Spherical harmonic domain methods for 3D audio production enhancement (check capitalization)

subtitle Required?

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

Write here your dedication dedication

Thanks thanks to.... thanks

Abstract

This is the abstract of the thesis in English. Please, use less than 150 words. abstract in english

Resum

Vet aquí el resum de la tesi en català. Si us plau, utilitzeu menys de 150 paraules. abstract in catalan

Preface is that really needed?

Contents

List of figures xviii

List of tables xix

1 INTRODUCTION **1**
- 1.1 Motivation 1
  - 1.1.1 3D arrays 1
  - 1.1.2 spherical microphone arrays 1
  - 1.1.3 ambisonics 1
  - 1.1.4 Current limitations of vr/ar production . . . 2
- 1.2 Goals 2
- 1.3 Context 2

2 SCIENTIFIC BACKGROUND **5**
- 2.1 Reference system and conventions 5
- 2.2 Spherical Harmonics 6
  - 2.2.1 Definition 6
  - 2.2.2 Spherical array processing 9
- 2.3 Ambisonics 10
  - 2.3.1 Ambisonics Theory 10
  - 2.3.2 Practical considerations 13
- 2.4 Parametric Spatial Audio Analysis 16
- 2.5 Spatial Coherence Analysis 19
- 2.6 Room Acoustics, Impulse Responses 20

2.7 SOFA 20
2.8 Practical Considerations 21

3 COHERENCE ESTIMATION 23
3.1 Introduction 23
3.1.1 Problem definition 24
3.2 Methods 24
3.2.1 Simulation 24
3.2.2 Recording 25
3.2.3 Data processing and metrics 25
3.3 Results and discussion 26
3.3.1 A-Format 26
3.3.2 B-Format 27
3.4 Conclusions 30

4 AUTOREGRESSIVE IMPULSE RESPONSE MODELS 31

5 SOUND EVENT LOCALIZATION AND DETECTION 33
5.1 Introduction 33
5.2 Method 35
5.2.1 DOA estimation 35
5.2.2 Association 37
5.2.3 Beamforming 38
5.2.4 Deep learning classification back-end 39
5.3 Experiments 40
5.3.1 Dataset, evaluation metrics and baseline system 40
5.3.2 Parametric front-end 41
5.3.3 Deep learning classification back-end 43
5.4 Results and Discussion 44
5.5 Conclusion 47

6 DATA GENERATION AND STORAGE 49
6.1 Recorded IRs 49
6.1.1 Ambisonics Recording 50
6.1.2 Ambisonics DRIRs 50

6.1.3 The SOFA Conventions 50
6.1.4 Convention Proposal 52
6.1.5 Results 53
6.1.6 Summary 54
6.2 Simulated IRs 54
6.3 High-level scene description 54
6.3.1 Motivation 54
6.3.2 Evaluation data 56
6.3.3 AMBISCAPER 59
6.3.4 Sound scene description 59
6.3.5 Architecture 60
6.3.6 Reverberation 61
6.3.7 AmbiScaper and experiment reproducibility 62
6.3.8 Sample Dataset 62
6.3.9 CONCLUSIONS AND FUTURE WORK . . . 63

7 CONCLUSIONS **65**
7.1 Summary of Contributions 65
7.2 List of related publications 66

List of Figures

1.1 todo caption 3

2.1 Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values. 8

2.2 Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from T. Carpentier [cite]. 15

3.1 *A-Format* coherence between microphone signals. Left: MSC as a function of the frequency of theoretical, simulated and recorded *((BLD,BRU)*, $N = 5, I = 64$) signals. Right: mean error $\bar{\varepsilon}$ of the recorded signals' MSC *(BLD,BRU)* compared to the simulated values, for all values of N and I.REDO FIGURE WITH N AND I 26

3.2 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. 27

3.3 *B-Format coherence* between microphone signals. Left: Δ of simulated and recorded ($N = 5, I = 64$) signals. Right: $\bar{\varepsilon}$ of the recorded signals coherence across all values of N and I.REDO FIGURE 29

5.1 System architecture. 35
5.2 DOA estimation architecture. 35
5.3 Association architecture. 37
5.4 Back-end architecture. 41
5.5 DCASE2019 Challenge Task 3 results, evaluation set. 46

6.1 Source and Listener position diagram for the "York Guildhall Council Chamber" Ambisonics DRIR (original diagram attribution to [of York Council, 2018] . 55
6.2 AmbiScaper architecture. 60

List of Tables

2.1 Cartesian and spherical representation of characteristic points along the unit sphere. 6

5.1 Parameter values for the selected configuration. Top: *DOA analysis* parameters. Bottom: *Association* parameters. 42
5.2 Results for development (top) and evaluation (bottom) sets. 45

6.1 Summary of audio data used across Ambisonics-based Source Localization (above) and Source Separation (below) algorithm proposals. update table? fix references 58

Chapter 1

INTRODUCTION

1.1 Motivation

1.1.1 3D arrays

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

1.1.2 spherical microphone arrays

- even distribution of capsules
- mathematical convenience: spherical harmonics

1.1.3 ambisonics

advantages on the vr/ar context

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

1.1.4 Current limitations of vr/ar production

1.2 Goals

Research question: How can we exploit the characteristics of ambisonic recordings in order to manipulate them more adequately?

1.3 Context

Different levels of application/contribution:

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

introduce rest of chapters?

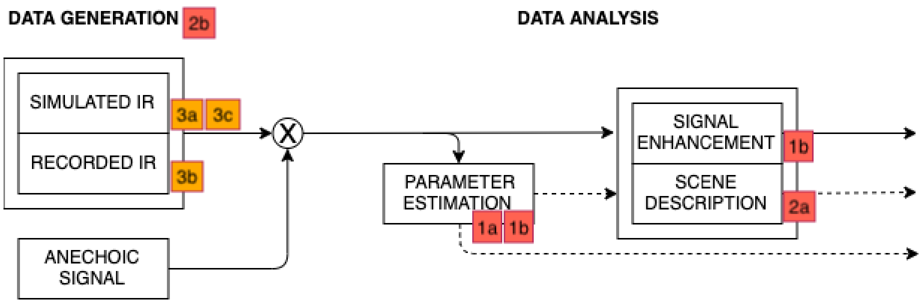


Figure 1.1: todo caption

Chapter 2

SCIENTIFIC BACKGROUND

2.1 Reference system and conventions

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

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

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

The transformation between spherical and cartesian coordinate

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

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

systems is given by the following relationship:

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

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

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

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

2.2 Spherical Harmonics

2.2.1 Definition

Spherical harmonics are continuous functions defined on the sphere surface. Due to their mathematical properties, any spherical function can be decomposed as a combination of spherical harmonics, in what is known as the *Spherical Harmonics Expansion* [Jarrett book, page 17].

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]:

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

where the *normalization factor* N_n^m is:

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

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

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

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

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

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

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

where δ_{xy} represents the Kronecker delta operator:

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

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

Through the spherical harmonic expansion, any sound field may be represented with a limited spatial resolution by the finite combination of all spherical harmonics up to order N. For a given order n, the number of spherical harmonic functions is $2n+1$. With the accumulation of all orders up to N, the total number of spherical harmonics is given by $M = (N+1)^2$. Figure 2.1 depicts all spherical harmonics from orders 0 to 3.

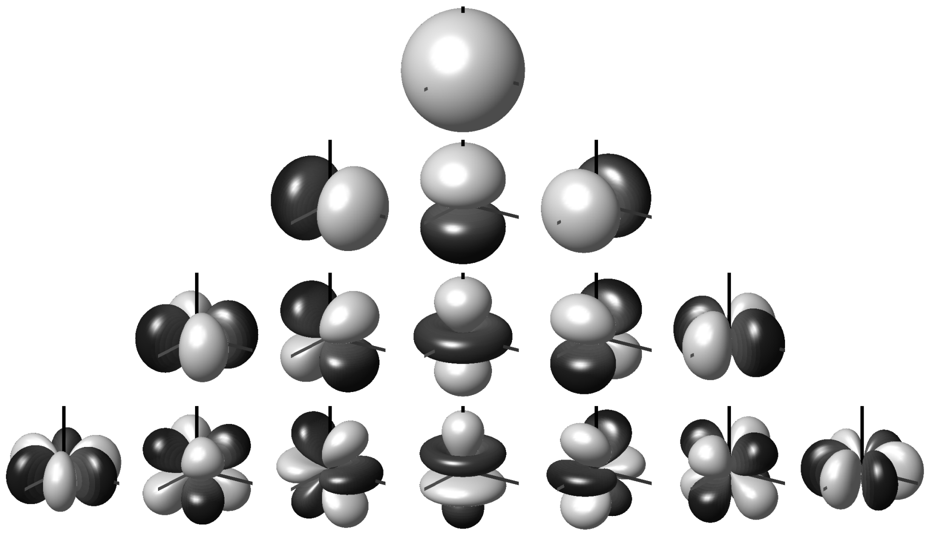


Figure 2.1: Spherical harmonics up to order $N = 3$. The rows correspond to the spherical harmonics of a given order n, and the columns span all possible degree values.

2.2.2 Spherical array processing

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

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

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

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

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

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

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

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

with c being the sound speed.

moreau has c/2Rgamma.

2.3 Ambisonics

2.3.1 Ambisonics Theory

Ambisonics is a spatial sound recording and playback technology initially developed during the 1970's, and further expanded into its modern formulation around the 2000s [ZOTTER, page 53]. 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, mainly due to practical limitations [CITE GERZON], as the so-called *First Order Ambisonics* (FOA). The technique was later formalized for arbitrary spherical harmonic orders, known as *Higher Order Ambisonics* or HOA [CITE DANIEL]. 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 can be expressed in terms of the spherical harmonic expansion of order N as: check equation, find references, how to explain the domain? extend also to multiple sources by superposition

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

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

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

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 $\boldsymbol{S}_n^m$:

$$\boldsymbol{S}_n^m = \boldsymbol{Y}_n^m(\Omega_s)S \tag{2.14}$$

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

Historically, the name *B-Format* was used as an opposite of *A-Format*, which describes the signals recorded by a tetrahedral microphone array [cite gerzon?]. The tetrahedron is the simplest and most common form of spherical microphone arrays (*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 [find ref] 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 [ambisonic book].

change L for other name Let us consider a loudspeaker located at the angular position Ω_ℓ. The signal feed L is *decoded* from the ambisonic signal as:

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

where α_n is a weighting factor which accounts for the beam directivity. Some of the weightings are widely used for its specific properties, such as *max-rE* or *in-phase* – the reader is referred to [find ref. daniel? zotter?] for more information.

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

$$L = \boldsymbol{S}_n^m \boldsymbol{Y}_n^m(\Omega_\ell)^T \alpha_n \tag{2.16}$$

where the superscript T represents the matrix transposition. This equation can be extended to the regular case of decoding to a loudspeaker array, comprised of L loudspeakers located at the positions $\boldsymbol{\Omega}_l = [\Omega_{\ell_1}, \ldots, \Omega_{\ell_L}]$. In such case, the loudspeaker feed vector $\boldsymbol{L}$ can be written as:

$$\boldsymbol{L} = \boldsymbol{S}_n^m \boldsymbol{D}, \tag{2.17}$$

where

$$\boldsymbol{D} = \text{diag}(\alpha_n)[\boldsymbol{Y}_n^m(\Omega_{\ell_1})^T, \ldots, \boldsymbol{Y}_n^m(\Omega_{\ell_L})^T] \tag{2.18}$$

is a $M \times L$ matrix known as the *decoding matrix*, and $\text{diag}(\alpha_n)$ is a diagonal matrix of size M containing the values of α_n along the main diagonal. $\boldsymbol{D}$ is frequency-independent and depends solely on the loudspeaker array geometry. The reader is referred to [zotter? daniel? idhoa?] for more information about the vast field of study of ambisonic decoding.

figure from daniel? to sum up the subsection). Explain the allrad method and stuff because it will appear later on.

2.3.2 Practical considerations

Due to historical and practical reasons, there are two aspects that must be taking into account when working with ambisonic signals: *channel normalization* and *channel ordering*. In the following, the term *channels* will be used as a synonym for spherical harmonics, as they are usually referred to in sound engineering contexts[1].

Channel normalization

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

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

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

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

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

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

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

[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.

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?], while the more modern *N3D* and *SN3D* were popularized after Daniel [daniel].

As a summary, Figure 2.2 displaysthe different normalization schemes. The reader is referred to [thibaut] for an extensive review on the topic.

Channel ordering

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

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

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

Historically, first-order ambisonic audio has followed what it might be called *traditional B-Format* channel ordering [ambisonics in multichannel broadcasting and video]. By following this scheme, the four channels are referred to by the axis where the corresponding spherical harmonic steers, plus the name W for the zeroth order component:

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

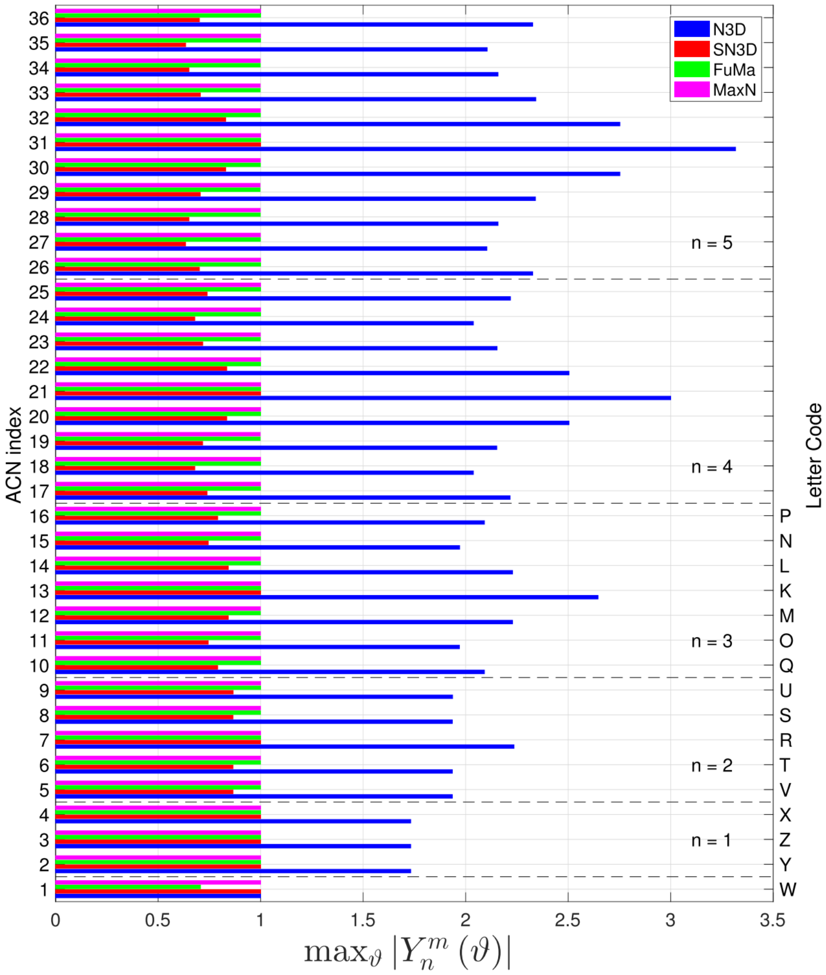


Figure 2.2: Maximum value of each ambisonic channel up to order 5, for all different normalization schemes. Image from T. Carpentier [cite].

with:

$$\begin{aligned} W &= Y_0^0(\Omega) \\ X &= Y_1^1(\Omega) \\ Y &= Y_1^{-1}(\Omega) \\ Z &= Y_1^0(\Omega) \end{aligned} \tag{2.24}$$

This nomenclature was extended to second and third order by Malham, 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 evident [Higher order Ambisonic systems]. Figure 2.2 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. In what follows, we will use indistinctly both *classical* and *ACN* conventions. is that true?

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 [book jarrett].

One of the most successful sound field parametric models is DirAC[pulkki2007 dirac]. Originally conceived as a method for impulse response processing and spatial sound reproduction [merima2005], it has been widely used in many different audio-related problems [find some references!].

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

DirAC (acronym for *Directional Audio Coding*) is a perceptually motivated time-frequency 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 arrivals (DOAs) [cite parametric book].

Let us consider a *N3D*-normalized first-order ambisonic signal in time-frequency domain $\boldsymbol{B}_n^m(k,n)$, begin k the frequency index and n the time index. For the sake of clarity, we will use *FuMa* channel notation and ordering:

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

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

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

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

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

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

The *active intensity* $\boldsymbol{I}(k,n)$, defined as the amount of transmitted acoustic energy, can be expressed in terms of sound pressure and velocity [fahy2002]:

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

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* $E(k,n)$ [stazial 1996]:

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

Finally, the *diffuseness* $\Psi(k,n)$ can be computed from the sound intensity and energy density [merimaa2005]:

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

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

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, pulkki, del galdo] and the more recent *COMEDIE* estimator [epain, craig].

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:

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

2.5 Spatial Coherence Analysis

put this chapter in context or something

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

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

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):

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

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

2.6 Room Acoustics, Impulse Responses

2.7 SOFA

maybe as subsection of room acoustics

2.8 Practical Considerations

Chapter 3

COHERENCE ESTIMATION

3.1 Introduction

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

In this paper, 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.

3.1.1 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* [Schorkhuber and Holdrich, 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.

3.2 Methods

3.2.1 Simulation

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

---

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

3.2.2 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.

3.2.3 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 $\varepsilon(k)$:

$$\varepsilon(k) = |X_1(k) - X_2(k)|^2, \tag{3.1}$$

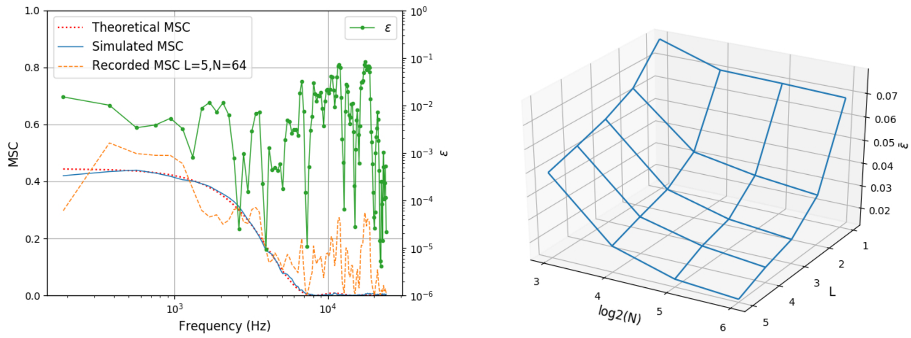


Figure 3.1: *A-Format* coherence between microphone signals. Left: MSC as a function of the frequency of theoretical, simulated and recorded *((BLD,BRU)*, $N = 5, I = 64$) signals. Right: mean error $\bar{\varepsilon}$ of the recorded signals' MSC *(BLD,BRU)* compared to the simulated values, for all values of N and I.REDO FIGURE WITH N AND I

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

$$\bar{\varepsilon} = \frac{1}{K}\sum_{k=1}^{K}|X_1(k) - X_2(k)|^2 \tag{3.2}$$

3.3 Results and discussion

3.3.1 A-Format

The coherence of the generated *A-Format* signals is exemplified in Fig. 3.1 (left), which shows the MSC between the capsule pair (*BLD,BRU*) for the theoretical, simulated and recorded cases. The theoretical coherence is derived from Eq. (2.36), while simulated and recorded MSC have been computed by Welch's method, using a *hanning* window of 256 samples and 1/2 overlap. The difference between theoretical and simulated coherence is negligible for practical applications. However, there is a noticeable difference when

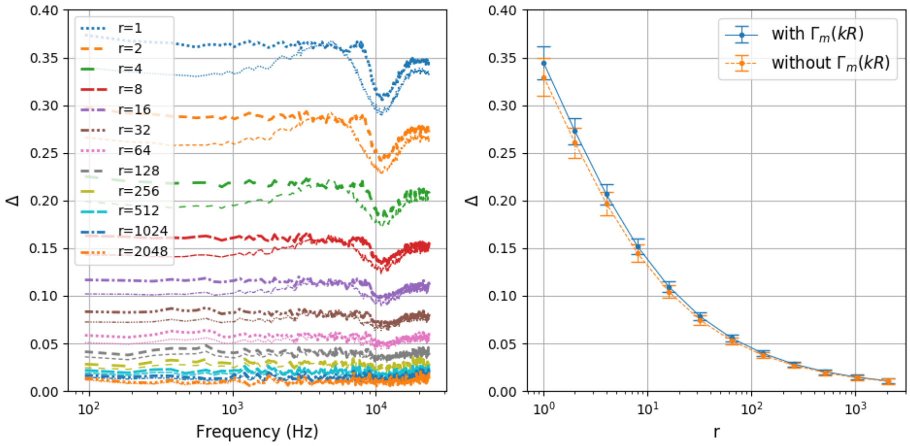


Figure 3.2: Estimated *B-Format* coherence (Δ) of a simulated diffuse sound field, as a function of the temporal averaging vicinity radius r. Left: $\Delta(k)$ for different values of r, with (coarse) and without (fine) application of radial filters. Right: mean and standard deviation of $\Delta(k)$ as a function of r.

compared to the recorded coherence. In general, the recorded MSC follows the tendency of the simulated curve up to around 5 kHz. Above this frequency, the recorded MSC presents several spectral peaks, which might be partially explained by the interference of the microphone itself in the recorded sound field, and by the non-ideal directivity of the capsules. The squared error $\varepsilon(k)$ with respect to the simulated curve is shown in Fig. 3.1 (left), while Fig. 3.1 (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.

3.3.2 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 3.3.1.

Figure 3.2 (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. 3.2 (right): the curve asymptotically decreases to a value $\Delta_{min} \approx 0$.

Another interesting observation comes from the frequency response of the curves. For all values of r, the coherence of the compensated *B-Format* signal (with $\Gamma_m(kR)$) is roughly flat up to around 7 kHz, which approximately corresponds to the operational spatial frequency range of the microphone [Gerzon, 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 [Schorkhuber and Holdrich, 2017].

As a remark, the comparison between Figs. 3.1 and 3.2 provides evidence that the application of the spherical harmonic transform

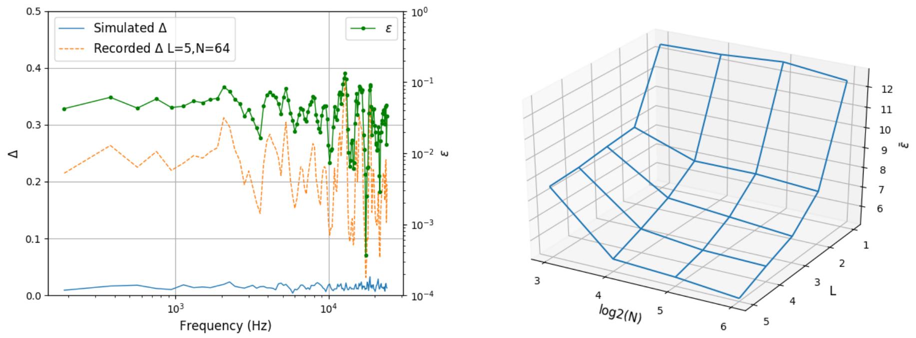


Figure 3.3: *B-Format coherence* between microphone signals. Left: Δ of simulated and recorded ($N = 5, I = 64$) signals. Right: $\bar{\varepsilon}$ of the recorded signals coherence across all values of N and I.REDO FIGURE

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

Figure 3.3 (left) shows the estimated coherence for the recorded sound field with $N = 5$ and $I = 64$, using a vicinity radius of $r = 1024$ ($\approx$ 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. 3.1.

This effect can be also appreciated in Fig. 3.3 (right): the mean squared error is around two orders of magnitude higher in *B-Format*. Nevertheless, similar as in Fig. 3.1 (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.

3.4 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 4

AUTOREGRESSIVE IMPULSE RESPONSE MODELS

Chapter 5

SOUND EVENT LOCALIZATION AND DETECTION

5.1 Introduction

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

SELD arises from the combination of two different problems: Sound Event Detection (SED) and Direction of Arrival (DOA) estimation. The number of works in the literature which jointly address SED and DOA problems is relatively small. It is possible to classify them by the type of microphone arrays used: distributed [Grobler et al., 2017, Butko et al., 2011, Chakraborty and Nadeu, 2014] or near-coincident [Hirvonen, 2015, Lopatka et al., 2016, Adavanne et al., 2018]. As mentioned in [Adavanne et al., 2018], the usage of near-coincident circular/spherical arrays enables the representation of the sound field in the spatial domain, using the spherical harmonic decomposition, also known as Ambisonics [Gerzon, 1973, Daniel, 2000]. Such spatial representation allows a flexible, device-independent comparison between methods. Furthermore, the number of commercially available ambisonic microphones has increased in recent years due to their suitability for immersive multimedia applications. Taking advantage of the compact spatial representation provided by the spherical harmonic decomposition, several methods for parametric analysis of the sound field in the ambisonic domain have been proposed [Pulkki, 2006b, Berge and Barrett, 2010, Politis et al., 2018b, Pulkki et al., 2018]. These methods ease sound field segmentation into direct and diffuse components, and further localization of the direct sounds. The advent of deep learning techniques for DOA estimation has also improved the results of traditional methods [Adavanne et al., 2018]. However, none of the deep learning-based DOA estimation methods explicitly exploits the spatial parametric analysis. This situation is further extended to the SELD problem, with the exception of [Lopatka et al., 2016], where DOAs are estimated from the *active intensity vector* [Pulkki, 2006b].

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

In what follows, we present the methodology and the architecture of the proposed system (Section 5.2). Then, we describe the design choices and the experimental setup (Section 5.3), and discuss the results in the context of DCASE2019 Challenge - Task 3 (Section 5.4). A summary is presented in Section 5.5. In order to support open access and reproducibility, all code is freely available at [cod,]. code ref.

5.2 Method

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

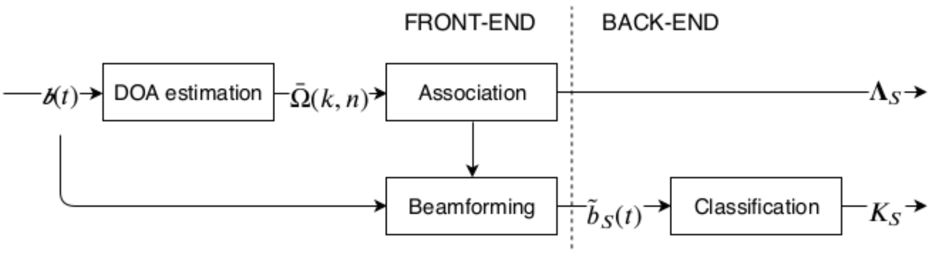


Figure 5.1: System architecture.

5.2.1 DOA estimation

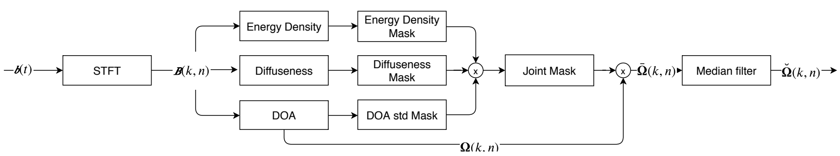


Figure 5.2: DOA estimation architecture.

The *DOA estimation* system (Figure 5.2) is based on parametric time-frequency (TF) spatial audio analysis. Let us consider a first-order ($L = 1$) ambisonic signal vector $\boldsymbol{b}(t)$ with N3D normalization [Carpentier, 2017b]:

$$\boldsymbol{b}(t) = [b_w(t), \sqrt{3}b_x(t), \sqrt{3}b_y(t), \sqrt{3}b_z(t)]. \tag{5.1}$$

From its short-time frequency domain representation $\boldsymbol{B}(k,n)$, the instantaneous DOA at each TF bin $\boldsymbol{\Omega}(k,n)$ can be estimated as:

$$\begin{aligned} \boldsymbol{I}(k,n) &= -\frac{1}{Z_0}\mathbb{R}\{[B_x(k,n), B_y(k,n), B_z(k,n)]B_w(k,n)^*\}, \\ \boldsymbol{\Omega}(k,n) &= [\varphi(k,n), \theta(k,n)] = \angle(-\boldsymbol{I}(k,n)), \end{aligned} \tag{5.2}$$

where $\boldsymbol{I}(k,n)$ stands for the *active intensity vector* [Pulkki, 2006b], Z_0 is the characteristic impedance of the medium, * represents the complex conjugate operator, and $\angle$ is the spherical coordinates angle operator, expressed in terms of azimuth φ and elevation θ.

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

The first mask is the *energy density mask*, which is used as an activity detector. The energy density $E(k,n)$ is defined as in [Stanzial et al., 1996] :

$$E(k,n) = \frac{|B_w(k,n)|^2 + ||[B_x(k,n), B_y(k,n), B_z(k,n)]||^2}{2Z_0c}, \tag{5.3}$$

with c being the sound speed. A gaussian adaptive thresholding algorithm is then applied to $E(k,n)$, which selects TF bins with local maximum energy density, as expected from direct sounds.

The *diffuseness mask* selects the TF bins with high energy propagation. Diffuseness $\Psi(k,n)$ is defined in [Merimaa and Pulkki, 2005] as:

$$\Psi(k,n) = 1 - ||\langle \boldsymbol{I}(k,n) \rangle||/(c\langle E(k,n) \rangle), \tag{5.4}$$

where $\langle \cdot \rangle$ represents the temporal expected value.

The third mask is the *DOA variance mask*. It tries to select TF regions with small standard deviation[1] with respect to their neighbor bins—a characteristic of sound fields with low diffuseness [Pulkki et al., 2018].

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

5.2.2 Association

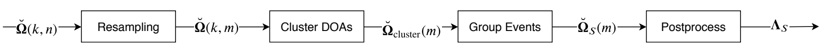


Figure 5.3: Association architecture.

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

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

Next, the standard deviation in azimuth (σ_φ) and elevation (σ_θ) of the frame-based DOA estimates $\breve{\mathbf{\Omega}}(k,m)$ are compared to

[1]In this work, all statistical operators for angular position refer to the 2π-*periodic* operator for azimuth, and the standard operator for elevation.

a threshold value (σ_{max}), and the result is used to estimate the frame-based event overlapping amount $o(m)$:

$$o(m) = \begin{cases} 1, & \text{if } \sigma_\varphi/2 + \sigma_\theta < \sigma_{max}, \\ 2, & \text{otherwise.} \end{cases} \tag{5.5}$$

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

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

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

The resulting DOAs $\mathbf{\Omega}_S(m)$ are subject to a postprocessing step with the purpose of delaying event onsets in frames where $o(m) > 2$, and discarding events shorter than a given minimum length. Finally, the frame-based event estimations are converted into *metadata annotations* in the form $\mathbf{\Lambda}_S = (\mathbf{\Omega}_S, \text{onset}_S, \text{offset}_S)$.

5.2.3 Beamforming

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

$$\tilde{b}_S(t) = \boldsymbol{Y}(\mathbf{\Omega}_S)\boldsymbol{b}^\intercal(t), \tag{5.6}$$

where $\boldsymbol{Y}(\boldsymbol{\Omega}_S) = [Y_w(\boldsymbol{\Omega}_S), Y_x(\boldsymbol{\Omega}_S), Y_y(\boldsymbol{\Omega}_S), Y_z(\boldsymbol{\Omega}_S)]$ is the set of real-valued spherical harmonics up to order $L = 1$ evaluated at $\boldsymbol{\Omega}_S$.

5.2.4 Deep learning classification back-end

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

The classification method is divided into two stages. First, the incoming signal is transformed into the log-mel spectrogram and split into TF patches. Then, the TF patches are fed into a single-mode based on a Convolutional Recurrent Neural Network (CRNN), which outputs probabilities for event classes $k \in \{1...K\}$, with $K = 11$. Then, we use a single model based on a Convolutional Recurrent Neural Network (CRNN), fed by the TF patches, and outputting probabilities for event classes $k \in \{1...K\}$, with $K = 11$. Predictions are done at the event-level (not at the frame level), since the temporal activities have been already determined by the front-end. Only one label is predicted for each incoming event, such that there is no binarization stage needed as in a standard SED task.

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

5.3 Experiments

5.3.1 Dataset, evaluation metrics and baseline system

We use the TAU Spatial Sound Events 2019 - Ambisonic, which provides first-order ambisonic recordings. Details about the recording format and dataset specifications can be found in [Adavanne et al., 2019]. The dataset features a vocabulary of 11 classes encompassing human sounds and sound events typically found in indoor office environments. The dataset is split into a development and evaluation sets. The development set consists of a four fold cross-validation setup.

The SELD task is evaluated with individual metrics for SED (F-score (*F*) and error rate (*ER*) calculated in one-second segments) and DOA estimation (DOA error (*DOA*) and frame recall (*FR*) calculated frame-wise) [Adavanne et al., 2018]. The *SELD score* is an averaged summary of the system performance.

The baseline system features a CRNN that jointly performs DOA and SED through multi-task learning [Adavanne et al., 2018]. Base-

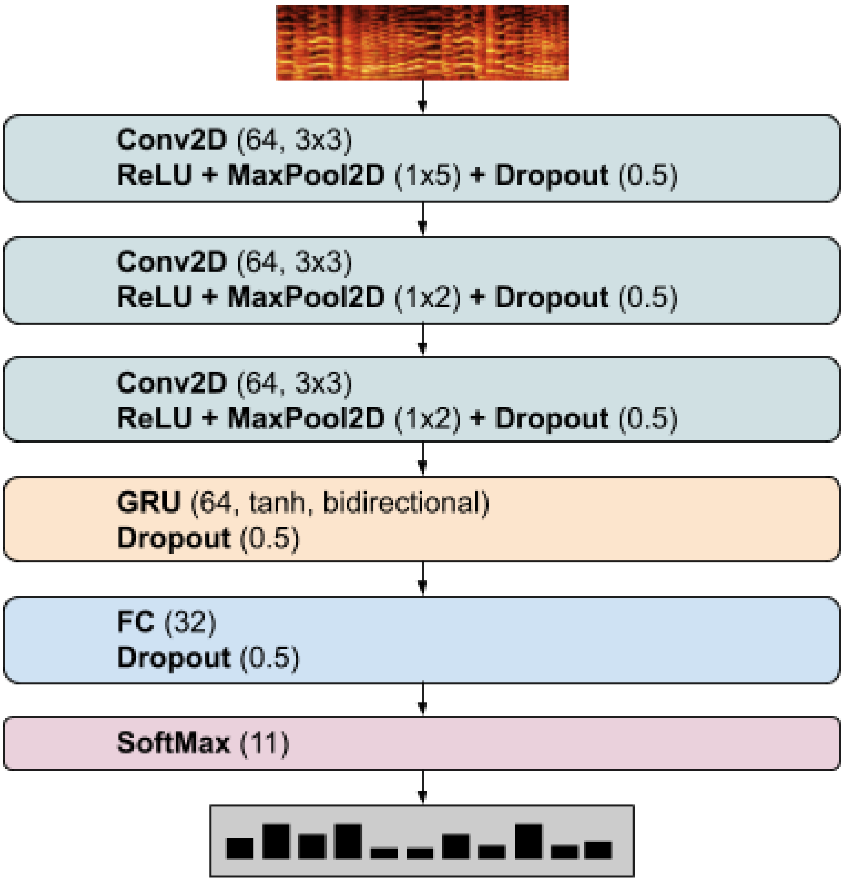


Figure 5.4: Back-end architecture.

line results are shown in Table 5.2.

5.3.2 Parametric front-end

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

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

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

5.3.3 Deep learning classification back-end

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

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

Sound events in the dataset last from $\sim$ 0.2 to 3.3 s. First, clips shorter than 2s are replicated to meet this length. Then, we compute TF patches of log-mel spectrograms of $T = 50$ frames (1 s) and $F = 64$ bands. The values come from the exploration of $T \in \{25, 50, 75, 100\}$ and $F \in \{40, 64, 96, 128\}$. $T = 50$ is the top performing value, roughly coinciding with the median event duration. In turn, more than 64 bands provide inconsistent improvements, at the cost of increasing the number of network weights.

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

For all experiments, the batch size was 100 and Adam optimizer was used [Kingma and Ba,] with initial learning rate of 0.001, halved each time the validation accuracy plateaus for 5 epochs. Earlystopping was adopted with a patience of 15 epochs, monitoring validation accuracy.

Prediction for every event was obtained by computing predictions at the patch level, and aggregating them with the geometric mean to produce a clip-level prediction.

Finally, we apply *mixup* [Zhang et al., 2017] as data augmentation technique. Mixup consists in creating virtual training examples through linear interpolations in the feature space, assuming that they correspond to linear interpolations in the label space. Essentially, virtual TF patches are created on the fly as convex combinations of the input training patches, with a hyper-parameter α controlling the interpolation strength. Mixup has been proven successful for sound event classification, even in adverse conditions of corrupted labels [Fonseca et al., 2019a]. It seems appropriate for this task since the front-end outcome can present leakage due to overlapping sources, effectively mixing two sources while only one training label is available, which can be understood as a form of label noise [Fonseca et al., 2019b]. Experiments revealed that mixup with $\alpha = 0.1$ boosted testing accuracy in $\sim 1.5\%$.

5.4 Results and Discussion

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

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

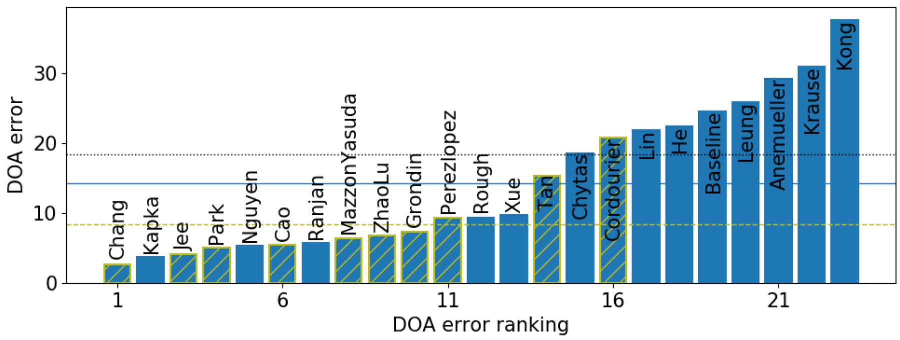
Method	*ER*	*F*	*DOA*	*FR*	***SELD***
Baseline	0.34	79.9%	28.5°	85.4%	0.2113
Proposed	0.32	79.7%	9.1°	76.4%	**0.2026**
Ideal front-end	0.08	93.2%	$\sim 0°$	$\sim 100\%$	0.0379
Baseline	0.28	85.4%	24.6°	85.7%	0.1764
Proposed	0.29	82.1%	9.3°	75.8%	**0.1907**

our method performs slightly worse than the baseline in evaluation mode, while marginally outperforming it in development mode.

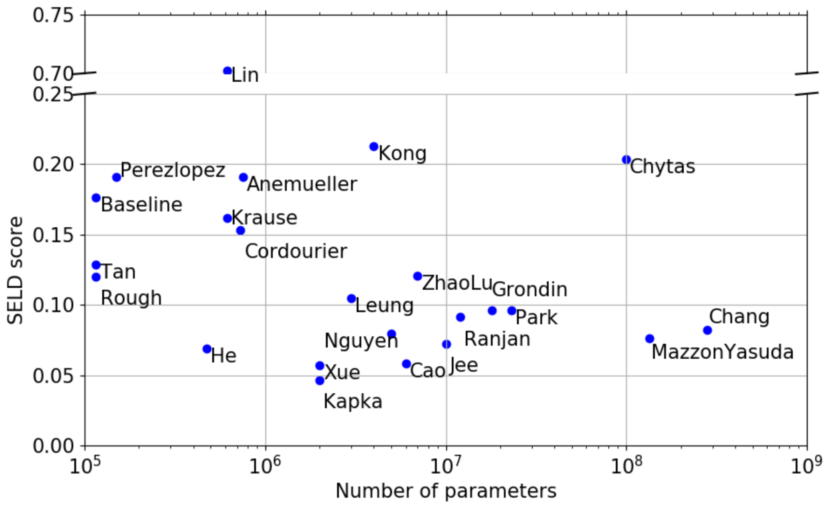
The most relevant observation is the great improvement in *DOA* error. Results suggest that using spatial audio parametric analysis as a preprocessing step can help to substantially improve localization. Figure 5.5a provides further evidence for this argument: Challenge methods using some kind of parametric preprocessing (*GCC-PHAT* with the microphone dataset, and *Intensity Vector-Based* in ambisonics) obtained in average better *DOA* error results.

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

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



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



(b) *SELD score* versus complexity.

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

on the back-end performance. Yet, the proposed system reaches similar performance to the baseline system in terms of SED metrics.

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

5.5 Conclusion

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

Chapter 6

DATA GENERATION AND STORAGE

Explain about mono files plus ambisonics IRs.

6.1 Recorded IRs

Impulse Responses (IRs) measurements constitute a compact way of representing the acoustic propertiesof a linear time-invariant system. When such measurements are performed in a specific room or enclosure, the so-called Room Impulse Responses (RIRs) are able to capture the intrinsic reverberation and acoustic characteristics of the enclosure, for which several methods have been developed these past years [Stan et al., 2002]. Furthermore, it is possible to account for different emitter/receiver positions in the measurement, usually performing the measurement with a microphone array. In that case, this kind of measurements is referred to as Directional Room Impulse Responses (DRIRs) [Embrechts et al., 2005]. DRIRs have a wide range of applications: auralization [Embrechts et al., 2005], room acoustics analysis [Embrechts, 2015,Clapp et al., 2011] and modelling [Romblom, 2017], spatial audio synthesis [Coleman et al., 2017], source separation and dereverberation [Baqué et al., 2016], acoustic heritage preservation [Gerzon, 1975b, Murphy, 2005], etc.

6.1.1 Ambisonics Recording

It is also possible to capture Ambisonics audio scenes by using specific recording devices. Spherical microphone arrays (also known as *Ambisonics microphones*) are a type of microphone arrays in which the capsules are located around a spherical surface, presenting rotational symmetry. Such geometric arrangement allows the recorded signals to be transformed to the Ambisonics domain, by means of projection into the Spherical Harmonics basis, and further equalization with radial filters [Bertet et al., 2006].

This process is known in the audio production domain as *A-B Conversion*. All major spherical microphone array manufacturers provide tools for achieving this transformation [Soundfield, 2018, Sennheiser, 2018, Zylia, 2018, VVAudio.com, 2018, Acoustics, 2018].

6.1.2 Ambisonics DRIRs

The intrinsic spatial capabilities of Ambisonics microphones might be applied to DRIR recording, as originally proposed by Gerzon in the context of acoustical heritage preservation [Gerzon, 1975b]. In recent years, several datasets of Ambisonics DRIRs have been publicly released, such as the OpenAIR database [Murphy and Shelley, 2010] or the set of measurements performed in the scope of the S3A project [Coleman et al., 2018, ope, 2018].

6.1.3 The SOFA Conventions

In general, the Ambisonics DRIR databases show a common approach for describing the measurements: given a specific room, usually represented as a folder, IR data consist of several multichannel audio files, with one audio channel per spherical harmonic, and one file per emitter/receiver combination. Furthermore, it is also usual to provide a *metadata* file, describing the different emitter and receiver positions, and eventually some information about the measurement setup, methodology, etc. Such files might be formatted as plain text or delimiter-separated value files.

Despite the common approach, it can be 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 situation hinders data manipulation and exchange, and forces users to write ad-hoc parsers and algorithms for each specific database.

The *Spatially Oriented Format for Acoustics* (SOFA) is a file format designed for a consistent, standardized storage and manipulation of IR data [Majdak and Noisternig, 2015]. The need for such standard arose from dealing with different databases of *Head-Related Transfer Functions* (HRTFs), in a similar manner as the one mentioned for Ambisonics DRIRs.

There are several SOFA conventions, each one addressing a particular type of IR measurement. In the case of Ambisonics DRIRs, given the data representation in existing databases, one could outline the following specificities:

- Presence of Ambisonics-related information (Ambisonics order, channel ordering and normalization)
- Audio stored in the Ambisonics (spherical harmonics) domain
- Data structure support for different combinations of source and receiver positions

However, none of the existing SOFA conventions meets those requisites. The potentially first candidate by name, *SingleRoomDRIR*, is limited to one source position per file. On the other hand, *MultiSpeakerBRIR* allows for multiple sound sources, but restricts the number of receivers (microphone capsules) to two, as expected in a binaural recording. *GeneralFIRE* is intended for "data which are too general to store in more specific conventions" [SOFA, 2018a].

6.1.4 Convention Proposal

Considerations

The SOFA specification defines some criteria that must be fulfilled in order to propose a new convention [SOFA, 2018b]. 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.

As we already mentioned, existing databases of Ambisonics DRIRs can be found in OpenAIR and S3A databases (criterium 1). Criterium 2 has been discussed in the previous paragraph. All available Ambisonics DRIR datasets are accompanied by explanations, pictures, diagrams and related information (criterium 3).

Specifications

We therefore propose a new SOFA convention, *AmbisonicsDRIR*, designed for DRIRs measured with spherical microphone arrays, and presented in the Ambisonics domain. In other words, we propose to store the instantaneous B_{mn} values for a given Ambisonics order L, provided that S is a unit impulse (Eq. **??**).

The *Listener* - as defined by the SOFA specification - is embodied by the spherical microphone array. The different *Receivers* are the different Ambisonics Components B_{mn}, so that their number is fixed provided L (more precisely, by following the relationship $R = (L+1)^2$), and their positions are not applicable. Furthermore, there might be many different *Sources*, all of them omnidirectional and consisting of one only *Emitter*. In that sense, *M* represents the number of different *Source* positions.

The proposed convention is based on *GeneralFIRE*, with the following modifications:

- Mandatory field *GLOBAL:AmbisonicsOrder* (type *double*, dimension *I*, default *1*). Indicates the order of the Spherical Harmonic expansion.

- Mandatory field *DATA:ChannelOrdering* (type *attribute*, default *acn*). Describes the ordering of the different Ambisonics Channels. Must be one of: *acn* or *fuma*.

- Mandatory field *DATAChannelNormalization* (type *attribute*, default *sn3d*). Describes the Ambisonics normalization convention used in the data. Must be one of: *sn3d*, *n3d*, *fuma* or *maxn*.

- Fields *ReceiverPosition*, *ReceiverPosition:Type* and *ReceiverPosition:Units* are not needed, since it is assumed that *Receivers* represent the Ambisonics channels. Furthermore, the values of *R* (number of receivers) are defined by $R = (L+1)^2$. Accordingly, the values of *GLOBAL:AmbisonicsOrder* and *R* are valid only if they follow the given equation.

- Field *DATA:Delay* is not mandatory.

- Relevant information about the microphone model and brand, Ambisonics encoding methodology, software used, etc, is recommended to be added into the field *GLOBAL:Comment*.

6.1.5 Results

As a preliminary result, we have extended the current SOFA C++ [Pérez-López, 2018b] and Matlab [Muynke, 2018] APIs to be fully compatible with the AmbisonicsDRIR convention. pysofaconventions

Furthermore, as a use-case, a selection of existing Ambisonics DRIRs have been transcoded to the proposed convention: *Emmanuel Main Church* from S3A database [Coleman et al., 2018], and *Heslington Church* and *York Guildhall Council Chamber* from the OpenAIRlib [ope, 2018]. Figure 6.1 shows a schematic diagram of the different *Emitter/Receiver* positions at the *York Guildhall Council Chamber* recordings. The data, as well as the tools used to perform the conversion, are available online under an open source license [Pérez-López, 2018a].

Finally, we must remark that the *AmbisonicsDRIR* SOFA convention proposal is currently under discussion, and the described specifications are subject to change with upcoming versions.

6.1.6 Summary

This document addresses the lack of compatibility among different databases of Ambisonics DRIRs. The present proposal consists in defining a new SOFA convention, specifically designed for Ambisonics DRIRs. That way we contribute to improve the ease of data manipulation and database interoperability. Software implementations and tools for automatic conversion are provided.

6.2 Simulated IRs

explain different methods and libraries. tell about masp.

6.3 High-level scene description

6.3.1 Motivation

introduce the topic

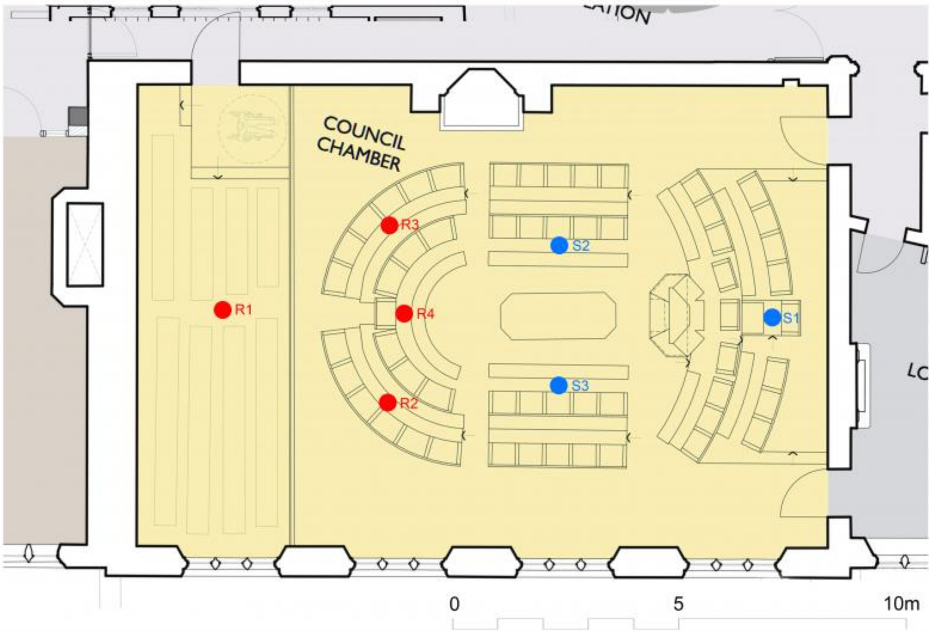


Figure 6.1: Source and Listener position diagram for the "York Guildhall Council Chamber" Ambisonics DRIR (original diagram attribution to [of York Council, 2018]

ARRANGE this introduction in a meaningful way
The work by Moore [Moore et al., 2015] was the first to combine Intensity Vector statistics with DPD-test preprocessing. A similar approach is found on the *Single Source Zones (SSZs)* algorithm, first presented by Pavlidi [Pavlidi et al., 2015]. Other recent proposals, as for example the work by He [He and Chen, 2017], consider the local DOA variance as an estimator of reliability.

Source Localization based on IV statistics has been also successfully used as a preprocessing step for Source Separation tasks. The proposal of Günel [Gunel et al., 2018], which uses beamforming over the DOA estimation, was a pioneer work on this scope. Other similar approaches, which rely on DOA-based Time-Frequency (TF) masking, can be found on the works by Shujau [Shujau et al., 2011], Riaz [Riaz, 2015] or Chen [Chen et al., 2015].

6.3.2 Evaluation data

Audio data used for evaluation in the works presented in Section **??** can be classified in three main categories:

- *IR simulations:* Impulse Responses simulated by numerical methods, using software tools such as *SMIR Generator* [Jarrett et al., 2012]. IR simulation provides the most flexible approach for evaluation, since it allows custom design and parametric analysis (algorithm evaluation as a function of reverberation time, source(s) distance(s), etc). However, the resulting audio data realism is restricted by the simulation algorithm, which is usually limited to empty *shoebox* rooms.

- *IR recordings:* Impulse Responses (IRs) of specific rooms recorded with an Ambisonics microphone, as first presented by Gerzon in the context of acoustic heritage preservation [Gerzon, 1975b]. IR recordings do not provide the flexibility of IR simulations, but allow realistic captures of the room's acoustic properties, while maintaining the potential of using different source contents and positions. OpenAIRlib [Murphy and Shelley, 2010] is the reference public repository for IRs, even including a dedicated Ambisonics IR section.

- *Audio recordings:* a sound scene recorded in a specific room, using an Ambisonics microphone and real sound sources. This approach represents the most accurate option for algorithm evaluation in a real scenario. However, its cost and lack of flexibility makes this approach only attractive for last stages of algorithm evaluation.

In the case of IR simulations and recordings, the Ambisonics IRs are convolved with anechoic monophonic recordings to produce the actual evaluation data. The anechoic recordings are usually taken from standard audio processing databases, such as TIMIT [Garofolo et al., 1993].

Table 6.1 summarizes the evaluation data types used in each work review in Section **??**, along with the audio content type and origin database (if any).

Although algorithm evaluation in real scenarios provides the most realistic approach to the problem, this methodology has not been widely adopted due to its cost and lack of flexibility. As an example, none of the works using real recordings in Table 6.1 performs the evaluation with more than one recording in each case. In contrast, when using IR-based scenes, the number of audios evaluated are usually one or two magnitude orders greater.

However, it is important to notice the lack of public availability of evaluation 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 Sound Localization, and original sound sources for Sound Separation). Only in the case of simulated IRs it is possible to partially replicate the experimental setup, since some of the parameters used in the simulation software are

Article	**Evaluation Data Type**	**Audio Content Type**	**Database**
Thiergart [Thiergart and Schultz-Amling, 2009]	Audio recording	Speech	-
Tervo [Tervo, 2009]	Audio recording	Noise, music	-
Jarret [Jarrett et al., 2010]	IR simulation, Audio recording	Noise	-
Nadiri [Nadiri and Rafaely, 2014]	IR simulation, Audio recording	Speech	TIMIT
Moore [Moore et al., 2015]	IR simulation	Speech	APLAWD
Pavlidi [Pavlidi et al., 2015]	IR simulation	Noise, speech	-
He [He and Chen, 2017]	IR simulation, Audio recording	Speech	TIMIT
Gunel [Gunel et al., 2018]	IR recording	Speech, music	Music for Archimedes
Shujau [Shujau et al., 2011]	Audio recording	Speech	TIMIT
Riaz [Riaz, 2015]	IR recording	Speech, music	Music for Archimedes
Chen [Chen et al., 2015]	IR simulation, IR recording	Speech	TIMIT

Table 6.1: Summary of audio data used across Ambisonics-based Source Localization (above) and Source Separation (below) algorithm proposals. update table? fix references

usually provided. Furthermore, the process of creation of the custom datasets (selecting an anechoic audio dataset, convolving with custom IRs, or performing real recordings) seems to be performed *ad-hoc* on each article.

Taking into account the flexibility offered by IR-based scenes, it would be desirable to have a tool for automatic generation of reverberant Ambisonics scenes (and their associated groundtruth) for analysis purposes. Such tool would help the scientific community in several ways: reducing the amount of time dedicated to build custom datasets, reusing publicly available resources and recordings, and enhancing experiment reproducibility by making easier the exchange of datasets. Furthermore, the capacity of producing a big number of diverse audio scenes for analysis will help the algorithm design and early testing stages,and specially the training stage of machine learning-based algorithms.

6.3.3 AMBISCAPER

this has been downgraded to subsection. maybe place it at the introduction or something to find a more meaningful explanation

AmbiScaper [Perez-Lopez, 2018] is a tool designed to provide a flexible way of creating complex Ambisonics sound scenes and their associated groundtruth, to be used in the context of Source Localization and Source Separation algorithms. AmbiScaper offers a high level control of the sound scene parameters, and provides an easy way of creating large datasets with custom characteristics. AmbiScaper is based on Scaper, a framework designed to generate ground truth information to train Sound Event Detection models [Salamon et al., 2017].

6.3.4 Sound scene description

One of the main features of AmbiScaper is that all parameters of the sound scene can be specified in a non-deterministic way.

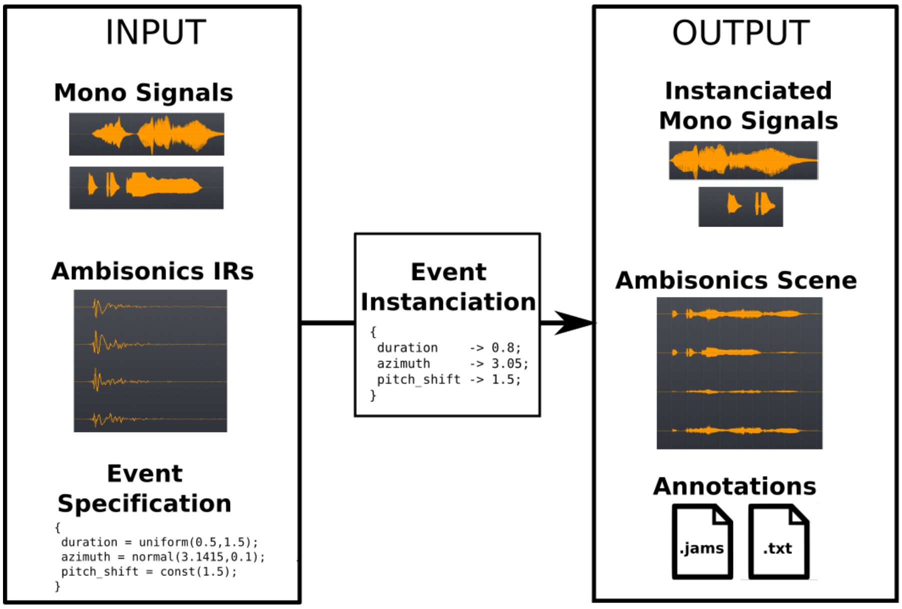


Figure 6.2: AmbiScaper architecture.

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 Instanciation*, 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.

6.3.5 Architecture

In order to generate a sound scene, AmbiScaper requires three different inputs: the original *mono signals*, which will be the basis for the scene’s audio content, an optional *Ambisonics IR*, and the *event specification*.

The process of dataset creation starts with the *event instanciation*, as described in Section 6.3.4. Once all values are sampled, three different types of output are generated: the *Ambisonics scene*, the *instanciated mono signals* (the original mono signals after data augmentation and duration changes), and the *annotations*, in the form of a JAMS file [Humphrey et al., 2014], containing all information about the scene specification and the instanciated values.

AmbiScaper's architecture is depicted in Figure **??**.

6.3.6 Reverberation

When no reverberation is specified, AmbiScaper can generate anechoic sound scenes, which can be useful for baseline performance evaluation. In this case, there is no upper limit on the Ambisonics order of the rendered scene. Furthermore, the anoechic case allows for the specification of source *spread* or apparent size through Ambisonics order downgrade [Carpentier, 2017a].

AmbiScaper supports the usage of recorded Ambisonics IRs, although it is currently limited to IRs from the S3A database [Coleman et al., 2018]. The development of a standardized file format for Ambisonics IRs, which is being discussed at the moment of writing, will provide the flexibility to work with arbitrary Ambisonics IRs.

Lastly, AmbiScaper features the possibility of using simulated Ambisonics IRs, through a wrapper to *SMIR Generator*. In this case, the reverberation model specifications might be defined as well in statistical terms, and the generated IRs are stored for evaluation purposes. A working copy of Matlab is required to run this option.

When a reverberant sound scene is created, the specific Ambisonics IRs used on the scene are also provided as an output. Other research problems such as dereverberation or room reflection modelling might therefore benefit from such data.

6.3.7 AmbiScaper and experiment reproducibility

As mentioned in Section 6.3.2, there is a generalized lack of publicly available datasets for Source Localization and Source Separation in the Ambisonics domain. Even when using general purpose audio/speech datasets (such as TIMIT), the actual reverberant Ambisonics evaluation data is usually not available. In that sense, the compatibility of AmbiScaper with public Ambisonics IR databases is a key aspect for reproducibility, since it allows for the reutilization of acoustical measurements in a systematic way.

Furthermore, the output of the AmbiScaper dataset generation process is not limited to the actual dataset. In fact, as explained in Section 6.3.5, the resulting annotation file does not only contain the *instanciation* (the actual values of each parameter in the sound scene), but also the *specification* (the statistical distributions from which the instanciated values are sampled). In the scope of experiment reproducibility, the exchange of *specification files* instead of actual audio files greatly reduces the storage capacitiy and bandwith required to transfer big databases.

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

6.3.8 Sample Dataset

As an example of the potential capabilities of AmbiScaper, we have created and published a dataset for the evaluation of source localization algorithms [1].

The dataset contains 300 first order Ambisonics sound scenes, each one containing from one to three static sound sources (be they simultaneous or not), with different gains and SNRs, and placed

[1]https://zenodo.org/record/1186907 link

at random positions around the sphere. Each sound scene has a duration between 1 and 2 seconds.

The sources are randomly chosen from a subset of the Anechoic OpenAirlib database, which mostly contains recordings from baroque musical instruments. The AudioBooth IR database from S3A has been used for all scenes. It features SoundField recordings from an acoustically treated room with a geodesic dome structure, to which 17 speakers have been attached.

6.3.9 CONCLUSIONS AND FUTURE WORK

AmbiScaper is a tool designed for easy dataset creation and exchange, in the context of reverberant Ambisonics Source Localization and Source Separation. It responds to the lack of public datasets for algorithm design, evaluation and reproducibility. An example dataset generated with AmbiScaper, and its analysis with state-of-the-art Source Localization algorithms are also provided. Further studies in Ambisonics IR position interpolation would allow creating non-static audio scenes, which might be an interesting feature. Another feature in progress is the support for a standardized Ambisonics IR format [Perez-Lopez and De Muynke, 2018], which will ease data reusability and provide the potential to cover a big variety of acoustic scenarios.

Chapter 7

CONCLUSIONS

7.1 Summary of Contributions

- Academic Contributions

  1. Analysis of spherical isotropic noise fields with an A-Format tetrahedral microphone ref
     - **Parameter estimation**: Contribution to the characterization of coherence with tetrahedral microphones (the most common spherical arrangement)
  2. Autoregressive B-Format Late IR Estimation ref
     - **Parameter estimation**: Novel technique for RT60 estimation from autoregressive models (subproduct of dereverberation)
     - **Signal enhancement**: Novel methodology to re-reverberate sound scenes (include new elements in the scene using the reverb of the recorded scene)
  3. A hybrid parametric-deep learning approach for sound event localization and detection ref
     - **Scene Description**: ovel State-of-the-Art methodology for Sound Event Localization and Detection

- Software Contributions

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

7.2 List of related publications

Bibliography

[cod,] https://github.com/andresperezlopez/DCASE2019_task3.

[ope, 2018] (Accessed on Marcg 8th, 2018). The open acoustic impulse response library www.openairlib.net.

[Acoustics, 2018] Acoustics, M. (Accessed on March 8th, 2018). Eigenstudio https://mhacoustics.com/download.

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

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

[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.

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

[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. *120th AES Convention*, pages 1–24.

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

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

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

[Carpentier, 2017b] Carpentier, T. (2017b). Normalization schemes in ambisonic: does it matter? In *Audio Engineering Society Convention 142*. Audio Engineering Society.

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

[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., 2018] Coleman, P., Remaggi, L., and Jackson, P. J. B. (2015 (accessed July 23th, 2018)). *S3A room impulse responses [Dataset]*. University of Surrey. http://epubs.surrey.ac.uk/808465/.

[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.

[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.

[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.

[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.

[Embrechts et al., 2005] Embrechts, J.-J., Lesoinne, S., and Werner, N. (2005). Computation of directional impulse responses in rooms for bettter auralization. In *Audio Engineering Society Convention 118*. Audio Engineering Society.

[Epain and Jin, 2016] Epain, N. and Jin, C. T. (2016). Spherical Harmonic Signal Covariance and Sound Field Diffuseness. *arXiv:1607.00211 [cs]*. arXiv: 1607.00211.

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

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

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

[Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). *TIMIT acoustic-phonetic continous speech corpus LDC93S1*. Philadelphia: Linguistic Data Consortium.

[Gerzon, 1975a] Gerzon, M. (1975a). The design of precisely coincident microphone arrays.

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

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

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

[Gunel et al., 2018] Gunel, B., Hachabiboglu, H., and Kondoz, A. M. (2018). Acoustic source separation of convolutive mixtures based on intensity vector statistics. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):748–756.

[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. page 4.

[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.

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

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

[Howard et al., 2017] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and

Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*.

[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.

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

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

[Jarrett et al., 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.

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

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

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

[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.

[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.

[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, 2005] Murphy, D. T. (2005). Multi-channel impulse response measurement, analysis and rendering in archaeological acoustics. In *Audio Engineering Society Convention 119*. Audio Engineering Society.

[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.

[Muynke, 2018] Muynke, J. D. (Accessed on February 13th, 2018). Github, sofa matlab/octave api https://github.com/jdemuynke/API_MO.

[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.

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

[of York Council, 2018] of York Council, C. (Accessed on February 13th, 2018). Guildhall drawings https://www.york.gov.uk/downloads/file/11337/guildhall_drawings.

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

[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.

[Perez-Lopez, 2018] Perez-Lopez, A. (2018 (accessed July 23th, 2018)). *AmbiScaper*. https://andresperezlopez.github.io/ambiscaper/.

[Perez-Lopez and De Muynke, 2018] Perez-Lopez, A. and De Muynke, J. (2018). Ambisonics directional room impulse response as a new convention of the spatially oriented format for acoustics. In *144th Audio Engineering Society Convention*. AES.

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

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

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

[Pulkki, 2006a] Pulkki, V. (2006a). Directional audio coding in spatial sound reproduction and stereo upmixing.

[Pulkki, 2006b] Pulkki, V. (2006b). 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 et al., 2018] Pulkki, V., Delikaris-Manias, S., and Politis, A. (2018). *Parametric time-frequency domain spatial audio*. Wiley Online Library.

[Pérez-López, 2018b] Pérez-López, A. (Accessed on February 13th, 2018b). Github, sofa cpp api `https://github.com/andresperezlopez/API_Cpp`.

[Pérez-López, 2018a] Pérez-López, A. (Accessed on March 12th, 2018a). Github, ambisonicsdrirexamples. `https://github.com/andresperezlopez/AmbisonicsDRIRexamples`.

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

[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.

[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.

[Schorkhuber and Holdrich, 2017] Schorkhuber, C. and Holdrich, R. (2017). Ambisonic Microphone Encoding with Covariance Constraint. page 6.

[Sennheiser, 2018] Sennheiser (Accessed on March 8th, 2018). Ambeo plugins `https://en-us.sennheiser.com/ambeo-blueprints-downloads`.

[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, 2018a] SOFA (Accessed on February 13th, 2018a). Generalfire `https://www.sofaconventions.org/mediawiki/index.php/GeneralFIRE`.

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

[Soundfield, 2018] Soundfield (Accessed on March 8th, 2018). Soundfield product listing `http://www.soundfield.com/products`.

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

[Stan et al., 2002] Stan, G., Embrechts, J., and Archambeau, D. (2002). Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262.

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

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

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

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

[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 and Schultz-Amling, 2009] Thiergart, O. and Schultz-Amling, R. (2009). Localization of sound sources in reverberant environments based on directional audio coding parameters. In *127th Audio Engineering Convention*. AES.

[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. Audio Engineering Society.

[VVAudio.com, 2018] VVAudio.com (Accessed on March 8th, 2018). Vvtetravst `https://www.vvaudio.com/products/VVTetraVST`.

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

[Zylia, 2018] Zylia (Accessed on March 8th, 2018). Zylia ambisonics converter `http://www.zylia.co/zylia-ambisonics-converter.html`.