None of the analyzed articles provide a way to access neither the used audio dataset, nor the groundtruth (position annotations in the case of localization, and original sound sources for sound separation). Only when simulated IRs are used, it is possible to partially replicate the experimental setup — the parameters used in the simulation software are usually provided. Furthermore, the process of dataset creation seems to be performed *ad-hoc* in each case.

Taking into account the flexibility offered by IR-based scenes, it would be desirable to have a tool for automatic generation of ambisonics scenes, and their associated annotations, for analysis purposes. A tool with such characteristics would help the

scientific community in several ways: (i) reducing the amount of time dedicated to build custom datasets, (ii) reusing publicly available resources and recordings, and (iii) enhancing experiment reproducibility by making easier the exchange of datasets. Moreover, the capacity of batch processing of large ambisonic collections might also contribute to the development of data-driven approaches.

6.4.2 Implementation

AmbiScaper [Pérez-López, 2018a] is a tool designed to provide a flexible way of creating complex ambisonics sound scenes and their associated groundtruth annotations, to be used in the context of source localization and separation algorithms. *AmbiScaper* offers a high level control of the sound scene parameters, and provides a simple interface for the creation of large datasets with custom characteristics. *AmbiScaper* is based on *Scaper*, a framework designed to generate annotations for training Sound Event Detection models [Salamon et al., 2017].

One of the main features of *AmbiScaper* is that all parameters of the sound scene can be specified in a non-deterministic way. In that sense, the parameters for each *event* (sound source) are actually generated through a two-step process. First, in the *Event Specification*, all parameters related to an event are defined in terms of statistical distributions. During the *Event Instantiation*, the actual values for each parameter are then sampled from the statistical distributions. This two-step process allows the user to describe abstract *templates* of sound scenes, rather than manually assigning values to parameters. Therefore, a single *event specification* might produce potentially infinite different sound scenes.

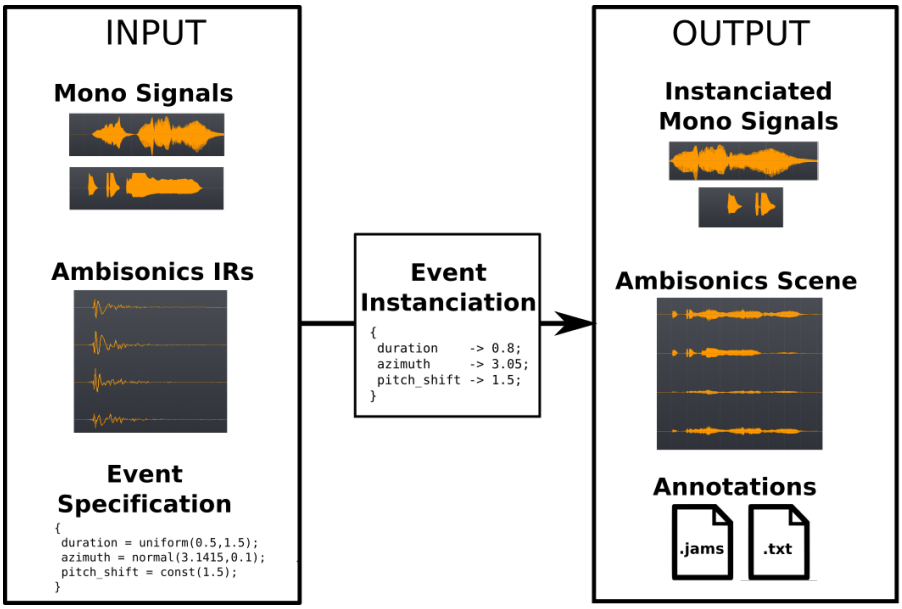

Figure 6.4: *AmbiScaper* architecture.

In order to generate a sound scene, *AmbiScaper* requires three different inputs: the original *mono signals*, which will provide the actual audio content; an optional ambisonic RIR; and an *event specification*.

The process of dataset creation starts with the *event instantiation*, as described above. Once all values are sampled, three different types of output are generated: the *ambisonic scene*; the *instantiated mono signals* (the original mono signals after data augmentation); and the *annotations*, in the form of a *jams* file [Humphrey et al., 2014], containing all information about the instantiated values.

The complete architecture of *AmbiScaper* architecture is depicted in Figure 6.4.

When no reverberation is specified, *AmbiScaper* can generate anechoic sound scenes, using the theoretical expression of the ambisonic encoding as in Eq. 2.12. In this case, there is no upper limit on the Ambisonics order of the rendered scene. Furthermore, the anechoic case allows the modification of the directivity of the source(s) through ambisonic order downgrade [Carpentier, 2017].

AmbiScaper partially supports RIR-based scene creation. It features a limited set of recorded ambisonic IRs from the S3A database [Coleman et al., 2015]; compatibility with the AmbisonicsDRIR (Subsection 6.3.2) could be easily included. On the other hand, the usage of simulated IRs is provided through a wrapper to the Matlab library *SMIR Generator* [Jarrett et al., 2012]. In this case, the reverberation model specifications are defined as well in statistical terms, and the generated RIRs are also stored for evaluation purposes.

6.4.3 Experiment reproducibility

As already mentioned, there is a generalized lack of publicly available datasets of ambisonic reverberant sound scenes. Even when using general purpose audio/speech datasets, the actual evaluation data is usually not available. In that sense, the potential compatibility of *AmbiScaper* with public ambisonic RIR databases is a key aspect for reproducibility, since it would allow the systematic reutilization of acoustical measurements in the analysis context.

Furthermore, the output of the *AmbiScaper* dataset generation process is not limited to the actual dataset. In fact, the resulting annotation file does not only contain the *instantiation* (the actual values of each parameter in the sound scene), but also the *specification* (the statistical distributions from which the instantiated values are sampled). In the scope of experiment reproducibility, the exchange of *specification files* instead of actual

audio files might greatly reduce the storage capacity and bandwidth required to transfer large databases.

AmbiScaper is implemented in the form of a Python package, and is publicly available through the *Python Package Index* repository under the GPL license, thus easing the software adoption and the potential engagement of the scientific community with the development.

As an example of the potential capabilities of the software, a sample dataset for the evaluation of source localization algorithms has been created [Pérez-López, 2018b]. The dataset contains 300 FOA sound scenes, with a duration between 1 and 2 seconds, each one containing a number of static sound sources between one and three. Sound sources might have different gains, and they are located at random positions around the sphere. The sources are randomly chosen from a subset of the Anechoic OpenAirlib database [OpenAIRlib, 2010], which mostly contains recordings from baroque musical instruments. Reverberation makes use of the *AudioBooth* RIR from S3A database [Coleman et al., 2015].

6.5 Conclusion

The present Chapter has presented the main contributions of the Thesis regarding the availability of ambisonic data, specifically targeting software packages and libraries developed for the generation and storage of ambisonic sound scenes.

The concept of *research reproducibility* is directly related with these topics. According to [Cannam et al., 2012], providing just an article as a research outcome would not be sufficient for reproducing the research in a feasible way. Instead, supplementary material, such as the code implementation, the dataset, and instructions on how to set up the environment, should be also conveniently provided.

In this regard, the software discussed in this Chapter try to foster reproducible methods and procedures in several different ways. First, all code is freely available online with open source licenses. Second, the contributed software is completely written in Python, which is itself an open-source project. Actually, all code used in this follows the same path, with with the exception of the particle tracker used in Chapter 5. And third, the motivation of the Ambiscaper project itself is the reproducibility of experimental conditions for signal processing methods using ambisonic data, as explained in detail in Section 6.4.3.

Regarding the lifecycle of the libraries described in this Chapter, their status is diverse. Although *pysofaconventions* has been recently published in an academic context (AES Convention in May 2020), it has been available online for more than a year. In fact, during that time it has gained a moderate amount of attraction: among others, it is being used as a dependency for the *Real-Time Spherical Microphone Renderer (ReTiSAR) for binaural reproduction*, developed by Chalmers University in collaboration with Facebook Reality Labs [Helmholz et al., 2019]. Furthermore, given the open nature of the project, several external researchers have contributed with bug fixes or new feature implementations.

The *masp* library has been also presented in the same context. However, in this case, the library has not been available beforehand, so the interest generated in the community has been moderate.

For its part, *Ambiscaper* has not been able to attract the community in a significant way. This might be due to a number of reasons, including not enough dissemination among potentially interested communities, or a lack of a paradigmatic use-case situation. On the contrary, the library that inspired this work, *Scaper*, has been used for the creation of the dataset in Task 4 of DCASE 2019 and 2020 (sound event detection in domestic environments). This situation generates a feedback loop for the maintenance and improvement of the software, motivated by

requests from the community. Given this situation, the future of *Ambiscaper* is linked to the evolution of the research topics and tools in the communities which miggt be potentially interested in using it.

Chapter 7

Conclusions

7.1 Summary of Contributions

In this thesis we have presented our contributions to different components of an ambisonics analysis and generation framework, with a focus on reproducibility and portability to real-world scenarios. The main scientific objectives of this thesis, as they were described in Section 1.3, are:

1. The development of methods to support and improve the characterization of acoustic parameters.
2. The research on parametric-based methodologies for sound event localization and detection.
3. The contribution in the generation and storage of annotated ambisonic datasets.

In what follows, we summarize the main contributions of the present thesis, both in the academic and software scopes.

Blind reverberation time estimation Chapter 3 presents a novel methodology for the blind estimation of reverberation time from ambisonic audio. The method is based on two main steps: first, dereverberation using a multichannel auto-recursive model (MAR), and second, estimation of the filter from reverberant and dereverberated signals. The actual reverberation time value is estimated from the energy decay of the estimated filter.

The proposed system is the first attempt in the literature to address the blind reverberation time estimation problem specifically for ambisonic signals. Compared with a state-of-the-art monophonic estimator, our method is able to improve in all the evaluation metrics under consideration.

Coherence estimation In Chapter 4, we have characterised the response of tetrahedral microphones to isotropic noise field, which is one of the most used models for diffuse sound. Furthermore, the capabilities of a spherical loudspeaker array with respect to the reconstruction of diffuse sound fields using ambisonics are also quantitatively analyzed.

Sound event localization and detection Chapter 5 describes an algorithm for sound event localization and detection (SELD), developed in the context of the DCASE 2020 challenge. The method estimates the localization and temporal activity of the sound events based on a particle filter that tracks event trajectories obtained from the parametric analysis of the ambisonic sound field. Each event is assigned to a sound class by a machine learning classifier that uses low- and mid- level audio features.

Results show a significant performance increase in all evaluation metrics under consideration, compared with a state-of-the-art deep learning baseline. This suggests that our approach, substantially different to the baseline and the majority of state-of-the-art methods, represents a feasible alternative in situations with low-complexity or small database constraints.

Data generation and management Finally, the thesis contributions to more practical aspects are presented in Chapter 6. Those contributions comprise two software libraries written in Python: one of them focused on spherical microphone array and acoustic simulation, and another one implementing the SOFA standard, which has also been revised and modified for allowing the representation of ambisonic data. Finally, a novel software tool for the procedural creation of annotated reverberant ambisonic datasets has ben also presented.

7.2 Future Work

This thesis has tackled several research problems associated with the analysis of ambisonic recordings, making use primarily of the parametric sound field modelling analysis. The presented techniques improve existing state-of-the-art methodologies, or present novel approaches for known research problems, which in turn bring new research questions.

A novel blind reverberation time estimation method for first-order ambisonic recordings is introduced in Chapter 3. Among others, the method could be straightforwardly extended to higher order ambisonic signals, which might still improve the reported results due to the availability of many more audio channels. Moreover, the usage of online MAR methods would enable the possibility of analyzing sound scenes with moving sources; the statistical time-invariance property of late reverberation supports this hypothesis. Finally, it is important to mention that the proposed method is resource-intensive. An analysis of the trade-off between computation time and evaluation performance, mostly dependent on the estimation filter length, remains to be done.

Given the current interest in the field of augmented reality, new related research topics emerge. One of them, which has been recently baptized as *acoustic matching* [Su et al., 2020], deals with the analysis of acoustic properties of real enclosures, with the aim of later introduction of virtual elements whose reverberation would match real conditions. The application of our method to the acoustic matching problem is straightforward: given an ambisonic recording with a target reverberation, estimate its reverberation time and synthesize a reverberant tail with the target energy decay; early reflections might be generated by various methods, including physical models or perceptually motivated approaches. We can foresee a growing interest on the topic in the near future; our contribution might help to establish the foundation of a new family of methods.

Regarding the diffuse field characterization performed in Chapter 4, an immediate extension of the work would include the study of different spherical microphone array geometries, from the ones that are commercially available. The usage of different models of diffuse fields, specially extending to the anisotropic case [Alary et al., 2019], might also constitute an interesting research continuation direction. Both cases could be also applied to the experiment of diffuse sound field reconstruction using loudspeaker arrays.

The wide scope of the SELD problem, as described in Chapter 5, allows for a wide range of possibilities regarding a potential follow-up of the proposed method. For instance, one of the major problems of our algorithm is the inaccurate parametric estimation when two events are simultaneously active. Although it is a known problem in the literature [Epain and Jin, 2016] , a successful solution in the given context remains still to be explored.

Another source of potential improvements is the refinement of the particle filter applied to this specific task. A deeper understanding of control theory, as well as collaborations with experts on the field, might bring a noticeable improvement on the overall system performance.

The performance of the event classifier might be also further analyzed. Although in this case we opted for a classical feature-based machine learning approach, different methods and architectures could be compared, including more modern deep-learning approaches.

Finally, we discuss briefly on the practical contributions described in Chapter 6. Apart from the straightforward task of software maintenance, the engagement of the research community with the usage and eventual contribution of the libraries would represent a desired situation in the near future. Moreover, the proposed file format convention is being currently discussed by the *Standardisation Committee on AES-69 Standard*, which can be considered a great achievement of the original proposal. The results of the discussion might lead to the addition of a modified version of our proposal into version 2 of the *AES-69 Standard*.

In the medium term, the application at commercial level of some of the methods described in this thesis would be a highly desirable outcome of our research. Such application would probably imply a re-implementation at the software level, intended for compatibility with the workflows and formats used in the VR/AR content production industry.

7.3 List of Contributions

In the following list we show the main scientific contributions, as main author, related to this dissertation:

- Peer-reviewed journal articles

  "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.* [Pérez-López and Stefanakis, 2019].

- Peer-reviewed conference articles

  "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.* [Pérez-López et al., 2020].

  "PAPAFIL: a low complexity sound event localization and detection method with parametric particle filtering and gradient boosting". A. Pérez-López and R. Ibañez-Usach. Submitted to *Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)* [Pérez-López and Ibanez-Usach, 2020].

  "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)* [Pérez-López et al., 2019].

  "Ambiscaper: A tool for automatic generation and annotation of reverberant ambisonics sound scenes". A. Pérez-López. In *Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, 2018* [Pérez-López, 2018a].

- Conference engineering briefs

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

  "pysofaconventions, a Python API for SOFA". A. Pérez-López. In *Proceedings of the 148th Audio Engineering Society Convention. Audio Engineering Society, 2020* [Pérez-López, 2020b].

  "A Python library for multichannel acoustic signal processing". A. Pérez-López and A. Politis. In *Proceedings of the 148th Audio Engineering Society Convention. Audio Engineering Society, 2020* [Pérez-López and Politis, 2020].

Moreover, although not strictly aligned with the research direction of this thesis, the author has supervised the following publication:

"Sound event localization and detection based on CRNN using dense rectangular filters and channel rotation data augmentation". F. Ronchini, D. Arteaga and A. Pérez-López. Submitted to *Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)* [**?**].

7.4 List of Software Resources

As a result of the development of this thesis, the following open-source libraries and repositories have been produced. All of them are freely available through the author's GitHub page [Pérez-López, 2020a] under open-source licenses.

- Software tools:

  masp: Multichannel Acoustic Signal Processing library Tools for acoustic simulation and spherical array processing.

  pysofaconventions Implementation of the SOFA convention in Python.

  AmbisonicsDRIR Ambisonic SOFA Convention proposal.

  AmbiScaper Tool for automatic generation of annotated ambisonic datasets.

- Method implementations:

  ambisonic_rt_estimation Contains the code implementing the methods described in Chapter 3.

  DCASE2020 Full code implementing the SELD algorithm from Chapter 5.

  DCASE2019_task3 Implementation of the method submitted to DCASE2019 (not included in this thesis).

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*, 13(1):34–48.

[Adavanne et al., 2019a] Adavanne, S., Politis, A., and Virtanen, T. (2019a). Aamulti-room reverberant dataset for sound event localization and detection. In *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, page 10.

[Adavanne et al., 2019b] Adavanne, S., Politis, A., and Virtanen, T. (2019b). Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network. *arXiv preprint arXiv:1904.12769*.

[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.

[Alary et al., 2019] Alary, B., Massé, P., Välimäki, V., and Noisternig, M. (2019). Assessing the anisotropic features of spatial impulse responses. In *EAA Spatial Audio Signal Processing Symposium*.

[Allen and Berkley, 1979] Allen, J. and Berkley, D. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

[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.

[AV_INFO, 1995] AV_INFO (1995). Reverberation time. http://www.bnoack.com/. Accessed on June 26th, 2020.

[Baqué et al., 2016] Baqué, M., Guérin, A., and Melon, M. (2016). Separation of direct sounds from early reflections using the entropy rate bound minimization algorithm. In *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. Audio Engineering Society.

[BBCNews, 2016] BBCNews (2016). Mwc 2016: Zuckerberg takes samsung into a vr future. https://www.bbc.com/news/technology-35629322. Accessed on June 11th, 2020.

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

[Bertet et al., 2006] Bertet, S., Daniel, J., and Moreau, S. (2006). 3D Sound Field Recording With Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone. In *Proc. 120th AES Convention*, pages 1–24, Paris, France.

[Bogdanov et al., 2013] Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., Serra, X., et al. (2013). Essentia: An audio analysis library for music information retrieval. In *14th Conference*

of the International Society for Music Information Retrieval (ISMIR); p. 493-8., Curitiba, Brazil.

[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 *Proc. IEEE ICASSP*, Barcelona, Spain. IEEE.

[Cannam et al., 2012] Cannam, C., Figueira, L. A., and Plumbley, M. D. (2012). Sound software: Towards software reuse in audio and music research. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2745–2748. IEEE.

[Cao et al., 2019] Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W., and Plumbley, M. (2019). Polyphonic sound event detection and localization using a two-stage strategy. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 30–34, New York University, NY, USA.

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

[Carpentier, 2018] Carpentier, T. (2018). Api_cpp. https://github.com/sofacoustics/API_Cpp. Accessed on June 26th, 2020.

[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.

[Chen et al., 2015] Chen, X., Wang, W., Wang, Y., Zhong, X., and Alinaghi, A. (2015). Reverberant speech separation with probabilistic time-frequency masking for b-format recordings. *Speech Communication*, 68:41–54.

[Clapp et al., 2011] Clapp, S. W., Guthrie, A. E., Braasch, J., and Xiang, N. (2011). Investigations of room acoustics with a spherical microphone array. In *Audio Engineering Society Convention 131*. Audio Engineering Society.

[Coleman et al., 2017] Coleman, P., Franck, A., Menzies, D., and Jackson, P. J. (2017). Object-based reverberation encoding from first-order ambisonic rirs. In *Audio Engineering Society Convention 142*. Audio Engineering Society.

[Coleman et al., 2015] Coleman, P., Remaggi, L., and Jackson, P. J. B. (2015). S3a room impulse responses [dataset]. http://epubs.surrey.ac.uk/808465/. Accessed on June 26th, 2020.

[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.

[DCASE, 2013] DCASE (2013). Research on detection and classification of acoustic secenes and events. http://dcase.community/community_info. Accessed on June 27th, 2020.

[DCASE, 2020] DCASE (2020). Sound event localization and detection, task description. http://dcase.community/challenge2020/task-sound-event-localization-and-detection-results. Accessed on July 2th, 2020.

[Drossos et al., 2017] Drossos, K., Adavanne, S., and Virtanen, T. (2017). Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 374–378. IEEE.

[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.

[Elliott, 2019] Elliott, T. (2019). The state of the octoverse: machine learning. https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/. Accessed on June 26th, 2020.

[Embrechts, 2015] Embrechts, J.-J. (2015). Measurement of 3d room impulse responses with a spherical microphone array. In *Proceedings of the Euronoise 2015 Congress*, pages 143–148.

[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.

[Evers et al., 2020] Evers, C., Loellmann, H., Mellmann, H., Schmidt, A., Barfuss, H., Naylor, P. A., and Kellermann, W. (2020). The locata challenge: Acoustic source localization and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[Facebook, 2014] Facebook (2014). Facebook to acquire oculus. https://about.fb.com/news/2014/03/facebook-to-acquire-oculus/. Accessed on June 11th, 2020.

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

[Fortune, 2019] Fortune (2019). The fall and rise of vr: The struggle to make virtual reality get real. https://fortune.com/longform/virtual-reality-struggle-hope-vr/. Accessed on June 11th, 2020.

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

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

[Gerzon, 1975b] Gerzon, M. A. (1975b). Recording concert hall acoustics for posterity. *Journal of the Audio Engineering Society*, 23(7):569–571.

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

[Grondin et al., 2019] Grondin, F., Glass, F., Sobieraj, I., and Plumbley, M. D. (2019). Sound event localization and detection using crnn on pairs of microphones. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 84–88, New York University, NY, USA.

[Gunel et al., 2008] Gunel, B., Hachabiboglu, H., and Kondoz, A. M. (2008). Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):748–756.

[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 et al., 2006] 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.

[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 and Chen, 2017] He, S. and Chen, H. (2017). Closed-form doa estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing. *IEEE Sensors Journal*, 17(4):1046–1060.

[Helmholz et al., 2019] Helmholz, H., Andersson, C., and Ahrens, J. (2019). Realtime implementation of binaural rendering of highorder spherical microphone array signals. *Fortschritte der AkustikDAGA 2019*.

[Humphrey et al., 2014] Humphrey, E. J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R. M., and Bello, J. P. (2014). Jams: A json annotated music specification for reproducible mir research. In *15th International Society for Music Information Retrieval Conference*, pages 591–596. ISMIR.

[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.

[Jarrett et al., 2010] Jarrett, D. P., Habets, E. A. P., and Naylor, P. A. (2010). 3d source localization in the spherical harmonic domain using a pseudointensity vector. In *18th European Signal Processing Conference*, pages 442–446. IEEE.

[Jarrett et al., 2012] Jarrett, D. P., Habets, E. A. P., Thomas, M. R. P., and Naylor, P. A. (2012). Rigid sphere room impulse

response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America*, 132(3):1462–1472.

[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.

[Kapka and Lewandowski, 2019] Kapka, S. and Lewandowski, M. (2019). Sound source detection, localization and classification using consecutive ensemble of crnn models. *arXiv preprint arXiv:1908.00766*.

[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 algorithms 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.

[Madmoni and Rafaely, 2018] Madmoni, L. and Rafaely, B. (2018). Direction of arrival estimation for reverberant speech based on

enhanced decomposition of the direct sound. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):131–142.

[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.

[Mesaros et al., 2019] Mesaros, A., Adavanne, S., Politis, A., Heittola, T., and Virtanen, T. (2019). Joint measurement of localization and detection of sound events. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA.

[Mesaros et al., 2018] Mesaros, A., Heittola, T., and Virtanen, T. (2018). A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Detection and Classification of*

Acoustic Scenes and Events 2018 Workshop (DCASE2018), pages 9–13.

[Moore et al., 2015] Moore, A. H., Evers, C., Naylor, P. A., Alon, D. L., and Rafaely, B. (2015). Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. In *23th European Signal Processing Conference*, pages 2296–2300. IEEE.

[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.

[Murphy and Shelley, 2010] Murphy, D. T. and Shelley, S. (2010). Openair: An interactive auralization web resource and database. In *129th Audio Engineering Convention*. AES.

[Nadiri and Rafaely, 2014] Nadiri, O. and Rafaely, B. (2014). Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1494–1505.

[Nguyen et al., 2020a] Nguyen, T. N. T., Jones, D. L., and Gan, W. S. (2020a). Dcase 2020 task 3: Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking. Technical report, DCASE2020 Challenge.

[Nguyen et al., 2020b] Nguyen, T. N. T., Jones, D. L., and Gan, W.-S. (2020b). A sequence matching network for polyphonic sound event localization and detection. In *Proc. IEEE ICASSP*, pages 71–75. IEEE.

[Noisternig et al., 2003] Noisternig, M., Sontacchi, A., Musil, T., and Holdrich, R. (2003). A 3d ambisonic based binaural sound reproduction system. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society.

[OpenAIRlib, 2010] OpenAIRlib (2010). The open acoustic impulse response library. www.openairlib.net. Accessed on June 26th, 2020.

[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.

[Pavlidi et al., 2015] Pavlidi, D., Delikaris-Manias, S., Pulkki, V., and Mouchtaris, A. (2015). 3d localization of multiple sound sources with intensity vector estimates in single source zones. In *23th European Signal Processing Conference*, pages 1556–1560. IEEE.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Pérez-López, 2018a] Pérez-López, A. (2018a). Ambiscaper: A tool for automatic generation and annotation of reverberant ambisonics sound scenes. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–9. IEEE.

[Pérez-López, 2018b] Pérez-López, A. (2018b). Ambiscaper example dataset [dataset]. https://zenodo.org/record/1186907. Accessed on July 26th, 2020).

[Pérez-López, 2018c] Pérez-López, A. (2018c). Ambisonics directional room impulse response as a new convention of the spatially oriented format for acoustics. In *144th Audio Engineering Society Convention*. AES.

[Pérez-López, 2020a] Pérez-López, A. (2020a). Andres perez-lopez github page. https://github.com/andresperezlopez. Accessed on June 27th, 2020.

[Pérez-López, 2020b] Pérez-López, A. (2020b). pysofaconventions, a python api for sofa. In *Proceedings of the 148th Audio Engineering Society Convention*. Audio Engineering Society.

[Pérez-López and de Muynke, 2018] Pérez-López, A. and de Muynke, J. (2018). Ambisonicsdrir. https://github.com/andresperezlopez/AmbisonicsDRIR. Accessed on June 26th, 2020.

[Pérez-López et al., 2019] Pérez-López, A., Fonseca, E., and Serra, X. (2019). A hybrid parametric-deep learning approach for sound event localization and detection. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, pages 189–193, New York University, NY, USA.

[Pérez-López and Ibanez-Usach, 2020] Pérez-López, A. and Ibanez-Usach, R. (2020). Papafil: a low complexity sound event localization and detection method with parametric particle filtering and gradient boosting. In *Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020) (submitted)*.

[Pérez-López and Politis, 2020] Pérez-López, A. and Politis, A. (2020). A python library for multichannel acoustic signal processing. In *Proc. 148th Audio Engineering Society Convention*. Audio Engineering Society.

[Pérez-López et al., 2020] Pérez-López, A., Politis, A., and Gómez, E. (2020). Blind reverberation time estimation from ambisonic recordings. In *IEEE 22nd International Workshop on Multimedia Signal Processing, 2020 (submitted)*. IEEE.

[Pérez-López and Stefanakis, 2019] Pérez-López, A. and Stefanakis, N. (2019). Analysis of spherical isotropic noise fields with an a-format tetrahedral microphone. *The Journal of the Acoustical Society of America. 2019; 146: EL329.*

[Politis, 2016] Politis, A. (2016). *Microphone array processing for parametric spatial audio techniques*. PhD thesis, Aalto University.

[Politis, 2020] Politis, A. (2020). Github repository. https://github.com/polarch. Accessed on June 26th, 2020.

[Politis et al., 2020] Politis, A., Adavanne, S., and Virtanen, T. (2020). A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. *arXiv e-prints: 2006.01919*.

[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., 2018] Politis, A., Tervo, S., and Pulkki, V. (2018). 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.

[Postma et al., 2016] Postma, B. N., Poirier-Quinot, D., Meyer, J., and Katz, B. F. (2016). Virtual reality performance auralization in a calibrated model of notre-dame cathedral. *Euroregio*, 6:1–10.

[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, 1997] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, 45(6):456–466.

[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.

[PYPL, 2020] PYPL (2020). Pypl popularity of programming language. http://pypl.github.io/PYPL.html. Accessed on June 26th, 2020.

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

[Rafaely, 2015] Rafaely, B. (2015). *Fundamentals of spherical array processing*, volume 8. Springer.

[Riaz, 2015] Riaz, A. (2015). *Adaptive blind source separation based on intensity vector statistics*. PhD thesis, University of Surrey.

[Robinson, 2017] Robinson, D. (2017). The overflow: The incredible growth of python. https://stackoverflow.blog/2017/09/06/incredible-growth-python/. Accessed on June 26th, 2020.

[Romblom, 2017] Romblom, D. (2017). Diffuse field modeling using physically-inspired decorrelation filters: Improvements to the filter design method. *Journal of the Audio Engineering Society*, 65(11):943–953.

[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.

[Salamon et al., 2017] Salamon, J., MacConnell, D., Cartwright, M., Li, P., and Bello, J. P. (2017). Scaper: A library for soundscape synthesis and augmentation. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348. IEEE.

[Särkkä et al., 2004] Särkkä, S., Vehtari, A., and Lampinen, J. (2004). Rao-blackwellized monte carlo data association for

multiple target tracking. In *Proceedings of the Seventh International Conference on Information Fusion*, volume 1, pages 583–590, Stockholm, Sweden.

[Scheibler et al., 2018] Scheibler, R., Bezzam, E., and Dokmanić, I. (2018). Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355. IEEE.

[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.

[Sennheiser, 2020] Sennheiser (2020). Ambeo vr mic. https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic. Accessed on June 26th, 2020.

[Shujau et al., 2011] Shujau, M., Ritz, C. H., and Burnett, I. S. (2011). Separation of speech sources using an acoustic vector sensor. In *13th International Workshop on Multimedia Signal Processing*. IEEE.

[SOFA, 2018] SOFA (Accessed on March 12th, 2018). Sofa conventions. https://www.sofaconventions.org/mediawiki/index.php/SOFA_conventions.

[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.

[Su et al., 2020] Su, J., Jin, Z., and Finkelstein, A. (2020). Acoustic matching by embedding impulse responses. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 426–430. IEEE.

[Tervo, 2009] Tervo, S. (2009). Direction estimation based on sound intensity vectors. In *17th European Signal Processing Conference*, pages 700–704. IEEE.

[TheVerge, 2018] TheVerge (2018). Former vr film company jaunt is giving up on vr to focus on augmented reality. https://www.theverge.com/2018/10/15/17980420/jaunt-vr-layoffs-ar-focus-switch-restructuring-xr-platform. Accessed on June 11th, 2020.

[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.

[Tho et al., 2014] Tho, N. T. N., Zhao, S., and Jones, D. L. (2014). Robust doa estimation of multiple speech sources. In *Proc. IEEE ICASSP*, pages 2287–2291, Florence, Italy. IEEE.

[Time, 2015] Time (2015). The surprising joy of virtual reality. https://time.com/3987059/in-the-latest-issue-41/. Accessed on June 11th, 2020.

[TIOBE, 2020] TIOBE (2020). Tiobe index for june 2020. http://www.tiobe.com. Accessed on June 26th, 2020.

[Trowitzsch et al., 2019] Trowitzsch, I., Taghia, J., Kashef, Y., and Obermayer, K. (2019). The nigens general sound events database. *arXiv preprint arXiv:1902.08314*.

[Wikipedia, 2020] Wikipedia (2020). List of ambisonic hardware. https://en.wikipedia.org/wiki/List_of_Ambisonic_hardware. Accessed on June 11th, 2020.

[Zotter and Frank, 2012] Zotter, F. and Frank, M. (2012). All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820.

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