# 3D PERCEPTUAL SOUNDFIELD RECONSTRUCTION VIA SOUND FIELD EXTRAPOLATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

EGE ERDEM

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
MULTIMEDIA INFORMATICS

JANUARY 2020

Approval of the thesis:

**3D PERCEPTUAL SOUNDFIELD RECONSTRUCTION VIA SOUND FIELD EXTRAPOLATION**

submitted by **EGE ERDEM** in partial fulfillment of the requirements for the degree of **Master of Science in Multimedia Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, Graduate School of **Informatics** _______________

Assist. Prof. Dr. Elif Sürer
Head of Department, **Multimedia Informatics, METU** _______________

Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu
Supervisor, **Multimedia Informatics, METU** _______________

**Examining Committee Members:**

Prof. Dr. Alptekin Temizel
Multimedia Informatics, METU _______________

Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu
Multimedia Informatics, METU _______________

Assoc. Prof. Dr. Banu Günel Kılıç
Information Systems, METU _______________

Assist. Prof. Dr. Zühre Sü Gül
Department of Architecture, Bilkent University _______________

Assist. Prof. Dr. Elif Sürer
Multimedia Informatics, METU _______________

**Date: 08.01.2020**

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Surname: Ege Erdem

Signature :

# ABSTRACT

## 3D PERCEPTUAL SOUNDFIELD RECONSTRUCTION VIA SOUND FIELD EXTRAPOLATION

Erdem, Ege
M.S., Department of Multimedia Informatics
Supervisor: Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu

January 2020, 51 pages

Perceptual sound field reconstruction (PSR) is a spatial audio recording and reproduction method based on the application of stereophonic panning laws in microphone array design. PSR allows rendering a perceptually veridical and stable auditory perspective in the horizontal plane of the listener, and involves recording using near-coincident microphone arrays. This thesis extends the two dimensional PSR concept to three dimensions and allows reconstructing an arbitrary sound field based on measurements with a rigid spherical microphone array. This work offers a method for emulating near coincident microphone recordings by using rigid spherical microphone arrays via sound field extrapolation carried out in the spherical harmonic domain.

An active-intensity-based analysis of the rendered sound field shows that the proposed approach can render direction of monochromatic plane waves accurately even with a straightforward extension of PSR directivity patterns designed for the 2D case. For the real recordings, listening tests are conducted using multi-channel audio recordings of the reconstructed sound field and compared with higher order Ambisonics recordings.

Keywords: sound field reconstruction, sound field extrapolation, spherical harmonics, spherical microphone arrays

# ÖZ

## SES ALANI DIŞDEĞERLENDİRMESİ İLE 3 BOYUTLU ALGISAL SES ALANI OLUŞTURMA

Erdem, Ege
Yüksek Lisans, Çokluortam Bilişimi Anabilim Dalı Bölümü
Tez Yöneticisi: Doç. Dr. Hüseyin Hacıhabiboğlu

Ocak 2020 , 51 sayfa

Algısal ses alanı oluşturma, iki kanallı kaydırma kurallarına dayanan uzamsal bir ses alanı kaydetme ve yeniden üretme yöntemidir. Algısal ses alanı oluşturma, yakın mikrofon dizileri kullanılarak yatay düzlemde dinleyiciye algısal olarak gerçeğe uygun ve kararlı bir dinleme deneyimi sunar. Bu tez, iki boyut için geçerli olan bu yöntemi 3. boyut için de gerçekleştirebilir hale getirmekte; kapalı mikrofon dizisi ölçümleri kullanarak herhangi bir ses alanını yeniden oluşturmayı mümkün kılmaktadır. Bu bağlamda, kapalı mikrofon dizisi ölçümlerini kullanarak küresel harmonik alanda yapılan ses alanı dış değerlendirmesi ile açık mikrofon dizisinden elde edilecek ses kayıtları tahmin edilmiş, diğer bir deyişle sanal olarak elde edilmiştir.

İki farklı yükseklikte bulunan beşer adet hoparlör kullanılarak oluşturulan ses alanının aktif yeğinlik analizi yapılarak; tek frekanslı düzlemsel dalgaların yönlerinin, 2 boyutlu PSR için tasarlanmış olan yönsellik örüntüsünün doğrudan kullanılmasına rağmen yüksek doğrulukla oluşturulabildiği sayısal olarak gösterilmiştir. Gerçek ses kayıtları ile oluşturulan ses alanı ise çeşitli dinleyici tesleri yapılarak ve yüksek dereceli Ambisonics kayıtları ile karşılaştırılmıştır.

Anahtar Kelimeler: ses alanı oluşturma, ses alanı dışdeğerlendirmesi, küresel harmonikler, küresel mikrofon dizinleri

This thesis is dedicated to the people who care about me.

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Hüseyin Hacıhabiboğlu, who is the sole reason for my decision about doing masters and being a research assistant, the ideas that were entirely far to me.

I owe a very important debt to Enzo De Sena and Zoran Cvetković for their guidance and informative inputs to our ICASSP 19' paper [1], which is my first academical contribution to this world, that would not have been possible.

Special thanks to Arne Leijon, the author of PairedCompCalc, giving incredibly fast, helpful, and spot-on responses about the usage of their Python package just when I needed most, for the first pilot experiment. The detailed and kind technical support of Daniel Rudrich for the IEM Plugin and Ambisonics decoding is truly appreciated.

Another special thanks also to my old friend Çağrı, who worked very hard to load my preliminary experiment on his online server, helped me collecting the experimental data very smoothly, and found a critical bug for the real experimental setup.

Many thanks to Mehmet Çalışkan and Zühre Sü Gül, the founders of my former company, and all of the members of Mezzo Stüdyo; for providing comfortable, flexible, and very high-quality job life.

I would also thank all of my friends who are part of our routine event called "synergy nights", for creating a productive working environment altogether.

# TABLE OF CONTENTS

CHAPTERS

# LIST OF TABLES

TABLES

# LIST OF FIGURES

FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 2D | 2 Dimensional |
| 3D | 3 Dimensional |
| PSR | Perceptual Soundfield Reconstruction |
| RSMA | Rigid Spherical Microphone Array |
| SHD | Spherical Harmonic Decomposition |
| SH | Spherical Harmonic |
| ISTFT | Inverse Short-time Fourier Transform |
| STFT | Short-time Fourier Transform |
| ITCD | Inter-Channel Time Difference |
| ILCD | Inter-Channel Level Difference |
| DMA | Differential Microphone Array |
| TF | Time-Frequency |

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation and Problem Definition

Recreating the auditory experience of an acoustic performance has been of interest for broadcasters, artists and academics for over a century. Blumlein was the first to attempt rendering of sound sources accurately in space using a pair of figure-eight microphones. Since then, many solutions have been proposed [2], exploring different microphone configurations, directivity patterns, and channel-mixing strategies. Apart from approaches based on Ambisonics recording, these methods are predominantly heuristic, and thus fundamentally reliant on the skills and taste of sound engineers.

Ambisonics and WFS are the notable examples of the methods that aim to reconstruct an accurate physical approximation of a sound field [3] however they have high equipment load requirements especially in terms of loudspeaker numbers. Various other methods of lower complexity, that are perceptually motivated multichannel audio reproduction systems that rely on psychoacoustics and human auditory perception have also been proposed such as vector-based amplitude panning (VBAP) [4], spatial impulse response rendering (SIRR) [5, 6], directional audio coding (DirAC) [7], spatial decomposition method (SDM) [8], and perceptual soundfield reconstruction PSR [9, 10].

The latter (PSR) uses a set of microphones of specifically designed directivity patterns, each connected to a loudspeaker in a corresponding direction without additional mixing [10]. It was shown that PSR performs on a par with VBAP [4] and second-order Ambisonics in the centre of the sweet-spot, but has a more graceful performance degradation away from the sweet-spot. More specifically, PSR provides better locatedness of phantom sources than techniques based on intensity alone. Using results of a computational model [11], the increased locatedness was attributed to the higher naturalness of the presented binaural cues [12].

The current formulation of PSR has two main limitations:

- It is confined to the horizontal plane,
- Each channel requires a dedicated (2nd or higher order) microphone.

The latter becomes particularly problematic for extending PSR to the full 3D case, as the number of required microphones increases substantially. The motivation of this thesis is to overcome these limitations and to investigate extending PSR to three dimensions using a single, spherical harmonics based coincident microphone array such as the Eigenmike.

## 1.2 Proposed Methods and Models

A novel reproduction strategy is proposed. In order to obtain emulated virtual near coincident microphone recordings, that is the "sound field extrapolation" results, pressure and intensity of the sound field around the rigid spherical microphone is calculated at the positions of the corresponding PSR microphones. This method enabled us to use original microphone directivity patterns in PSR [10] and reconstruct the sound field directly.

## 1.3 Contributions and Novelties

The main contributions of this thesis are:

- A novel method of using sound field extrapolation to reconstruct any arbitrary field is proposed.
- Near coincident microphone array recordings are emulated virtually, with using only a single rigid spherical microphone.

Following conference publication has been written and presented with the work reported in this thesis:

- E. Erdem, E. De Sena, H. Hacıhabiboğlu and Z. Cvetković, "Perceptual Soundfield Reconstruction in Three Dimensions via Sound Field Extrapolation," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 8023-8027.

## 1.4 The Outline of the Thesis

Firstly, a mathematical and an acoustical background is presented in 2.1 and 2.2 respectively, aiming to give all the necessary information to be able to understand the proposed work. Section 2.3 introduces rigid spherical microphone arrays and gives an insight to understand the pressure field around the microphone. Chapter 2 is concluded with some of the spatial audio reproduction methods. Chapter 3 explains PSR in detail, the prior work of this thesis which is the "perceptual" part of the proposed audio reproduction method.

The core of this thesis, sound field extrapolation is described in Chapter 4 and Chapter 5 for two different cases: for plane waves and for real recordings, respectively. Plane wave reproduction is numerically analyzed in Section 4.3.1 whereas real recordings are evaluated by performing a binaural listening experiment in Section 6.6. Finally, Chapter 7 concludes the thesis also discussing future work directions.

# CHAPTER 2

# BACKGROUND

In this chapter, technical background is given. Definitions of spherical harmonics, spherical harmonics decomposition and special spherical functions are presented. Then plane-wave composition of a sound field and rigid microphone arrays are introduced. Finally the most common state of the art methods are summarized.

## 2.1 Mathematical Background

Due to the fact that this work is based on spherical microphone and loudspeaker arrays, it is necessary to discuss the representation of sound fields using a series of spherical harmonics. That is to say, spherical coordinate system is used throughout this thesis. In addition, spherical harmonics and special spherical Bessel and Hankel functions are central components to proposed method. Section 2.1 aims to give required mathematical background to be able to understand the methods in the subsequent sections. After that, Section 2.2 establishes a connection between mathematical background and physical acoustics.

### 2.1.1 Spherical Coordinate System

Consider a point defined in Cartesian coordinates as $\mathbf{x} = (x, y, z)$, this point is transformed to spherical coordinates using equations in 2.1. Spherical coordinate of the point is defined as $\mathbf{r} = (r, \theta, \phi)$ where radial distance, inclination and azimuth angles are defined as $r$, $\phi$, $\theta$ respectively The relation of spherical coordinates with Cartesian coordinates is visualized in Fig. 2.1.

$$\begin{aligned} x &= r \sin\theta \cos\phi \\ y &= r \sin\theta \sin\phi \\ z &= r \cos\theta \end{aligned} \tag{2.1}$$

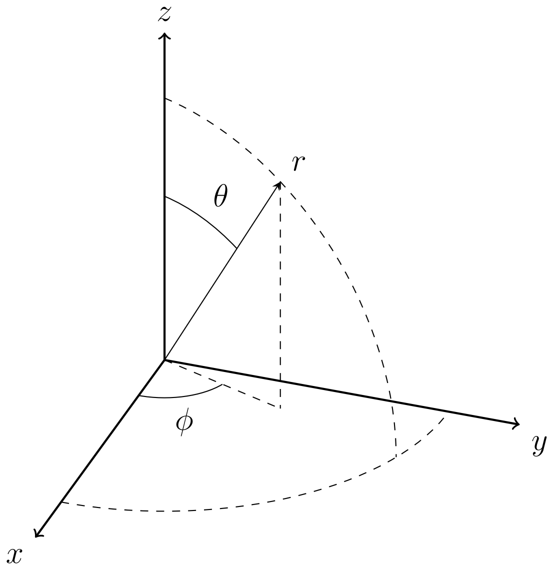


**Figure 2.1:** Spherical coordinate system variables ($r$, $\phi$, $\theta$) correspondents in Cartesian coordinate system.

### 2.1.2 Spherical Bessel and Hankel Functions

Solution of the acoustic wave equation in spherical coordinates results in special functions of spherical Bessel and Hankel functions. In other words, solutions to the wave equation can be represented as a linear combination of spherical Bessel and Hankel functions. These special functions are crucial for obtaining pressure field due the sound waves and will be frequently encountered in the following chapters.

Spherical Hankel functions of the first kind, $h_n(x)$, and the second kind, ${h_n}^{(2)}(x)$ can be written as a combination of spherical Bessel function of the first kind, $j_n(x)$ , and of the second kind, $y_n(x)$, as follows :

$$h_n(x) = j_n(x) + iy_n(x) \tag{2.2}$$

$${h_n}^{(2)}(x) = j_n(x) - iy_n(x) \tag{2.3}$$

Following relations can also be inferred from Eq. 2.2 and Eq. 2.3:

$$j_n(x) = \text{Re}\{h_n(x)\} \tag{2.4}$$

$$y_n(x) = \text{Im}\{h_n(x)\} \tag{2.5}$$

In order to keep the exposition in this thesis simple, the individual definitions of spherical Bessel and Hankel functions are not given since the properties of the function is

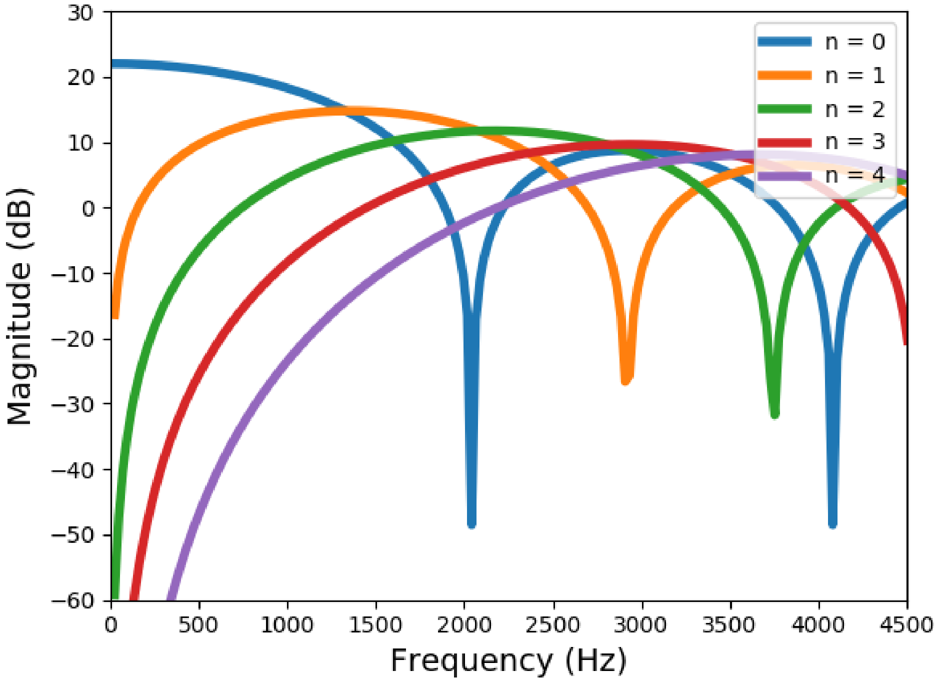


**Figure 2.2:** The magnitude of $4\pi j_n(kr)$ for $n = 0, ..., 8$ as a function of frequency, with $r = 8.4$ cm and $k = 2\pi f/c$.

the important part rather than the definition itself. That being said, first 4 orders of the spherical Bessel function is depicted in Fig. 2.2 giving the magnitude and frequency relation. Note that frequency is proportional to the wave number $k = 2\pi f/c$ and forming the x axis of the figure.

### 2.1.3 Spherical Harmonics

The functions on the unit sphere mentioned in Section 2.1 are defined as a weighted sum of basis functions, also analogous to Fourier basis functions applied a the sphere. These basis functions are the spherical harmonics, a set of special functions of the form [13]:

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta)e^{im\phi} \tag{2.6}$$

where $P_n^m(\cdot)$ are the associated Legendre polynomials of order $n \in \mathbb{Z}^+$ and degree $-n \le m \le n$, $\theta \in [0, \pi]$ is the inclination angle with respect to the $+z$ axis, $\phi \in [0, 2\pi)$ is the azimuth angle defined from the $+x$ axis.

Fig. 2.3 depicts real parts of the first five orders of spherical harmonic function where Mollwiede projection for visual representation.

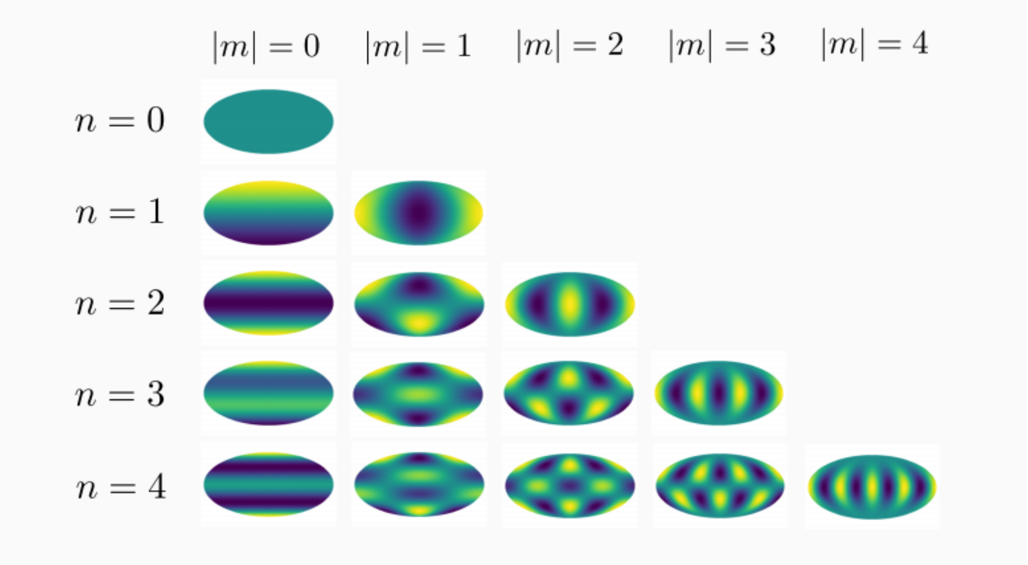


**Figure 2.3:** Real parts of the first five orders of the spherical harmonic function, $\mathrm{Re}[Y_n^m(\theta, \phi)]$.

## 2.2 Acoustical Background

### 2.2.1 Spherical Harmonic Decomposition

Spherical harmonics of order $n \in \mathbb{N}$ and degree $m \in \mathbb{Z}$ are defined in 2.6. Series of spherical harmonics can be used to approximate wide range of functions defined on the sphere by the following relation [14]:

$$p(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm} Y_n^m(\theta, \phi) \tag{2.7}$$

where $p(\theta, \phi)$ is the function which is being projected to spherical harmonic basis and $p_{nm}$ is the spherical harmonic decomposition coefficients. In other words, $p_{nm}$ can be considered as the weights that compose the spherical Fourier Transform of the function $p(\theta, \phi)$. Any function on the sphere can be represented once the SHD coefficients $p_{nm}$ are obtained with the following surface integral :

$$p_{nm} = \int_0^{2\pi} \int_0^{\pi} p(\theta, \phi) Y_n^m(\theta, \phi)^* \sin\theta d\theta d\phi \tag{2.8}$$

If the function $p(\theta, \phi)$ represents a pressure distribution due to a sound field, then:

$$p(\theta, \phi) = p(k, r, \theta, \phi) \tag{2.9}$$

$$p_{nm} = p_{nm}(k, r) \tag{2.10}$$

where $p(k, r, \theta, \phi)$ is the sound pressure at a frequency $\omega = kc$ and $c$ is the speed of sound on a spherical surface of radius $r$.

Since pressure distribution on a spherical surface deduced from microphone recordings, one can only have finite number of microphone recordings in real life scenarios. As a result, double integral in 2.8 becomes a summation that is introduced in the following sections.

### 2.2.2 Plane-wave Composition of a Sound Field

#### 2.2.2.1 A Single Plane Wave

An arbitrary sound field can be represented as a linear combination of an infinitely many plane waves. Starting from exponential representation of a single plane wave (Eq. 2.11), plane wave composition of a sound field in spherical harmonics domain can be derived.

Consider a unit-amplitude, single-frequency plane wave, arriving from direction $(\theta_k, \phi_k)$ with a wave vector $-\mathbf{k} = (k, \theta_k, \phi_k)$. Sound pressure at $r = (r, \theta, \phi)$ due to this plane wave is $e^{i\mathbf{k}\cdot\mathbf{r}}$ in exponential from and can be written as a summation of spherical harmonics and spherical Bessel functions [13]:

$$p(k, r, \theta, \phi) = e^{i\mathbf{k}\cdot\mathbf{r}} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n j_n(kr)[Y_n^m(\theta_k, \phi_k)]^* Y_n^m(\theta, \phi) \tag{2.11}$$

In practice, representation of plane waves with an infinite summation is overcome by approximating it as a finite summation, *i.e.* Eq. 2.11 becomes:

$$p(k, r, \theta, \phi) = e^{i\mathbf{k}\cdot\mathbf{r}} \approx \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n j_n(kr)[Y_n^m(\theta_k, \phi_k)]^* Y_n^m(\theta, \phi) \tag{2.12}$$

As the spherical Hankel functions diverge at the origin , spherical Bessel functions are used to represent a plane-wave sound field.

Equation 2.11 provides the sound pressure of a single plane wave. Now, if the sound pressure is evaluated at the surface of a sphere of radius $r$, function $p(k, r, \theta, \phi)$ can be represented with SHD coefficients $p_{nm}(k, r)$ defined in 2.2.1, satisfying :

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi) \tag{2.13}$$

Comparing Eq. 2.7 and 2.11, spherical harmonic decomposition coefficients $p_{nm}(k, r)$ for a sound field composed of a single plane wave can be written as:

$$p_{nm}(k, r) = 4\pi i^n j_n(kr)[Y_n^m(\theta_k, \phi_k)]^* \tag{2.14}$$

#### 2.2.2.2 Multiple Plane Waves

If the sound field composed of multiple plane waves, the sound pressure can be written using 2.11, similar to Eq. 2.8 with and amplitude density denoted by $a(k, \theta_k, \phi_k)$:

$$\begin{aligned} p(k, r, \theta, \phi) &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n j_n(kr) Y_n^m(\theta, \phi) \\ &\times \int_0^{2\pi} \int_0^{\pi} a(k, \theta_k, \phi_k)[Y_n^m(\theta_k, \phi_k)]^* \sin\theta_k d\theta_k d\phi_k \qquad (2.15) \\ &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi i^n a_{nm}(k) j_n(kr) Y_n^m(\theta, \phi) \qquad (2.16) \end{aligned}$$

where $a_{nm}(k)$ is the spherical harmonic decomposition or spherical Fourier Transform of $a(k, \theta_k, \phi_k)$. Comparing Eq. 2.11 with 2.16, the following relation for single, unit amplitude plane wave can be derived:

$$a_{nm}(k) = [Y_n^m(\theta_k, \phi_k)]^* \qquad (2.17)$$

All the introduced equations considered, spherical harmonic decomposition coefficients for a combination of plane waves evaluated at the surface of a sphere of radius $r$ can be written as,

$$p_{nm}(kr) = 4\pi i^n j_n(kr) a_{nm}(k) \qquad (2.18)$$

whereas finite summation approximation of Eq. 2.16 is,

$$\begin{aligned} p(k, r, \theta, \phi) &= \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n a_{nm}(k) j_n(kr) Y_n^m(\theta, \phi) \qquad (2.19) \\ &= \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi) \end{aligned}$$

From these results it is worth noting the direct relation between $p_{nm}$ and $a_{nm}$, which are the measured function and the function generating the sound field respectively. Equation 2.18 relates SHD coefficients of the sound pressure $p_{nm}$ to the SHD coefficients of the plane-wave amplitude density $a_{nm}$, that are both in spherical harmonics domain facilitating the calculation process. With some alterations, these relations are employed to obtain pressure distribution around spherical surface of a rigid sphere in the following sections.

## 2.3 Rigid Spherical Microphone Arrays (RSMAs)

The first step of the proposed work is to record an auditory event with a rigid spherical microphone. To that end, understanding the structure of this type of microphone as well as other similar types of microphones such as open sphere microphone arrays and near-coincident pairs are crucial to apprehend the advantages of the work introduced in this thesis. This section divided into three parts; Section 2.3.1 to be able to visualise different types of microphone arrays, 2.3.3 to understand the relation between microphone numbering and the order of spherical harmonics and 2.3.2 to calculate the sound pressure around the surface of the rigid sphere.

### 2.3.1 Microphone Types

Spherical microphone arrays can be classified into two groups as open and closed (*i.e.* rigid spherical microphone arrays) given in Fig. 2.4. Open arrays includes microphones positioned on the surface of an open sphere whereas closed arrays have microphones positioned on a rigid spherical surface. For the latter; due to the closeness of the structure, microphone itself become a spherical scatterer and this spherical rigid body imposes a boundary condition on its surface of zero radial particle velocity which result in a change in the expression of pressure and SHD coefficients for a plane wave given in Section 2.2.

The microphone array that was used in this thesis was Eigenmike32®, a microphone array with 32 electret microphones embedded onto the surface of a 4.2cm radius rigid spherical baffle (see Fig.2.4b), more specifically, located at the faces of a truncated icosahedron and at the center of each face [15]. Early prototype of Eigenmike32® was introduced in 2003 [16] by mh acoustics.

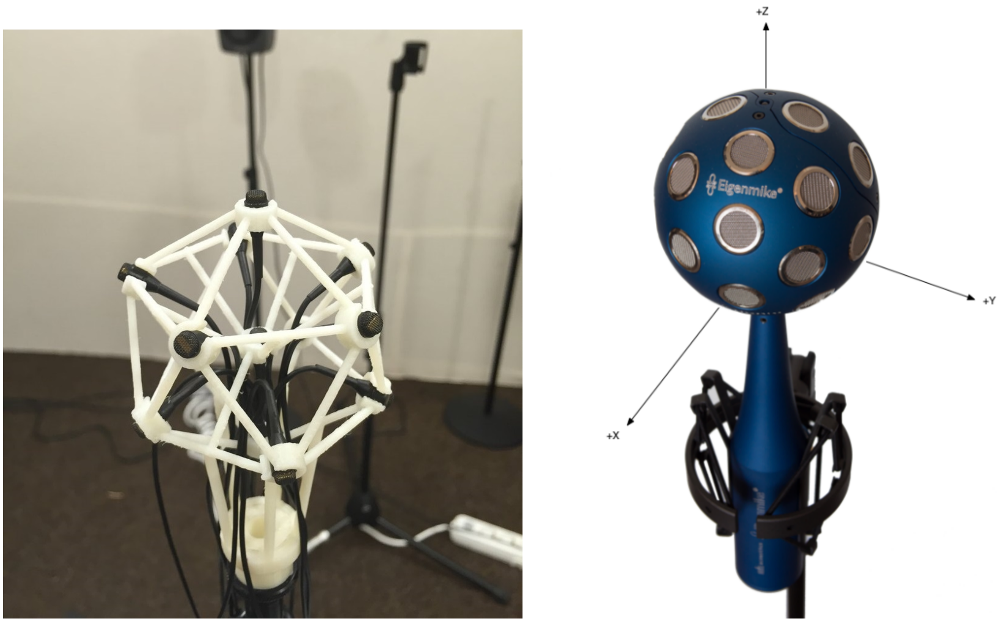
The open microphone array in Fig.2.4a consists microphones positioned over a spherical surface but this time with an open sphere instead of a rigid baffle, implemented by METU Spatial Audio Research Group (SPARG).

Finally, examples of near-coincident microphone configurations are given in 2.5. A near-coincident pair is a two microphone positioned nearly coincident but spaced few centimeters apart, angled symmetrically on either side of the centre. It is a stereo miking technique often preferred for their stereo image and especially used by audio recording engineers and musicians.

One of the promising outcomes of this thesis is about emulating near-coincident microphone arrays with only using a rigid spherical microphone array, which will be further explained in the subsequent chapters.

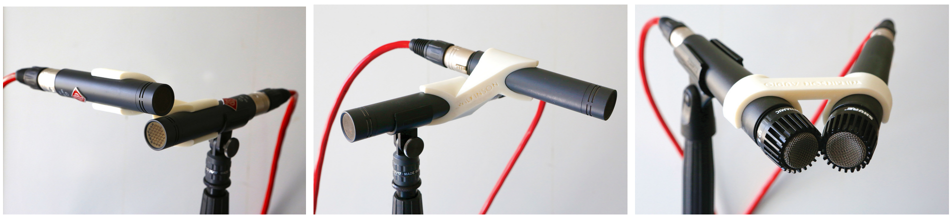
### 2.3.2 Sound Pressure Around a Rigid Sphere

In previous subsections of Section 2.2, sound pressure on a spherical surface due to plane waves are approximated with SHD coefficients. In this section, spherical har-



(a) (b)

**Figure 2.4:** Open (a) and closed (b) microphone array examples.



(a) (b) (c)

**Figure 2.5:** Near coincident microphone pair examples.

monic decomposition (SHD) coefficients are derived to obtain sound pressure around a rigid spherical surface.

The sound pressure around a rigid sphere has two components: incident field that is the sound field in free field without the rigid sphere, $p_i$, and the sound field that is scattered from the rigid sphere due to the incident field, $p_s$. The overall sound field is calculated using these two pressure field components. In addition, rigid sphere imposes a boundary condition due to infinite impedance at the sphere boundary, making sound pressure impossible to generate a radial motion at the boundary, causing a zero radial velocity. Since velocity and pressure are dependent via Euler equation (or conservation of momentum in fluid dynamics), incident and scattered pressure fields can be derived with the boundary conditions mentioned above.

Incident and scattered pressures are substituted to Euler equation, leading to

$$\frac{\partial}{\partial r}[p_i(k,r,\theta,\phi)+p_s(k,r,\theta,\phi)]\bigg|_{r=r_a}=0 \tag{2.20}$$

Now write the incident and scattered sound pressure as a spherical harmonic series as[13]

$$p_i(k,r,\theta,\phi)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}a_{nm}(k)4\pi i^n j_n(kr)Y_n^m(\theta,\phi) \tag{2.21}$$

$$p_s(k,r,\theta,\phi)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}c_{nm}(k){h_n}^{(2)}(kr)Y_n^m(\theta,\phi) \tag{2.22}$$

Rewriting 2.20 with 2.21 and 2.21 yields

$$c_{nm}(k)=-a_{nm}(k)4\pi i^n\frac{j_n^{'}(kr_s)}{h_n^{(2)'}(kr_s)}. \tag{2.23}$$

By summing up the incident and the scattered pressures as $p=p_i+p_s$, total pressure around a rigid sphere is obtained as,

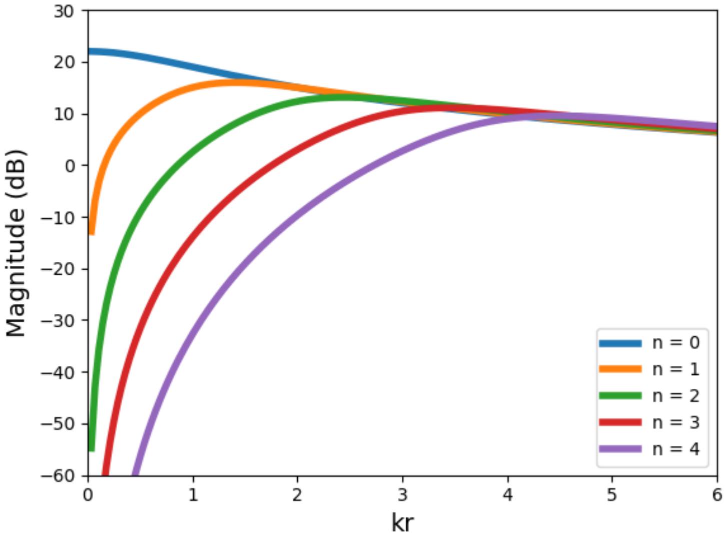
$$p(k,r,\theta,\phi)=\sum_{n=0}^{\infty}\sum_{m=-n}^{n}4\pi i^n a_{nm}(k)\left[j_n(kr)-\frac{j_n^{'}(kr_s)}{h_n^{(2)'}(kr_s)}h_n^{(2)}(kr)\right]Y_n^m(\theta,\phi) \tag{2.24}$$

The structure of the pressure function is similar to Eq. 2.19, except the spherical Bessel term $j_n(kr)$ becomes a new term denoted as $b_n(kr)$

$$b_n(kr)=4\pi i^n\left[j_n(kr)-\frac{j_n^{'}(kr_s)}{h_n^{(2)'}(kr_s)}h_n^{(2)}(kr)\right]. \tag{2.25}$$

where $j_n(\cdot)$ and $h_n^{(2)}(\cdot)$ are the spherical Bessel function of the first kind and spherical Hankel function of the second kind, respectively. The derivatives of these functions with respect to their arguments are given as $j_n^{'}(\cdot)$ and $h_n^{(2)'}(\cdot)$, respectively.

The behavior of the magnitude of $b_n(kr)$ with respected to $kr$ is presented in Fig. 2.6 In Fig. 2.7, the same Bessel function is drawn this time with respect to order $n$. Note that the order $n$ is also the argument of the Bessel function together with $k$ and $r$. An important result inferred from Fig.2.6 is that the magnitude of $b_n(kr)$ decreasing at the higher orders of $n$. Importance of this behaviour is explained in the following section.



(a)

**Figure 2.6:** (a)$|b_n kr|$ for a rigid sphere with $r = r_a$, for $n = 0, ..., 5$ and (b) $|b_n kr|$ with $r = r_a = 8.4cm$ and $kr = 8$.

Finally, pressure around a rigid spherical microphone array is obtained by substituting $b_n(kr)$ to 2.24 as

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k) b_n(kr) Y_n^m(\theta, \phi) \tag{2.26}$$

Fundamental spherical harmonics approximation of any sound pressure function given in Section 2.2 Eq.2.7 is rewritten here for convenience:

$$p(k, r, \theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}(k, r) Y_n^m(\theta, \phi) \tag{2.27}$$

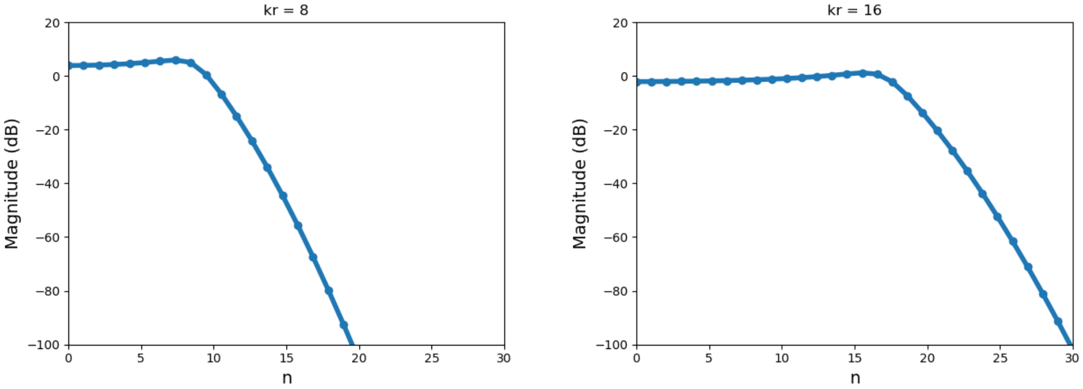
Comparing 2.26 with 2.27, following relation can be derived

$$p_{nm}(k, r) = a_{nm}(k) b_{nm}(kr) \tag{2.28}$$

$$a_{nm}(k) = \frac{p_{nm}(k, r)}{b_{nm}(kr)} \tag{2.29}$$

Note the similarity between 2.18 and 2.29, with $4\pi i^n j_n(kr)$ replaced by $b_{nm}(kr)$.

In summary, SHD coefficients obtained from RSMA recordings include the scattering effect of the rigid sphere which depends both on frequency and on the radius of the



(a)

(b)

**Figure 2.7:** (a) $|b_n(kr)|$ for a rigid sphere with $r = r_a$, for $n = 0, ..., 5$ and (b) $|b_n(kr)|$ with $r = r_a = 8.4cm$ and $kr = 8$.

sphere. This effect is eliminated with $b_n(kr)$ term, also providing some advantages in terms of spatial aliasing and reconstruction accuracy that is explained in the following section.

### 2.3.3 Spatial Aliasing

Pressure field over a spherical surface due to single plane wave and multiple plane waves is given in Section 2.2. Then, pressure field around a rigid spherical surface is derived in Section 2.3.2. Resulting pressure function approximations include SHD coefficients terms, $p_{nm}(k, r)$, that should be obtained via RSMA recordings. As a result, the sampling of sound pressure functions in space requires microphone measurements. The positions and the number of microphones used in RSMA determine the sampling points. The design of a rigid microphone array introduces a trade-off between hardware complexity and the system accuracy. The total number of microphone on a spherical surface determines the accuracy of the reconstructed sound pressure function.

In order to achieve perfect representation of a sound field from the finite number of microphone samples, sampling theorems require the sound pressure functions to be order-limited. This means the function should be represented by a finite number of spherical harmonics, *i.e.* the basis functions [13].

Note that Eq.2.29 shows the magnitude of $p_{nm}$ is proportional to the magnitude of $b_n(kr)$. It is therefore expected that $p_{nm}$ for a multiple plane-wave sound field decays as a function of $n$ for $n > kr$, as suggested by Fig. 2.6, and more explicitly, as illustrated in Fig. 2.7.

This is an important result, as it suggests that the sound field represented by the infinite summation in Eq. 2.11 can be represented by a finite summation as in Eq. 2.12

with little error. The spherical harmonics series for a plane-wave sound field can therefore be considered as nearly order limited, so that sampling theories for order-limited functions can be applied [13].

Moving to RSMA recordings, SHD coefficients can be obtained using a finite set of $Q$ pressure signals captured on a rigid sphere of radius $r_a$. This involves the calculation of an approximation to the surface integral using a numerical quadrature:

$$\widetilde{p}_{nm}(k) = \sum_{q=1}^{Q} w_q p(\theta, \phi) \left[Y_n^m(\theta, \phi)\right]^* \tag{2.30}$$

where $w_q$ are the quadrature weights.

Having obtained $\widetilde{p}_{nm}(k)$ including the effect of the rigid sphere, $p_{nm}(k)$ can be calculated eliminating the scattering effect due to rigid sphere as

$$a_{nm}(k) = \widetilde{p}_{nm}(k)/b_n(kr_a) \tag{2.31}$$

With Eq.2.31, SHD coefficients for a sound pressure around a rigid spherical surface due to sound field of multiple plane waves is presented.

If $N$ is the maximum harmonic order that is used during the spherical harmonic decomposition calculations, then SHD coefficients that approximate the pressure function has $(N+1)^2$ spherical harmonics whereas the number of microphones on array, $Q$, should satisfy the following condition:

$$Q \geq (N+1)^2 \tag{2.32}$$

Having 32 microphones on Eigenmik em32, maximum order that can be used is found as $N = 4$.

## 2.4 Spatial Audio Reproduction

Spatial audio reproduction is a term for regenerating an auditory event through loudspeakers or headphones, aiming to reconstruct the sound field indistinguishable from the real sources and from the real sound field that is previously recorded. This experience give the impression of sounds coming from different directions and locations around the listener. There are broad range of approaches with different reproduction strategies and employing various representations for the recorded sound field. To begin, stereophonic reproduction which is an important aspect of spatial audio is introduced. Secondly, some of the spatial audio reproduction methods are explained.

### 2.4.1 Stereophony

Stereophony is the perception of a phantom source between independent loudspeakers. With stereophonic reproduction, illusion of multi-directional audio is created at the listener position even though there are only two loudspeakers. It is the most used reproduction method also for domestic usage. In spatial audio, generally multichannel systems are used to achieve surround sound. In order to realize stereophony with an accurate spatial rendition, various panning methods are developed.

### 2.4.2 Methods

This section explains several state-of-the-art methods for spatial audio reproduction. Since there is no example of audio production with sound field extrapolation, the most widely used approaches are summarized.

### 2.4.3 Perceptually Motivated Approaches

In perceptually motivated approaches such as VBAP, DirAC/SDM/SIRR; the purpose is to reconstruct a field with equal perceptual attributes as the original sound field, instead of aiming to reconstruct the original sound field precisely.

**Vector Based Amplitude Panning (VBAP)** is one of the most widely used, straightforward, and efficient stereophonic amplitude panning technique developed by Pulkki [4]. VBAP sets individual gains for multiple loudspeaker reproduction systems by positioning virtual sound sources using a minimum number of loudspeakers at a time. VBAP does not require a fixed number of loudspeakers; an arbitrary loudspeaker configuration can be used as long as the locations of the loudspeakers are equidistant with respect to the listener. Another common requirement for audio reproduction systems is that the listening room should not be very reverberant, to prevent unwanted reflections that decrease sound source localization ability. Vectors and vector bases are used in VBAP, which leads a simple and computationally efficient technique.

The intensity panning technique introduced in VBAP [4] was a two-dimensional amplitude panning, calculates the individual gain factors for loudspeakers. The amplitude relation of these gains determines the direction of the virtual sound source. If there is an equal-loudness constraint of the virtual source during the movement from one loudspeaker to another, gain factors are normalized, setting a constant sound power value. The amplitude panning method of the VBAP is the same as previously introduced stereophonic law of sines by Blumlein [17], this time introduced in a vector base form together with the tangent panning version of it. One of the differences of VBAP among other amplitude panning methods is that the gain factors may have negative values, implying an anti-phase signal.

**DirAC** is a flexible spatial audio system which mainly based on the direction of arrival (DOA) and diffuseness features of the sound field. These features can be related to spatial hearing of human by the following psychoacoustical cues:

- Interaural time difference (ITD)
- Interaural level difference (ILD)
- Interaural coherence

DirAC makes an energy analysis by assigning a direction and a diffuseness level to each output channel of a filter bank with an equivalent rectangular bandwidth scale.

Similar to DirAC, there are methods considering the room impulse responses such as **Spatial Decomposition Method (SDM)** and **Spatial Impulse Response Rendering (SIRR)**. In these methods, the pressure observed in a point of the recorded space is divided into short overlapping time-frames. Using direction-of-arrival estimation methods, the dominant direction of the recorded sound field is calculated for each time frame. In the reproduction space, the pressure signal in each time-frame is rendered using the closest loudspeaker to the estimated direction or panned between loudspeakers triplets using intensity panning such as vector-based amplitude panning (VBAP) [4].

Even though the physical reproduced sound field is possibly quite different from the recorded one, these methods have been used successfully in rendering various sound scenes like concert acoustics, car cabin acoustics and also used at a broad range of applications such as teleconferencing, mobile phone audio optimisation and VR.

### 2.4.4 Sound Field Synthesis Approaches

Different than VBAP and DirAC/SDM/SIRR, **Ambisonics** and **Wave Field Synthesis (WFS)** are sound field synthesis approaches which aim to reconstruct the original sound field directly.

**Ambisonics** is an elegant approach based on the spherical harmonics decomposition of the sound field [18], introduced by Gerzon [19]. Spherical harmonic decomposition is explained in Section 2.2.1. As sound fields can be regarded as a superposition of plane waves, which can be represented as an infinite series of spherical harmonic functions, Ambisonics aims approximate the sound field by reconstructing it in the center of a loudspeaker setup. The listening area of which the accurate reproduction is obtained is called the sweet spot. Since spherical microphone arrays can be used to record spherical harmonic components only up to a certain order, the area of the sweet spot depends mainly on the order of the approximation, which is directly related to the number of loudspeakers. Among spatial audio reproduction methods, the sweet spot is a concern for multi-channel systems as spatial fidelity decreases when the listener moves toward off-center positions even slightly. The optimal listening area is tried to be extended as much as possible to have a more flexible and comfortable system.

In first-order Ambisonics, directional information is encoded into 4 separate channels. This format is the so-called B-Format, and the corresponding four channels are W, X, Y, Z. These channels carry different directional informations of the sound field where W channel is omnidirectional information, and the others are directional information at x,y, and z axes, respectively. As one might guess, extending Ambisonics to **Higher Order Ambisonics (HOA)** introduces new channels, 5 additional channels for second-order, and 7 channels for third-order. Each introduced channel improves the approximation quality to the original sound field. The commercially-available Soundfield microphone allows recording up to the first order, whereas the Eigenmike32® up to the fourth-order [20] harmonics. Once the sound field has been recorded, it is first encoded using suitable ambisonics channel ordering and normalization format; then, the sound field is reproduced using various decoding strategies. The mode-matching decoding, for instance, aims to reconstruct a certain number of spherical harmonic components in the center of the loudspeaker array. The main issue with HOA is that for large orders, it requires large number ($(N+1)^2$ for $N$-th order) of carefully positioned and calibrated loudspeakers.

In Chapter 6, second-order Ambisonics recordings are performed and compared with the proposed 3D PSR method by introducing binaural listening experiments.

# CHAPTER 3

# PRIOR WORK: PERCEPTUAL SOUNDFIELD RECONSTRUCTION (PSR)

This chapter introduces PSR [10], a prior work presenting a systematic framework for design of multichannel surround sound systems. It proposes a methodology for the design of circular microphone arrays in the same configuration as the corresponding loudspeaker array, as two elements of the array shown in Fig. 3.2. In order to create an accurate rendition of the auditory scene, PSR aims to capture inter-channel time and intensity differences. Rather than designing a system based on empirical observation or hands-on tuning, PSR concentrates on psychoacoustics to understand the underlying physical and perceptual phenomena for human auditory perception. In addition, it presents a powerful directivity pattern design method to be used for higher-order microphones and offers a new recording strategy. Directivity design described in PSR is adapted for this thesis.

PSR is performed with a pentagonal setup consists of 5 loudspeakers and 5 microphones, placed $\pi/5$ radians apart. One of the main reason for using pentagonal array setup is due to the fact that it is optimal for reconstruction of first and second-order circular harmonics [21]. During recording stage, 5 microphones are positioned towards the loudspeaker locations and during reproduction stage 5 loudspeaker is positioned inward, towards the corresponding microphone locations as given in Fig. 3.1. Thus, microphone and loudspeaker numbers and configurations match with each other. An important advantage of practical property of the PSR method is that each loudspeaker plays back the exact signal recorded by corresponding microphone, without any post-processing or any mixing.

The concept of active intensity is used when analyzing the reproduction capability of PSR. Complex intensity is the product of pressure and complex conjugate velocity

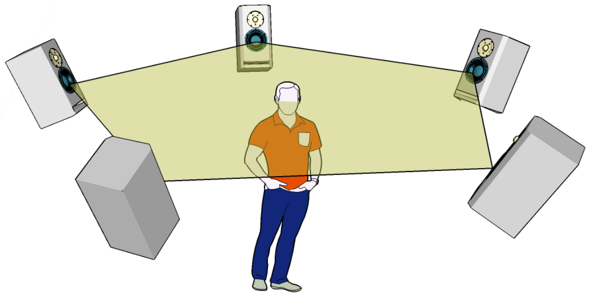
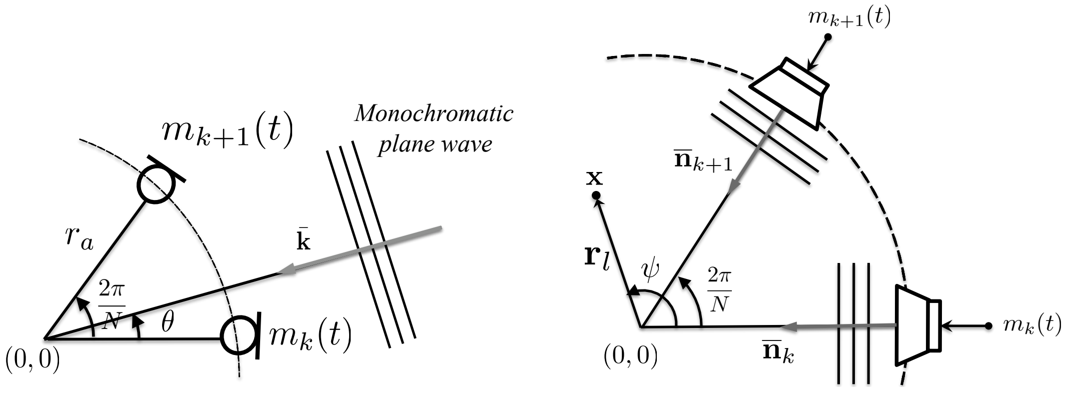


**Figure 3.1:** Reproduction stage of PSR.



(a)

(b)

**Figure 3.2:** Reproduction and recording stages shown for each two of the five microphones and loudspeakers, respectively © 2013, IEEE.

whereas active intensity is the real part of the complex intensity term. Active intensity is co-directional with the sound wave propagation, thus giving the direction of sound at that particular coordinate in the space. Constructing a sound field with a uniform active intensity field is crucial for a pleasing listening experience, as fields with fluctuating intensity has an adverse effect on the size and the stability of the sweet-spot. To achieve this, cross-channel components in the active intensity field representation should be minimized as much as possible. Minimizing the energy of the cross-talk terms is an optimisation problem which reveals that solutions have only two active channel at a time. Thus, plane wave in the direction of $\phi$ is rendered only by the adjacent two loudspeaker channels $(m, m+1)$ satisfying $\phi_m < \phi < \phi_{m+1}$ where $\phi_m = m\frac{2\pi}{5}, m = 0, \ldots, 4$ are the azimuth angles for each loudspeakers.

Directivity patterns can be optimized such that undesirable cross-talks are suppressed by making microphones more selective. Directivity patterns are formulated so that the system operates along the chosen psychoacoustic curves to record and render acoustic sources at locations between loudspeaker pairs in the context of multichannel systems. These curves are time-intensity stereophonic panning curves as established by Franssen [22] and Williams [23] to design directivity patterns, based on the summing localisation effect [24].

Panning curves provide the pairs of time-level differences between loudspeaker pairs that result in the phantom image being perceived in the direction of one loudspeakers or the other. In Figure 3.3, time-intensity curves by Franssen [22] is presented where $L(\tau)$ (left), $R(\tau)$ (right) and $M(\tau)$ (middle) represent the pairs of inter-channel time difference (ICTD) and inter-channel level difference (ICLD) which the auditory event is localized at. For a system with ICTD of $|t_{max}| = 1$ ms, auditory event is localized at left and right loudspeakers $A_L$, $A_R$ respectively. Two example curves to achieve gradual source movement from one loudspeaker to other are showed with a dashed and solid lines connected $A_L$ and $A_R$. ICTD is denoted as $G_R - G_L$ and ICLD as $\tau_L - \tau_R$.

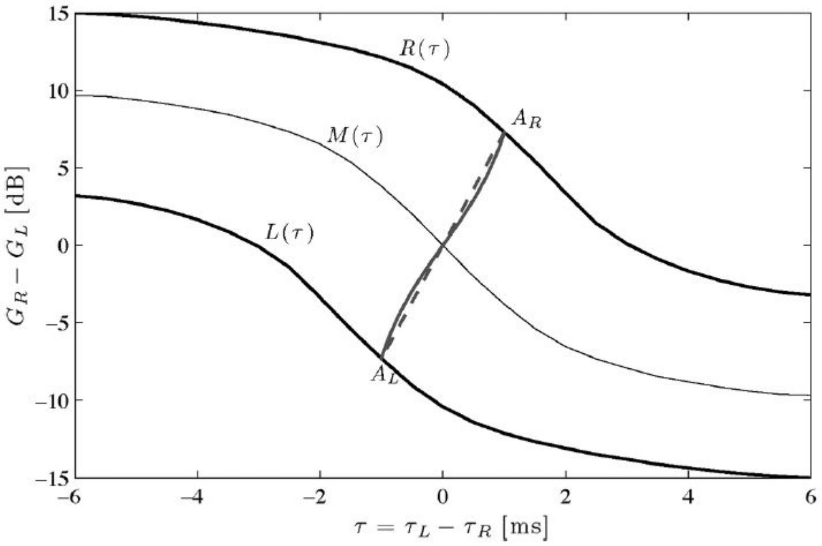


**Figure 3.3:** Time-intensity psychoacoustic curves, adapted from Franssen [22], takenfrom [10] © 2013, IEEE.

To design directivity patterns, first, the maximal time delay between channels due to a source in the direction of one of the loudspeakers is calculated; this delay is specified by the radius of the array and the particular angular placement of channels. Then, the level difference in that direction is set to be the level difference that is needed in combination with the time delay to create a phantom source in that direction. Finally, an equal-power constraint is imposed, giving the following directivity pattern which effectively interpolates time-level difference pairs between the end points:

$$\Gamma_d(\Theta) = \begin{cases} \left[1 + \frac{\sin^2(|\Theta|+\beta)}{\sin^2(|\Theta|-(\phi_0+\beta))}\right]^{-1/2} & \Theta \in [-\phi_0, \phi_0] \\ 0 & \text{elsewhere} \end{cases} \tag{3.1}$$

where $\phi_0$ is the angle between loudspeakers in the horizontal plane and $\beta = \arctan \frac{\eta \sin(\phi_0)}{1-\eta\cos(\phi_0)}$, where $\eta$ is a value calculated from psychoacoustic curves to achieve the desired level differences at loudspeaker directions. The parameters used in the PSR formulation are $\phi_0 = 2\pi/5$ and $\eta = 0.302$. The reader is referred to [10] for details of how these parameters are calculated. Whereas different interpolating functions can be used, the formulation in (3.1) has a form of a generalized tangent panning law, *i.e.* the tangent panning law is its special case for $\beta = 0$, which physically means infinite level differences (expressed in dB) in loudspeaker directions.

Similar to representing functions the on sphere using spherical harmonics as in 2.1.3, microphone directivity patterns can also be approximated with linear combination of spherical harmonic functions as:

$$\Gamma(\theta', \phi') = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm} Y_n^m(\theta', \phi') \tag{3.2}$$

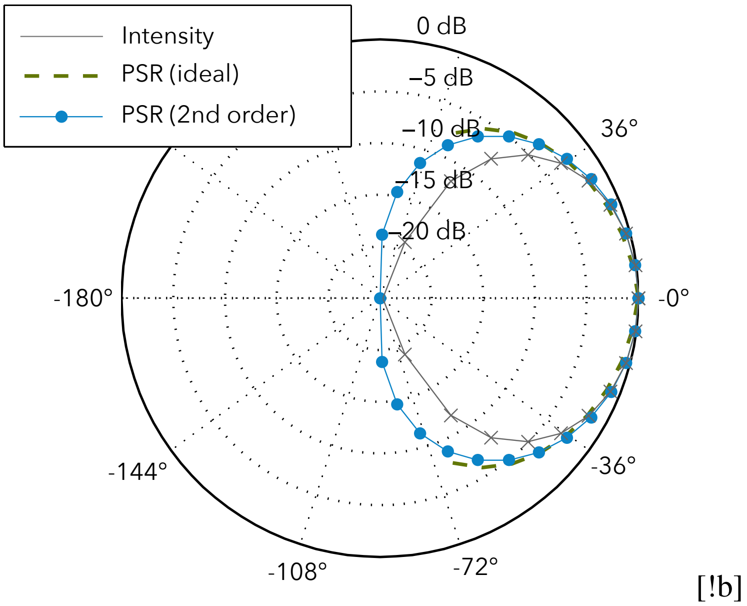


**Figure 3.4:** Original PSR directivity patterns [10] © 2013, IEEE.

where $(\theta', \phi')$ is the local spherical coordinates defined with respect to the microphone axis and the coefficients $\alpha_{nm} \in \mathbb{C}$ have to satisfy $\alpha_{nm} = \pm\alpha^*_{nm}$ to obtain an axisymmetric pattern. In order for $\Gamma(\theta, \phi)$ to be real the coefficients should satisfy $\alpha_{nm} = \alpha^*_{nm}$. The design space can be constrained to use the spherical harmonics of degree $m = 0$, resulting in [25]:

$$\Gamma(\Theta) = \sum_{n=0}^{N} \alpha_n \left(\frac{2n+1}{4\pi}\right) P_n(\cos\Theta) = \sum_{n=0}^{N} \beta_n \cos^n \Theta \tag{3.3}$$

where $\Theta$ is the angle between the acoustic axis of the microphone and the wave of a plane wave. The directivity pattern can be constrained to have unit response in the direction of its acoustic axis by imposing $\sum_{n=0}^{N} \beta_n = 1$. Whereas the method proposed here is not limited to axisymmetric directivity patterns, time-intensity directivity pattern described in Equation 3.1 [10] is employed.

The PSR pattern, $\Gamma_d(\Theta)$, is approximated here as a pattern in the form (3.3) by jointly minimizing the $L^2$-distance in the pick-up region and the $L^2$-norm in the rejection region:

$$\underset{\beta_1,\beta_2,\ldots,\beta_N}{\text{argmin}} \; \lambda \int_0^{\phi_0} |\Gamma(\Theta) - \Gamma_d(\Theta)|^2 + (1-\lambda) \int_{\phi_0+\epsilon}^{\pi} |\Gamma(\Theta)|^2 \, d\Theta \tag{3.4}$$

The resulting pattern for $\lambda = \frac{1}{2}$, $\epsilon = \frac{\pi}{10}$, and $N = 4$ has coefficients $\beta_0 = 0.001$, $\beta_1 = 0.458$, $\beta_2 = 0.536$, $\beta_3 = 0.040$ and $\beta_4 = -0.126$. This pattern (see Fig. 3.4 is used in Chapter 5 to obtain loudspeaker signals.

# CHAPTER 4

# 3D SOUND FIELD EXTRAPOLATION OF MONOCHROMATIC PLANE WAVES

In this chapter, the 3D extension of PSR using sound field extrapolation is proposed. As a proof of concept, this chapter introduces 3D reconstruction of a virtual sound field composed of monochromatic plane waves whereas Chapter 5 realises the same concept with real recordings by RSMAs.

## 4.1 Motivation

The 3D extension of PSR studied here uses a reproduction setup of 10 loudspeakers, arranged in two horizontal layers above and below the ear level, as shown in Fig. 4.1. Two layers rotated by $\pi/5$ with respect to each other, consisting 5 pentagonally placed loudspeakers. The loudspeakers at the top and bottom layers have common inclination angles of $\theta_t \approx 0.352\pi$ and $\theta_b \approx 0.648\pi$, respectively. The azimuth angles for the top and bottom layers are $\phi_t = 2m\pi/5$ and $\phi_b = (2m+1)\pi/5$ for $m = 0, \ldots, 4$.

Similar to 2D PSR recording and reproduction method [10] described in Chapter 3, the aim is to record the sound field with microphones with their acoustic axes pointing towards the corresponding 10 loudspeakers directions and afterwards to reconstruct the recorded field through loudspeakers. 2D PSR strategy requires number of microphones to be equal to number of loudspeakers in the array, positioned on the surface

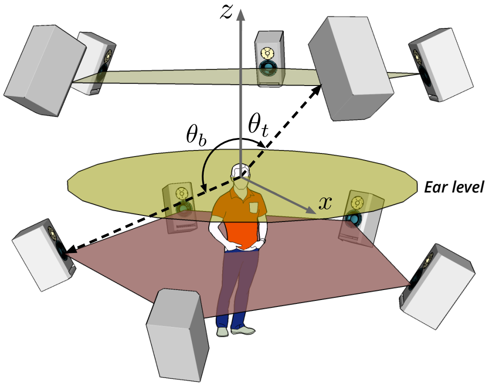


**Figure 4.1:** The proposed reproduction setup with 10 loudspeakers. Taken from [1] © 2019, IEEE

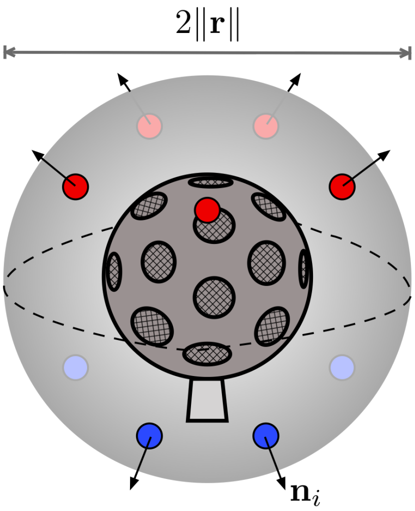


**Figure 4.2:** Virtual microphone positions over $r = 15.5$ cm spherical surface around the Eigenmike, $\mathbf{n}_i$ vectors showing acoustical axes of the microphones.

of an open sphere of radius $15.5$ cm. For 10 loudspeakers, corresponding design of a bespoke 10 channel near-coincident microphone array is challenging. If the design is carried out using differential microphone arrays (DMAs) [26] of order $M$, at least $10M + 1$ microphones would be needed (assuming that the center microphone is shared among all channels), that are equalised for their frequency responses. The positioning of these DMAs would also present practical problems. The same directivity patterns can also be designed via steered beamforming using the SHD coefficients obtained from an RSMA. However, the lack of inter-channel time delays [10] eliminates one of the fundamental premises of the original PSR design. In order to keep the robustness and flexibility offered by the RSMAs whilst capturing interchannel time differences, we propose an approach based on sound field extrapolation using SHD coefficients.

## 4.2 Sound Field Extrapolation

In this section; instead of using an open spherical microphone array or a near coincident microphone array of at least 10 dedicated microphones, emulating these microphones "virtually" via sound field extrapolation is introduced. The aim is to obtain intensity vectors and sound pressure at the locations of specified virtual microphones. The virtual microphones are positioned on a $15.5$ cm radius spherical surface given in Fig. 4.2, based on pyschoacoustical design choices proposed in [9, 27, 28].

Sound field extrapolation is carried out in two steps. First, the pressure is extrapolated at positions of the corresponding PSR microphones. This is done in the spherical harmonic domain in a manner similar to other prior art methods [29]. The direction of the active intensity field is then also extrapolated at the same positions. Second, the directions of the active intensity vectors at locations of PSR microphones are used to weigh the pressure signal according to the PSR directivity patterns.

As explained in Sec. 2.2.2.2, a pressure field composed of plane waves can be approximated using a linear combination of spherical harmonic functions as

$$p(k, \mathbf{r}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n a_{nm} j_n(kr) Y_n^m(\theta, \phi). \tag{4.1}$$

where $\mathbf{r} = (r,\ \theta,\ \phi)$ represents a point in spherical coordinates. Note that $\lim_{N\to\infty} p(k, r, \theta, \phi)$ characterises the sound field exactly at all points while the truncated series in (4.1) will provide a good approximation only up to a finite radius around the origin [13].

Note also that $a_{nm}$ in (4.1) is the same as in the case of plane wave representation of a sound field described in Sec. 2.2.2.2, Eq.4.2, giving

$$a_{nm}(k) = [Y_n^m(\theta_k, \phi_k)]^* \tag{4.2}$$

Particle velocity and pressure are related by conservation of momentum (*i.e.* Euler equation) which can be expressed in time and frequency domains as:

$$-\nabla p(t, \mathbf{r}) = \rho_0 \frac{\partial \mathbf{u}(t, \mathbf{r})}{\partial t} \overset{\mathcal{F}}{\leftrightarrow} -\nabla p(k, \mathbf{r}) = j\rho_0 kc\mathbf{u}(k, \mathbf{r}), \tag{4.3}$$

where $\rho_0$ is the ambient density. Particle velocity can be calculated at an arbitrary point around the origin using this relation as:

$$\mathbf{u}(k, \mathbf{r}) = -\frac{1}{j\rho_0 kc} \left[ \frac{\partial p}{\partial r}\hat{\mathbf{u}}_r + \frac{1}{r}\frac{\partial p}{\partial \theta}\hat{\mathbf{u}}_\theta + \frac{1}{r\sin\theta}\frac{\partial p}{\partial \phi}\hat{\mathbf{u}}_\phi \right] \tag{4.4}$$

where $\hat{\mathbf{u}}_r$, $\hat{\mathbf{u}}_\theta$, and $\hat{\mathbf{u}}_\phi$ are the unit vectors in the radial, inclination and azimuth directions, respectively. The partial derivatives are:

$$\frac{\partial p}{\partial r} = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm} 4\pi i^n k j_n^{'}(kr) Y_n^m(\theta, \phi) \tag{4.5}$$

$$\frac{\partial p}{\partial \theta} = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm} 4\pi i^n j_n(kr) \frac{\partial Y_n^m(\theta, \phi)}{\partial \theta} \tag{4.6}$$

$$\frac{\partial p}{\partial \phi} = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm} 4\pi i^{n+1} m Y_n^m(\theta, \phi) \tag{4.7}$$

where

$$\frac{\partial Y_n^m(\theta, \phi)}{\partial \theta} = \left[ m\cot\theta Y_n^m(\theta, \phi) + \sqrt{(n-m)(n+m+1)} e^{-i\phi} Y_n^{m+1}(\theta, \phi) \right] \tag{4.8}$$

and $Y_n^m = 0$ for $|m| > n$.

Once particle velocity is extrapolated, active intensity, that represents the direction and strength of energy, can be obtained at any point at which the approximation is sufficiently accurate as:

$$\mathbf{I}_{virtual}(k,\mathbf{r}) = \frac{1}{2}\operatorname{Re}\left\{p(k,\mathbf{r})\mathbf{u}^*(k,\mathbf{r})\right\}. \tag{4.9}$$

This will be used to obtain directional responses of emulated microphones located at the desired 10 different positions.

## 4.3 Emulated off-centre microphone recordings for 3D PSR

Once the direction of the sound field is calculated via extrapolation, it is possible to obtain virtual microphone recordings at points around the sphere, and thus emulate near-coincident recording setups. The time-intensity microphone directivity pattern described in Chapter 3 is used in the form of linear combination of cosine terms as

$$\begin{aligned}\Gamma(\Theta) &= \sum_{n=0}^{N}\beta_n\cos^n\Theta \\ &= 0.001 + 0.458\cos(\Theta) + 0.536\cos^2(\Theta) + 0.040\cos^3(\Theta) - 0.126\cos^4(\Theta)\end{aligned} \tag{4.10}$$

The local pressure, $p(k,\mathbf{r})$ and the active intensity vector, $\mathbf{I}_a(k,\mathbf{r})$ can be calculated at any point $\mathbf{r}$ within a region where the sound field extrapolation is accurate. The directional response of the emulated microphone depends on $\Theta_r$, the angle between the local intensity vector and the acoustic axis of the microphone, demonstrated visually in Fig. 4.3 as

$$\Theta_{\mathbf{r}} = \arccos\frac{\langle\mathbf{r},\mathbf{I}_a(k,\mathbf{r})\rangle}{|\mathbf{I}_a(k,\mathbf{r})||\mathbf{r}|} \tag{4.11}$$

Then, the emulated microphone signal is obtained as

$$p_{\text{rec}}(k,\mathbf{r}) = \Gamma(\Theta_{\mathbf{r}})p(k,\mathbf{r}). \tag{4.12}$$

3D PSR involves the calculation of emulated microphone signals at a distance of $15.5$ cm with the directivity pattern given in (4.10). The acoustic axes of the emulated microphones are radially outwards. It is assumed that listener's head would be positioned at the centre of the loudspeaker rig.

### 4.3.1 Evaluation of 3D PSR of Plane Waves

The proposed method 3D PSR is evaluated in terms of their directional reproduction accuracy for monochromatic plane wave fields. Sound field due to such a wave is

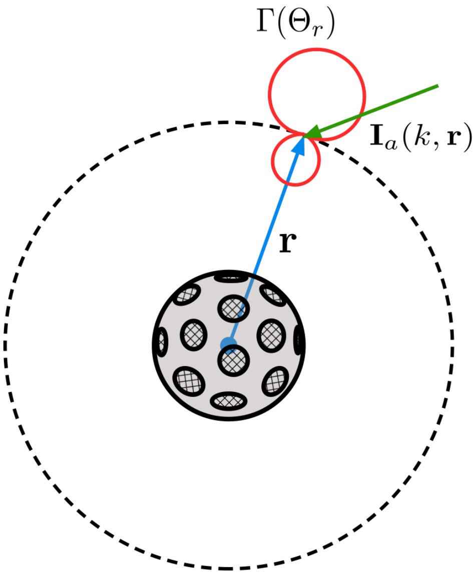


**Figure 4.3:** Example vectors of local intensity (green) and acoustical microphone axis (blue).

homogenous and the active intensity is aligned with the direction of propagation of the wave. Since the direction of active intensity vectors represent the direction of the sound field, numerical simulations are carried out showing active intensity vector field together with a contour map of directional error.

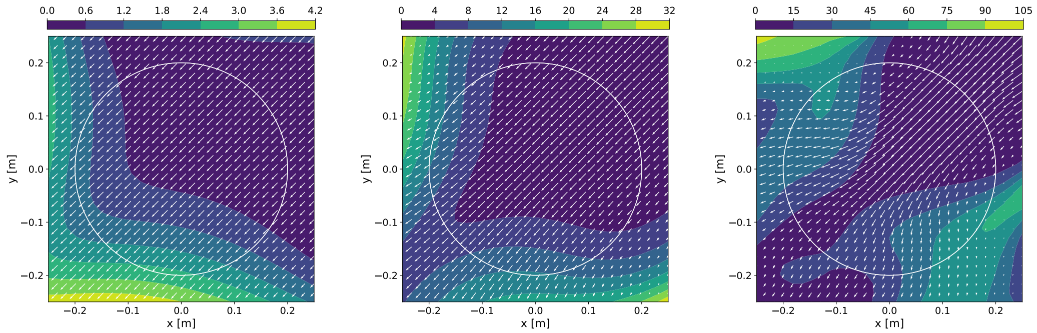
Absolute angular error used in the discussion below is defined as:

$$\epsilon(\mathbf{r}) = \arccos \langle \mathbf{n}_{PSR}(\mathbf{r}), \mathbf{n}_{pw}(\mathbf{r}) \rangle \tag{4.13}$$

where $\mathbf{n}_{PSR}$ is the unit vector in the direction of the reproduced field and $\mathbf{n}_{pw}$ is the unit vector in the direction of the plane wave.

We evaluated the directional accuracy of reproduced sound fields due to monochromatic plane waves at frequencies $250$ Hz, $500$ Hz, and $1$ kHz, incident from $\mathbf{n}_k = (\pi/2, \pi/4)$ and reconstructed using 3D PSR. The simulations use emulated microphone recordings using the sound field extrapolated $15.5$ cm away from the origin. The maximum order of extrapolation used were $N = 1$, $N = 2$, and $N = 3$, for $250$ Hz, $500$ Hz, and $1$ kHz waves, respectively. Fig. 4.4 shows the angular reproduction error as a contour plot and intensity vectors as a vector plot for the tested cases. The average directional errors calculated within a spherical volume of radius $0.2$ m around the center of the simulated volume are $0.66°$, $3.68°$, $27.75°$, for the three tested frequencies respectively. Figures also indicate that the directional accuracy decreases as the frequency increases.

In order to assess the dependence of system accuracy on the direction of incidence, monochromatic plane waves ($f = 500$ Hz) incident from different directions in the horizontal and the median planes are also simulated. Fig. 4.5 shows average absolute angular errors (in degrees) within a sphere of radius $0.2$ m as polar plots for incidences in the horizontal and median planes. The maximum order used in the extrapolation

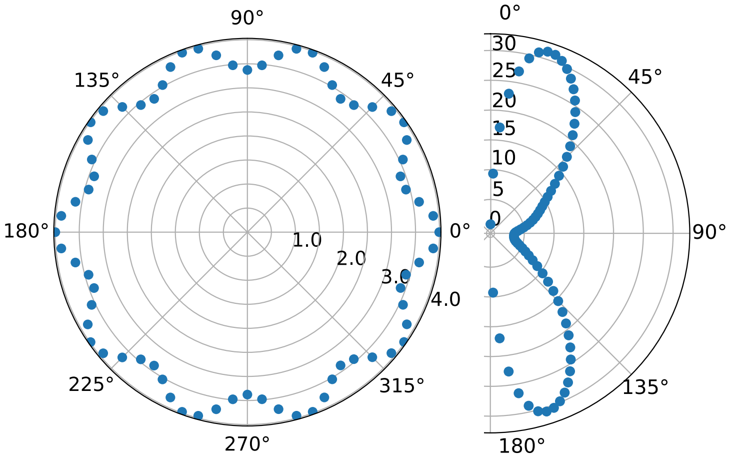


(a) (b) (c)

**Figure 4.4:** Reconstructed sound field around the optimal listening area for a monochromatic plane wave with (a) $f = 250$ with $N = 1$, (b) $f = 500$ with $N = 2$, and (c) $f = 1$ kHz with $N = 3$. The contour plot shows the direction error (in degrees). The vector plot shows the local active intensity. The circle shows a circular region with a radius of $0.2$ m. Note that the color bars are different across figures. © 2019, IEEE

was $N = 2$. The figures indicate a good reproduction accuracy in terms of active intensity directions, specifically in the horizontal plane.

Potentially improved performance could be achieved by including two loudspeakers positioned at the apex and nadir of the sphere, respectively. While average absolute angular error can be high for elevated sources, the perceptual impact is not likely to be high since localisation blur for directions above the horizontal plane are generally higher [24].



(a) (b)

**Figure 4.5:** Average absolute angular error (in degrees) for different directions of incidence for a monochromatic plane wave with $f = 500$ Hz and $N = 2$ in (a) horizontal plane ($\theta = \pi/2$ and $\phi \in [0, 2\pi)$), and (b) median plane ($\theta \in [0, \pi]$ and $\phi = 0$). © 2019, IEEE

# CHAPTER 5

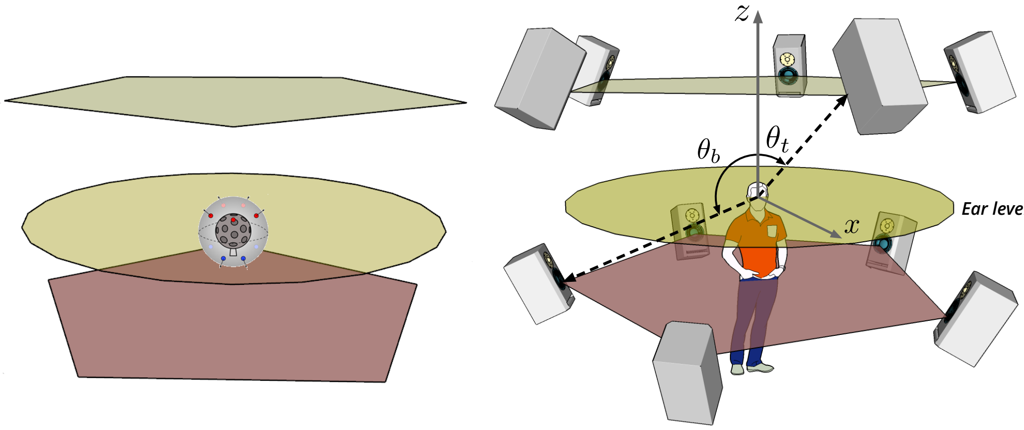
# 3D PSR VIA SOUND FIELD EXTRAPOLATION OF REAL RECORDINGS

In this chapter, real sound scenes recorded by Eigenmike© are used for 3D PSR via sound field extrapolation. During recording process, Eigenmike is positioned at the center of a sound scene. Later, listener are placed at the center of loudspeaker array and reconstructed audio is played through the 10 channel loudspeaker system. Recording and reproduction schemes are shown in Fig. 5.1; block diagram of the proposed method is given in Fig. 5.2.

Main difference of using real recordings rather than simulated monochromatic plane waves involves the calculation of SHD coefficients, as also explained in Section 2.2 in detail. For convenience, the procedure of calculating SHD coefficients described in Section 2.3.2 for rigid spherical surfaces are reintroduced here in a more specific manner. In order to obtain SHD coefficients, first, audio signals should be converted to frequency domain.

As a rigid spherical microphone array, Eigenmike32® provides 32 channel of audio recording files in uncompressed, PCM-encoded wave format for each 32 microphone positioned over its surface.

In most of the audio applications, time domain signal is converted to frequency do-



(a) (b)

**Figure 5.1:** (a) Recording Stage (b) Reconstruction Stage.

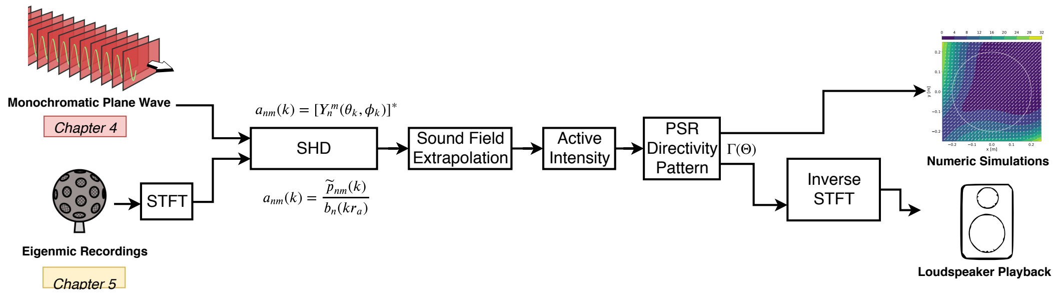


**Figure 5.2:** The flow diagram of the algorithm for plane waves and real recordings.

main by applying appropriate Fourier Transform algorithms in order to capture the time varying frequency composition (*i.e. frequency spectrum*). In this work, STFT is used to obtain the change in frequency and phase content of a non-stationary signal by evaluating Fourier Transform at each local subsections of a signal.

STFT of the input audio signals are obtained with the following equation,

$$\textbf{STFT}\{a_i(n)\} \equiv A_m(\omega) = \sum_{n=0}^{L} a_i(n) w\left[n-m\right] e^{-j\omega n} \tag{5.1}$$

where $A_m(\omega)$ is the Short Time Fourier Transform (STFT) of the input signal $a_i(n)$ and $Q = 32$ is number of microphones. The window function is represented with $w\left[n-m\right]$, chosen as Hann window for this work. $7/8$ overlap ratio is used.

The input signal $a_i(n)$, for $i = 1, \ldots, 32$, consists of 32 channel audio recording with $L$ samples of each,

$$a_i(n) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{32,1} & \dots & \dots & a_{32,n} \end{bmatrix}_{32\times L} \tag{5.2}$$

Resulting STFT matrix $A_m(\omega)$ is a three dimensional matrix of the form $32 \times T \times N$, where $N$ is the associated time bins and $T$ is the frequency bins depend on STFT size used (1024), total number of samples and sampling rate of the signal. As an example, $1024$ sized STFT of a $30$ seconds audio signal with sampling rate $44100$ Hz gives the following number of frequency and time bins:

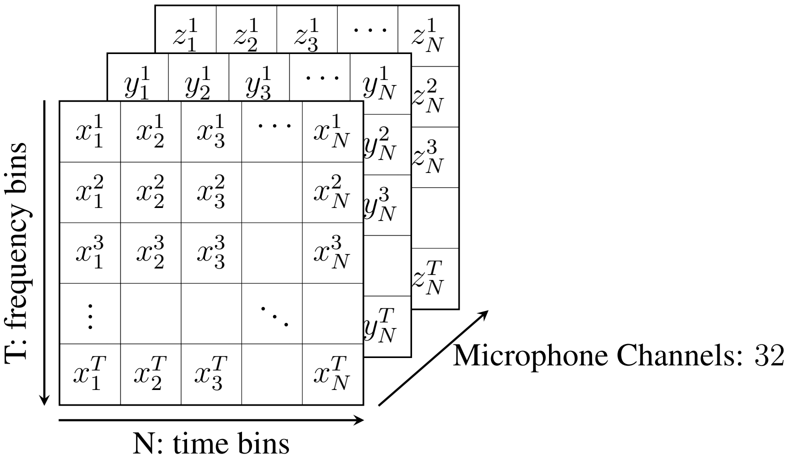
$$T = \frac{size}{2} = 512 \tag{5.3}$$

where each frequency bin are $f = \frac{44100}{1024} = 43.066$ Hertz apart ( $T_i = 43.066i$ Hz,

$i = 1, \ldots, T$).

$$N = \frac{time \times Fs}{size} = 30s \times \frac{44100}{s} \times \frac{1}{1024} \approx 1292 \tag{5.4}$$

$A(m, \omega)$ =


$z_1^1$ $z_2^1$ $z_3^1$ $\cdots$ $z_N^1$
$z_N^2$
$z_N^3$
$z_N^T$
$y_1^1$ $y_2^1$ $y_3^1$ $\cdots$ $y_N^1$
$y_N^2$
$y_N^3$
$y_N^T$
$x_1^1$ $x_2^1$ $x_3^1$ $\cdots$ $x_N^1$
$x_1^2$ $x_2^2$ $x_3^2$ $x_N^2$
$x_1^3$ $x_2^3$ $x_3^3$ $x_N^3$
$\vdots$ $\ddots$
$x_1^T$ $x_2^T$ $x_3^T$ $x_N^T$
T: frequency bins
N: time bins
Microphone Channels: 32

Afte obtaining STFT of the signal, sound recordings are represented in spherical harmonic domain with linear combination of spherical harmonics (5.5).

$$\widetilde{p}_{nm}(k) = \sum_{q=1}^{Q} w_q A_m(\omega) \left[Y_n^m(\theta, \phi)\right]^* \tag{5.5}$$

where $w_q$ are the quadrature weights, A is the complex-valued amplitude at a given TF bin.

After obtaining $\widetilde{p}_{nm}(k)$, scattering effect of the rigid sphere is removed with the method explained in Sec. 2.3 and Sec. 2.3.3, giving $a_{nm}(k)$ coefficient to represent the sound field,

$$a_{nm}(k) = \widetilde{p}_{nm}(k)/b_n(kr_a) \tag{5.6}$$

Having obtained SHD coefficients, pressure field and the particle velocity is calculated as described in the Section 4.2. Using those values, intensity vectors are obtained at virtual microphone positions with Equation 4.9, rewritten as

$$\mathbf{I}_{virtual}(k, \mathbf{r}) = \frac{1}{2} \operatorname{Re}\left\{p(k, \mathbf{r})\mathbf{u}^*(k, \mathbf{r})\right\}. \tag{5.7}$$

Notice that $I_{virtual}$ are the intensity vectors calculated for each time-frequency bin representing the sound field for each 10 virtual locations. This means that, for every time-frequency bin constructed according to chosen STFT algorithm, there are 10 different intensity vector calculated at 10 corresponding extrapolation locations (*i.e.* virtual microphone coordinates shown in Fig. 4.2).

$\mathbf{I}_{virtual}(k, \mathbf{r})$ is in the form of a three dimensional matrix, this time having 10 channels instead of the 32 channel input matrix $STFT\{a_i(n)\}$. In summary, dimensions of the input signal $32 \times T \times N$ becomes $25 \times T \times N$ after spherical harmonic decompostion; and finally output matrices of $10 \times T \times N$ are obtained via extrapolation:

$$32 \times T \times N \longrightarrow 25 \times T \times N \longrightarrow 10 \times T \times N \tag{5.8}$$

The signals that will be played back by the loudspeakers, $p_{\text{rec}}$, are generated in Equation 4.12 as,

$$p_{\text{rec}}(k, \mathbf{r}) = \Gamma(\Theta_{\mathbf{r}})p(k, \mathbf{r}). \tag{5.9}$$

These signals are also in the frequency domain in the form of $10 \times T \times N$. To convert frequency domain signals back into time domain signals, Inverse STFT (ISTFT) algorithm is used. As a result, real valued output signals S are obtained to be played directly through 3D PSR setup of $j = 10$ loudspeakers, without any addional mixing or processing.

$$\mathbf{S} = \mathbf{ISTFT}\{p_{rec}\} \equiv \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{10,1} & \cdots & \cdots & s_{10,n} \end{bmatrix}_{10 \times L} \tag{5.10}$$

where $s_j(n)$ for $j = 1, 2, \ldots, 10$ and $n = L$ where $L$ is the number of total samples of a recording.

# CHAPTER 6

# SUBJECTIVE EXPERIMENT: MULTI-CHANNEL AUDIO TEST FOR HOA VS. 3D PSR VIA SOUND FIELD EXTRAPOLATION

This chapter reports a subjective localisation experiment that aims to demonstrate the performance of 3D PSR in comparison with a state-of-the-art 3D audio technology, higher-order Ambisonics. Details of the localisation experiment as well as the obtained results are shown together with a discussion that follows. In the following:

- **PSR** refers to 3D PSR via sound field extrapolation for real recordings
- **AMB** refers to 2nd-order Ambisonics with maximum energy vector magnitude (Max-rE) using ALLRad decoding method [30].

The proposed work in this thesis is denoted as **PSR** (Chapter 4.2), whereas **AMB** (Ambisonics) was briefly explained in Section 2.4.4. The motivation behind this selection was that 2nd-order HOA is nominally the best state-of-the-art method that could allow a direct comparison with the employed multichannel setup.

## 6.1 Recordings

Three different anechoic recordings of musical instruments [31] were used in the experiments. These recordings were played through a Genelec 6010A studio monitor one by one, and recorded with the Eigenmike em32 microphone positioned at the center of the 3D PSR loudspeaker rig as shown in Fig. 6.1. The same three instrument recordings were played at four different directions on the sphere defined by the 3D PSR array as shown in Fig. 6.2. The directions of recorded sources with respect the listener's front direction are given in Table 6.1. Note that the look direction coincides with positive x-axis, with loudspeaker no.1 and also with the P2 source position. METU SPARG Lab has non-parallel walls and ceiling-ground, making difficult to understand x-axis/look direction by only looking at the setup photos. Origin of the sphere in Fig. 6.2 is the center point of the Eigenmike microphone in Fig. 6.1 while the x-axis is directed towards the P2 position.

Recordings are made under tightly controlled acoustic conditions in METU SPARG Lab which has a reverberation time of $T30 = 80$ ms and a background noise level less than $40$ dB SPL.

The three recordings that were used are given below:

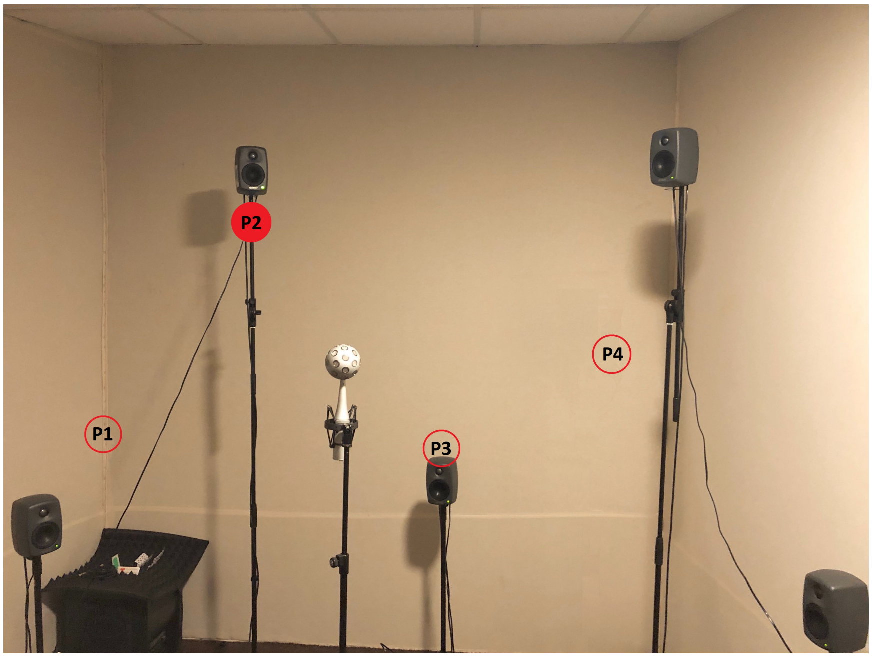
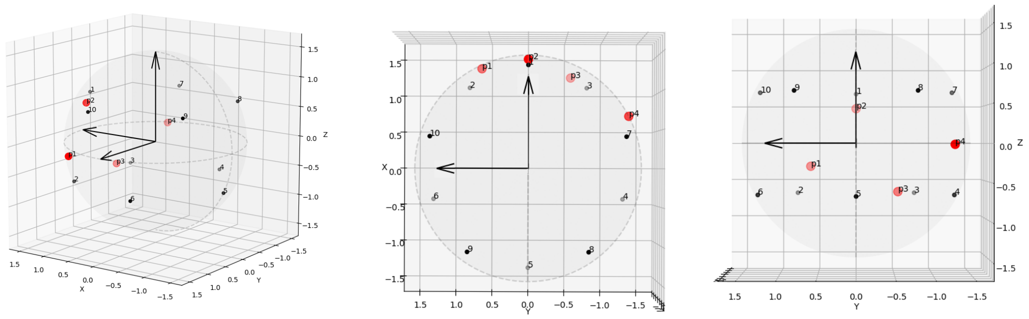


**Figure 6.1:** Recording stage of anechoic instrument samples in METU SPARG Lab. Approximate source directions P1, P2, P3 and P4 is denoted with red circles.



(a)

(b)

(c)

**Figure 6.2:** 4 different positions of sample anechoic instruments recordings, played only by single loudspeaker at a time. (a) Isometric view, (b) Top view: $\theta = 90$, (c) Rear view: $\phi = 180$

**Table 6.1:** Source positions of anechoic recording

| **Position no.** | $\theta$ | $\phi$ | **Radius [m]** |
|---|---|---|---|
| **p1** | 102.1 | -25 | 1.498 |
| **p2** | 71.6 | 0 | |
| **p3** | 116.1 | 25 | |
| **p4** | 90.6 | -62.5 | |

1. **b:** Latin American rhythm played with an African bongo [31].
2. **vc:** Cello recording of Theme by Weber [31].
3. **fl:** Flute passage from an aria of Donna Elvira from the opera Don Giovanni, W. A. Mozart. [32]

## 6.2 Methodology

Sound fields recorded using the described setup were reconstructed using AMB and PSR methods through 3D PSR loudspeaker array setup (see Fig. 3.1). Subjects were positioned at either i) the center of the loudspeaker array, or ii) at an off-center location that is $60$ cm away from the center. They listened to the reconstructed sound fields one by one, and showed the directions of the virtual sound sources (i.e instrument sounds) that they perceived, after listening to each test sample (recording). Since the head movement dramatically changes the perception as spatial cues become more natural and comfortable to be detected and result in a real-time adjusting of sound localization according to the head-body movement, the subjects were encouraged to turn their heads to localize sources if they needed to.

Tracking the direction data is realized using Leap Motion Controller, as will be explained in the following sections. Three sound recordings each reconstructed for four different positions, with two different reconstruction method (PSR and AMB) and two listening positions (center, off-center) resulted in a total of 48 questions per person.

Thirteen subjects with no reported hearing impairments participated in the experiment. 48 unique test items per listener resulted in $13 \times 48 = 624$ perceived sound source location in total. Subjects listened to the recordings in randomized order such that neither same instrument nor the same sound source position were used in successive presentations.

## 6.3 3D PSR Algorithm Parameters and Experiment Setup

Physical parameters for experiment setup are as follows:

1. Height of the upper pentagonal loudspeaker layers from the ground $= 2.24$ m
2. Height of the lower pentagonal loudspeaker layers from the ground$= 0.9$ m
3. Height of the center of the sphere composed of two pentagonal layers, i.e. height of the listening position (ear level) $= 1.57$ m
4. Radius of the sphere, i.e distance between center and any loudspeaker position $= 1.5$ m
5. Radius of the pentagon, i.e distance between center and any vertex of the pentagon $= 1.34$ m

The loudspeaker array consists of 10 Genelec 6010A studio monitors. Two MOTU 896 Mk3 Hybrid Audio Interfaces were used as an aggregate device for multi-channel audio routing, using Mac Pro running Mac OS X v$10.10$. Python 2.7 was used to design the experimental software. `sounddevice` and `pyaudio` modules were used for multi-channel audio playback whereas Python bindings of Leap Motion SDK version 2.3 were used for real-time direction data tracking in the pointing task explained below.

Sampling rate of $44.1$ kHz is used both for recording and reproduction stages. For 3D PSR, the following parameters were used for the STFT as explained in Chapter 5:

(i) Window size = $1024$

(ii) Overlap ratio = $7/8$

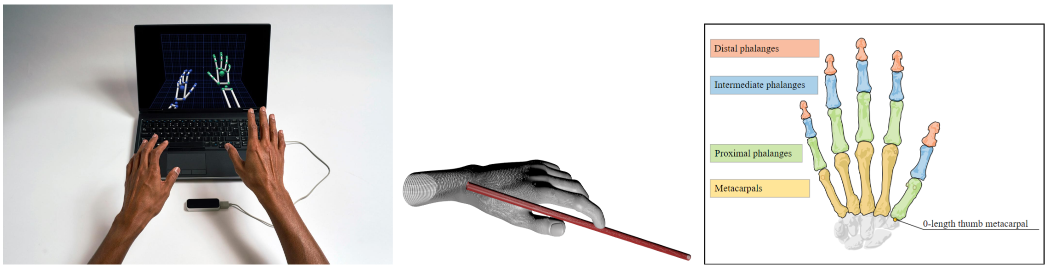
(iii) Window function = Hann

## 6.4 Ambisonics Decoded Signals

**AMB** signals were obtained using the following procedure:

1. Recorded input signals, $a_i(n)$, are encoded into an HOA format signal in Eigenstudio® interface [33] by mh acoustics.
2. Decoding matrix is obtained with the 3D PSR loudspeaker array configuration, using the IEM Plug-in Suite [34] using REAPER digital audio workstation (DAW) software [35].
3. Decoding matrix is multiplied with input signal, giving the final Ambisonics decoded output signal to be played through the loudspeakers.

For encoding and decoding; the parameters ACN and N3D were used for Ambisonic channel ordering and normalization parameters. As a decoding strategy, Max-Re (maximum energy vector magnitude) weighting [30] was used.



(a) (b) (c)

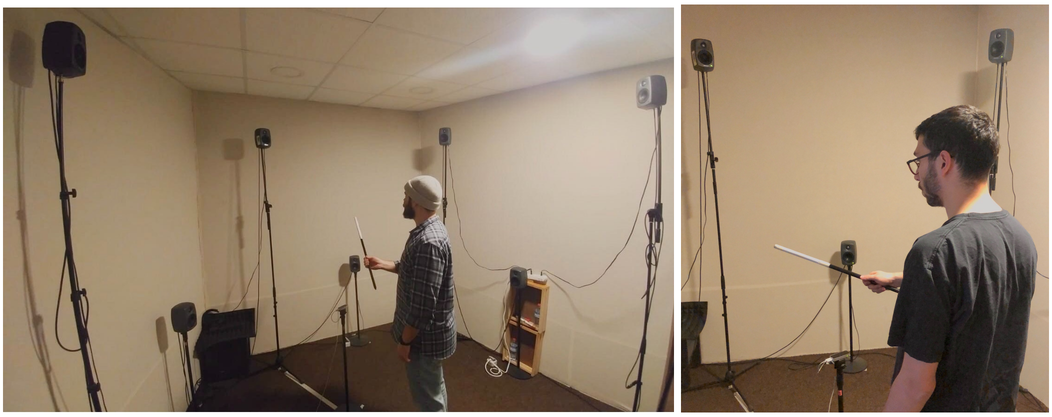
**Figure 6.3:** (a) Leap motion controller with hand tracking, (b) An example pointable object for tool tracking, (c) Finger bones that leap motion is able to track

## 6.5 Direction Data Tracking with Leap Motion

Leap motion [36] is an optical hand tracking controller (Fig. 6.3a) that captures the movement of hands with two near-infrared cameras operating at 120 Hz. The controller is capable of tracking hands within a 3D inverted pyramid shaped interactive zone extending from the device in a 120°×150° field of view with approximately 60 cm height.

Leap motion also has the ability to track tool-like objects such as a pencil, a stick, a rod, or a baton when grabbed by hand as shown in Fig. 6.3b. To obtain the data of the direction which subjects are pointing with their hand, index finger tracking was used in a first set of trials. As shown in Fig.6.3c, position data of the base and tip of the finger bones; i.e distal, intermediate and proximal phalanges were used to obtain direction vector. It is considered that the directions calculated using the position data of the finger bones may result in angle deviations due to the short length of finger bones as well as differences between each individual's hand anatomies. To gather more stable and less fluctuating direction data, a second set of trials were performed using tool tracking instead of finger tracking. Consequently, it is observed that pointing tasks can be tracked with higher accuracy and stability by using tool tracking.

Fig. 6.4 shows the usage of a rod like tool during the experiment. A sample visual data of tool tracking is given in Fig. 6.6. Leap motion controller is positioned 50 cm front to the listening position at a height of 87cm. Subjects indicated the sound sources they have perceived, by pointing out with their hands while grabbing the tool. In Fig. 6.5, the usage of a sample tool tracking data is explained. Red square shows 3D PSR loudspeaker setup origin and the center of the sphere. This point also representing the listening position which is the approximate location between the two ears of the listener. Blue square shows the origin of the leap motion, coincide with the controller. Leap motion provides the position data according to its own reference frame, with respect to the origin (blue square) of the device. Since there is a distance between the device origin and the tool tip position, the position vector starts from the tip of the rod, shown as the black circular point. Tool direction vector is plotted with blue arrow and intersects the sphere at the location of the blue star. By drawing a



(a)

(b)

**Figure 6.4:** Leap motion controller with rod like tool object for the experiment

new direction vector (red) from 3D PSR origin (red square) to this intersection point, inclination ($\theta$) and azimuth angles ($\phi$) with respect to listening position are calculated conveniently. Grayed points show 10 loudspeakers in the 3D PSR setup.

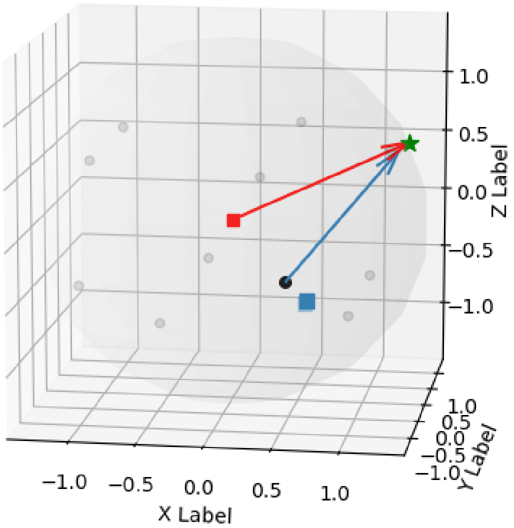
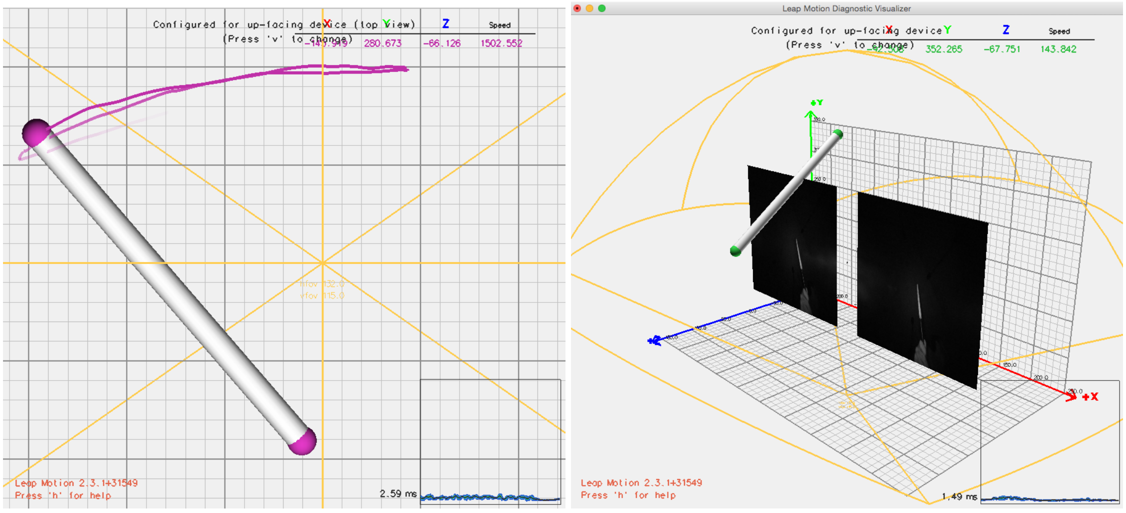


**Figure 6.5:** 3D PSR Setup origin and leap motion controller origin

## 6.6 Experiment Results

Excluding missing data from the overall data gathered, 586 individual direction vectors are obtained. From a single direction vector, both inclination and azimuth angle pairs ($\theta$, $\phi$) can be extracted. Thus, these two data are not evaluated independently although the accuracy of the inclination and azimuth angles are expected to differ



(a)

(b)

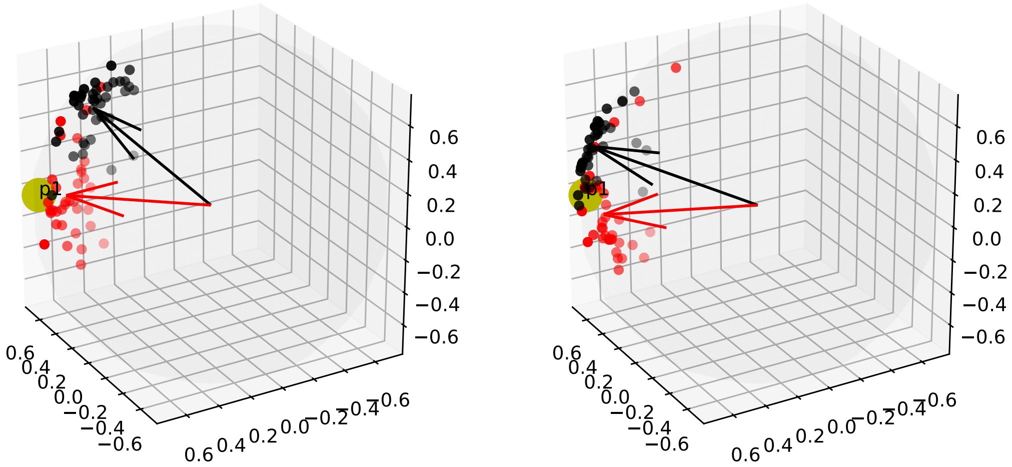
**Figure 6.6:** Visual representation of the tracked tool by Leap Motion Visualizer

considerably. To evaluate the accuracy of the angle pairs as dependent variables; 5-parameter Fisher-Bingham distribution, also known as Kent distribution was fitted to pooled direction observations. This method is suitable for fitting an asymmetrically distributed data over unit sphere and obtaining a bivariate probability distribution. Three parameters of the Kent distribution are relevant in the context of this study: $\gamma_1$ is the mean direction vector for the data points, $\kappa$ is the parameter that determines the concentration or spread of the distribution, and $\beta$ is the parameter which determines the ellipticity of the distribution.

Experimental data from subjects are plotted over the unit sphere together with the true sound source directions in Figs. 6.7, 6.8, 6.9 and 6.10. Black and red arrow vectors represent mean direction vectors, $\gamma_1$, for AMB and PSR, respectively. Left column shows the results for the center listening position while right column shows the corresponding off-center position for the same sound source position. The direction data obtained for AMB reconstructed sound recordings are represented with black points whereas PSR reconstructed signals are specified with red points.

The inclination and azimuth angles ($\theta$, $\phi$) of the mean direction vectors are compared with the directions of the real sound sources (P1, P2, P3 and P4) in Table 6.2. The angle between these direction vectors are calculated using the dot product, and given in Table 6.3 as an angular errors in degrees.

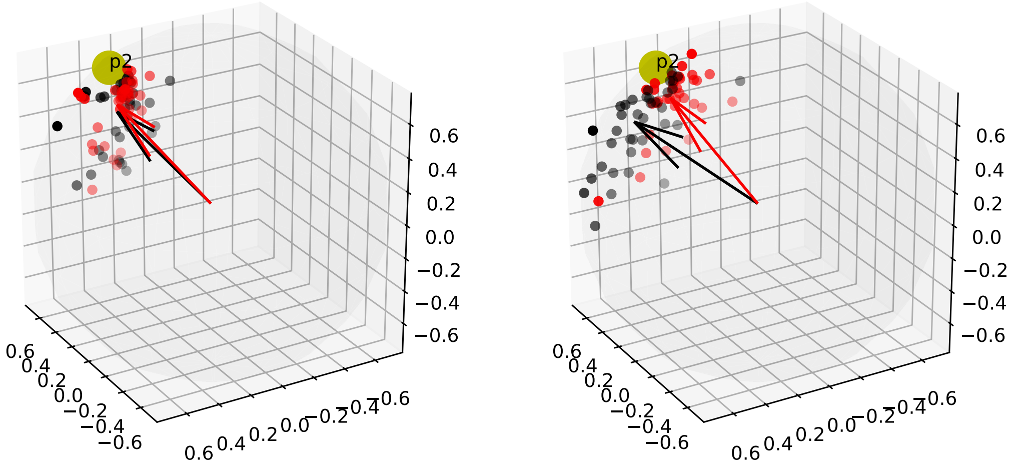
The parameters $\kappa$ and $\beta$ that represent the spread and ellipticity of the fitted Kent distributions are given in Table 6.4 and Table 6.5, respectively.



(a)

(b)

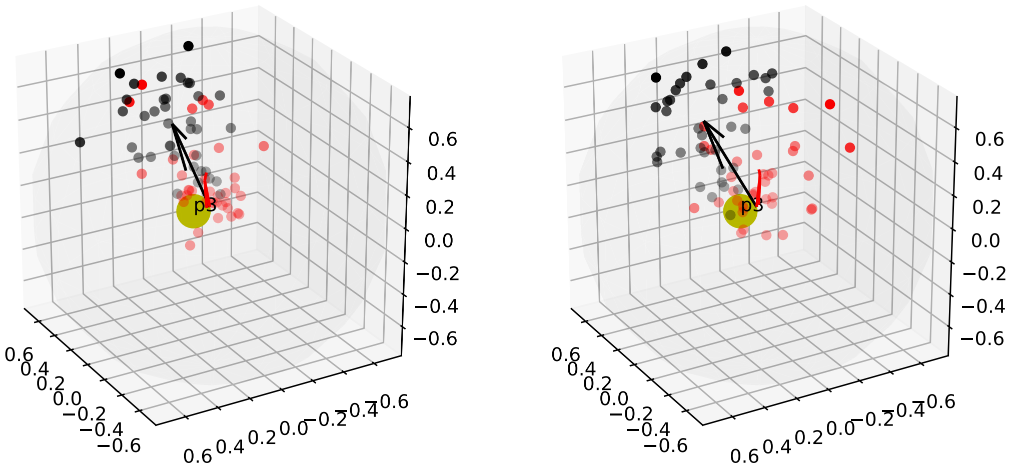
**Figure 6.7:** Perceived directions for Ambisonics (black) and PSR (red) for position: P1. (a) center (b) off-center listening locations.



(a)

(b)

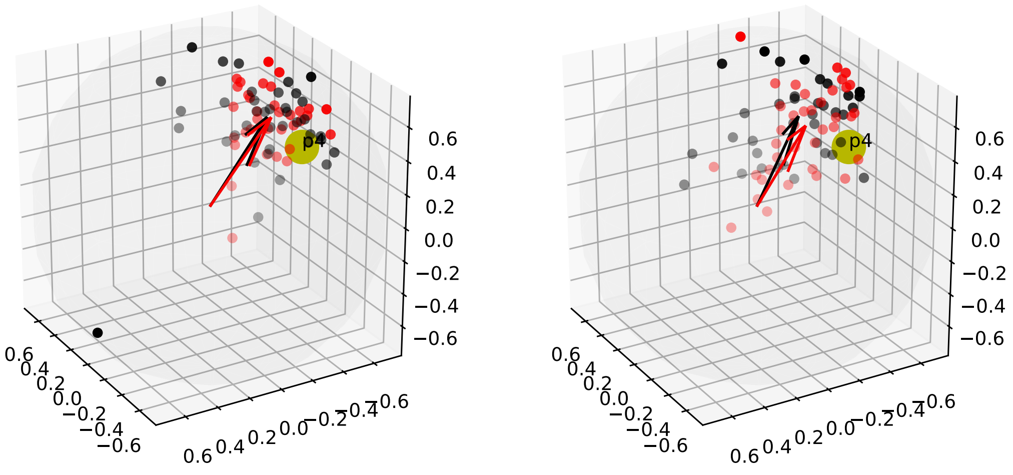
**Figure 6.8:** Ambisonics (black) and PSR (red) for P2, (a) center (b) off-center listening positions.



(a)

(b)

**Figure 6.9:** Ambisonics (black) and PSR (red) for P3, (a) center (b) off-center listening positions.



(a)

(b)

**Figure 6.10:** Ambisonics (black) and PSR (red) for P4, (a) center (b) off-center listening positions.

**Table 6.2:** Real source directions vs mean directions calculated with Kent distribution.

| | Center | | Off-center | |
|---|---|---|---|---|
| | $\theta$ | $\phi$ | $\theta$ | $\phi$ |
| **P1** | 102.1° | 25° | 102.1° | 25° |
| **AMB** | 72.5° | 15.1° | 77.4° | 41.4° |
| **PSR** | 100.5° | 27.9° | 104.3° | 35.3° |
| **P2** | 71.6° | 0° | 71.6° | 0° |
| **AMB** | 78.1° | 3.8° | 78.3° | 16.4° |
| **PSR** | 75.1° | 4.1° | 72.9° | 1.8° |
| **P3** | 116.1° | -25° | 116.1° | -s25° |
| **AMB** | 85.2° | -17.2° | 83.4° | -11.9° |
| **PSR** | 105.9° | -28.2° | 104.2° | -31.2° |
| **P4** | 90.6° | -62.5° | 90.6° | -62.5° |
| **AMB** | 81.8° | -49.8° | 82.2° | -44.3° |
| **PSR** | 81.7° | -51.2° | 85.6° | -46.8° |

**Table 6.3:** Angle between real source directions and mean directions (Kent distribution)

| **Angular Errors** | **Center** | | **Off-center** | |
|---|---|---|---|---|
| **Source Positions** | **AMB** | **PSR** | **AMB** | **PSR** |
| **P1** | 31.2° | 3.2° | 29.6° | 10.3° |
| **P2** | 7.4° | 5.3° | 17.2° | 2.2° |
| **P3** | 31.8° | 10.5° | 35.1° | 13.2° |
| **P4** | 15.4° | 14.3° | 19.9° | 16.5° |

**Table 6.4:** $\kappa$ parameter obtained from Kent distribution.

| $\kappa$ | **Center** | | **Off-Center** | |
|---|---|---|---|---|
| **Source Positions** | **AMB** | **PSR** | **AMB** | **PSR** |
| **P1** | 33.1° | 33.6° | 28.6° | 27.3° |
| **P2** | 31.1° | 49.2° | 21.4° | 35.9° |
| **P3** | 22.2° | 29.9° | 17.4° | 20.3° |
| **P4** | 10.6° | 28.1° | 15.1° | 19.3° |

**Table 6.5:** $\beta$ parameter obtained from Kent distribution.

| $\beta$ | **Center** | | **Off-Center** | |
|---|---|---|---|---|
| **Source Positions** | **AMB** | **PSR** | **AMB** | **PSR** |
| **P1** | 3.1° | 8.6° | 3.6° | 7.5° |
| **P2** | 4.5° | 10.2° | 4.8° | 6.7° |
| **P3** | 5.7° | 8.1° | 3.9° | 2.3° |
| **P4** | 0.3° | 5.1° | 2.7° | 6.3° |

During subjective experiments; in addition to pointing task, subjects are also asked to rate their confidence on perceived sound direction. Every subject evaluated each audio sample on a 5-point rating scale with:

- 5 = Very high confidence about the perceived location of the sound source
- 1 = Least confidence about the direction of the sound source.

**Table 6.6:** Confidence rates with respect to instruments.

| | **Confidence Rate - Instruments** | | | |
|---|---|---|---|---|
| | **Center** | **Off-center** | **Center** | **Off-center** |
| **Instrument** | **Ambisonics** | **PSR** | **Ambisonics** | **PSR** |
| **Bongo** | 3.96 | 4.17 | 3.29 | 4.08 |
| **Flute** | 4.02 | 4.38 | 3.73 | 4.44 |
| **Violoncello** | 4.26 | 4.39 | 3.80 | 4.36 |

**Table 6.7:** Confidence rates with respect to position

| | Confidence Rate - Positions | | | |
|---|---|---|---|---|
| | Center | Off-center | Center | Off-center |
| **Source Position** | **Ambisonics** | **PSR** | **Ambisonics** | **PSR** |
| **P1** | 4.18 | 4.23 | 3.62 | 4.27 |
| **P2** | 4.28 | 4.42 | 3.59 | 4.49 |
| **P3** | 4.05 | 4.37 | 3.76 | 4.22 |
| **P4** | 3.81 | 4.10 | 3.50 | 4.09 |

## 6.7 Discussion

The results of the experiment indicate that PSR provides a more accurate subjective localization on average. In this respect, it performs better than 2nd-order Ambisonics both at the center and off-center listener positions (see Table 6.2). As can be seen in Table 6.3, average angula errors for PSR are significantly less than the errors for Ambisonics indicating an even spread around the true source direction. Although the average localisation error is smaller for PSR, Table 6.4 indicates that the spread of the data is in general higher for PSR than AMB, meaning the data is more concentrated for the AMB method. Similarly, AMB results in a more elliptic distribution (see Table 6.5). Although the spread is small, sound source localization accuracy of the 2nd order Ambisonics is not satisfactory. From this, it may be inferred that Ambisonics is forcing a dominant location for the sound source yet lacking the true directional information.

In addition, subjects stated that they give their response more confidently for PSR reconstructed samples. This can be validated by looking the confidence ratings at the Tables 6.7 and 6.6. Regardless of position or instrument type, PSR achieved higher confidence levels. Since the confidence and spread results are not in parallel, we speculate that other factors such as tonal coloration may have played a part in this outcome. On the other hand, the accuracy of the pointing task by using the Leap Motion Controller is also unclear and this may have contributed to the high level of spread observed for both **AMB** and **PSR**.

**AMB** signals employ several different correction strategies (e.g. for near-field sources) and different decoding approaches to adapt it to arbitrary speaker layouts, whereas **PSR** is a raw method excluding any any such correction. In this work, in order to present similar experiment samples in terms of audio quality and content, and to prevent the subjects from noticing that there are two distinct methods in the experiment, frequency response of the reconstructed signals were very roughly equalised. To this end, a simple equalization is added to PSR recordings, attenuating the low frequency ($F < 100$ Hz) and amplifying the high frequency content ($F > 4000$ Hz).

# CHAPTER 7

# CONCLUSION

## 7.1 Conclusion

An extension of perceptual soundfield reconstruction to three dimensions was presented in this thesis. Such an extension is not trivial with classical microphone setups due to the required number of microphones and the necessity to equalize them. We proposed a conceptual framework that enables 3D PSR recordings from rigid spherical microphone arrays via sound field extrapolation. Numerical simulations with monochromatic sound fields using the directivity pattern designed for 2D-PSR show that 3D PSR via sound field extrapolation is feasible even with a straightforward application of 2D PSR directivity pattern that is confined to the horizontal plane.

We then realised 3D PSR via sound field extrapolation using real recordings obtained from an Eigenmike em32 and compared the results with a widely used state-of-the-art higher-order Ambisonics method in a multi-channel subjective localisation experiment. The results are promising for the considered perceptual attribute of subjective localization.

Notice also that the proposed extrapolation method is not limited to PSR microphone arrays, but can be used with any near-coincident array, including those typically used in the audio engineering community [2, 3].

## 7.2 Possible Use Cases of 3D PSR via Sound Field Extrapolation

The work reported in this thesis facilitates spatial audio production by decreasing the number of microphones and loudspeakers that should be used. Due to the fact that the method proposed is realised only with a single rigid spherical microphone array, it provides an opportunity to emulate any near coincident microphone array combinations, which gives great recording flexibility, especially for real recordings of live events. In this respect recordings made with RSMAs can act as a master recording where infinitely many different combinations of near-coincident microphone setups can be derived.

### 7.3 Future Work

The directivity pattern used here was designed using interchannel time and level differences needed for accurate horizontal localization. Obtaining an appropriate directivity pattern taking different psychoacoustic phenomena into account governing the perception of source elevation to enhance the reproduction in the vertical plane is the subject of future work.

Moreover, different extrapolation methods, as well as a formal analysis of the error measure of the extrapolation step, including white noise gain (WNG) analysis, can be assessed. For the plane-wave scenario, we expect a more significant error in extrapolated positions further away from the center. This means, in real recordings, there may be a trade-off between noise gain/extrapolation error and taking advantage of the perceptual improvement of time-intensity reproduction.

3D PSR has similarities to Dirac/SDM/SIRR [37][38][39], but with the crucial difference that both pressure and direction are estimated at a position different from the center of the array, so as to include time-delays between individual channels.

Like many other state-of-the-art methods, 3D PSR may also have an imaginary loudspeaker exactly at the top and at the bottom, to increase the perception of elevated sound sources. This may be achieved for example by creating a virtual speakervia vector base amplitude panning (VBAP).

In its current version, 3D PSR is not ideal for diffuse sound fields. Future work would involve a more thorough analysis of STFT parameters, especially the audio frame lengths and window functions. A more suitable window function with a higher overlap percentage may smooth out sharply changing intensity vectors due to diffuse sound fields. Reconstruction of this type of diffuse fields can further be enhanced by proposing a hybrid model such as direct-diffuse separation that enables using different parameters or methods for different subsections of the sound. In this way, 3D PSR can perform better also for recordings made in highly diffuse scenes with multiple incoherent sources by processing the contribution of diffuse components separately.

# REFERENCES

[1] E. Erdem, E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Perceptual soundfield reconstruction in three dimensions via sound field extrapolation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8023–8027, May 2019.

[2] J. Eargle, *The microphone book.* Focal Press, 2004.

[3] H. Hacıhabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. Smith, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36–54, 2017.

[4] V. Pulkki, "Virtual sound source positioning using vector-base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, June 1997.

[5] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, 2005.

[6] V. Pulkki and J. Merimaa, "Spatial impulse response rendering ii: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20, 2006.

[7] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, 2007.

[8] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28, 2013.

[9] J. D. Johnston and Y. H. Lam, "Perceptual soundfield reconstruction." AES $109^{th}$ Conv., Preprint #2399 Los Angeles, USA, Sep. 2000.

[10] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 21, pp. 1653–1665, August 2013.

[11] E. De Sena and Z. Cvetković, "A computational model for the estimation of localisation uncertainty," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process. (ICASSP-13)*, (Vancouver, Canada), pp. 388–392, May 2013.

[12] E. De Sena, *Analysis, Design and Implementation of Multichannel Audio Systems.* PhD thesis, King's College London, 2013.

[13] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8 of *Springer Topics in Signal Processing.* Berlin, Heidelberg: Springer-Verlag, Oct. 2015.

[14] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography.* Academic Press, 1999.

[15] mh acoustics, "em32 eigenmike® microphone array release notes (v17.0)," 2013.

[16] J. Meyer and G. W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process. (ICASSP-93)*, (Minneapolis, USA), April 1993.

[17] A. D. Blumlein, "Improvements in and relating to sound-transmissions, sound-recording and sound-reproducing systems." U.K. Patent 394,325,1931.Reprinted in Stereophonic Techniques (AES, New York, 1986) B1, 1931.

[18] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention 114*, Audio Engineering Society, 2003.

[19] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973.

[20] J. Meyer and G. W. Elko, "A spherical microphone array for spatial sound recording," *J. Acoust. Soc. Amer.*, vol. 111, no. 5, pp. 2346–2346, 2002.

[21] M. A. Poletti, "A unified theory of horizontal holographic sound systems," *J. Audio Eng. Soc.*, vol. 48, pp. 1155–1182, Dec. 2000.

[22] N. V. Franssen, *Stereophony*. Philips Research Laboratories, 1964.

[23] M. Williams and G. Le Du, "Microphone array analysis for multichannel sound recording." presented at the 107th Audio Eng. Soc. Conv., Preprint #4997, New York, USA, September 1999.

[24] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, Massachusetts: MIT Press, 1997.

[25] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "A generalized design method for directivity patterns of spherical microphone arrays," in *Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process. (ICASSP-11)*, (Prague, Czech Republic), May 2011.

[26] J. Benesty and C. Jingdong, *Study and design of differential microphone arrays*, vol. 6. Springer Science & Business Media, 2012.

[27] J. D. Johnston and E. R. Wagner, "Microphone array for preserving soundfield perceptual cues." United States Patent, US 6,845,163 B1, Jan. 2005.

[28] G. L. Rosen and J. D. Johnston, "Automatic speaker directivity control for sound field," in *Proc. AES 19th Intern. Conf.*, (Schloss Elmau, Germany), Jun. 2001.

[29] P. Samarasinghe, T. Abhayapala, and M. Poletti, "Wavefield analysis over large areas using distributed higher order microphones," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 3, pp. 647–658, 2014.

[30] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," 2012.

[31] Bang and Olufsen, "Music for archimedes. audio cd, 1992."

[32] J. P. Tapio Lokki and V. Pulkki, "Anechoic recordings of symphonic music." `https://users.aalto.fi/~ktlokki/Sinfrec/sinfrec.html`.

[33] mh acoustics, "Eigenstudio® application, a gui interface for control, record and playback." `https://mhacoustics.com/download`.

[34] Institute of Electronic Music and Acoustics, "IEM Plug-in Suite." `https://plugins.iem.at/`.

[35] Cockos, "REAPER Version 6.03," 2006.

[36] Leap Motion Company, "Leap Motion Controller," 2010.

[37] V. Pulkki, M.-V. Laitinen, J. Vilkamo, J. Ahonen, T. Lokki, and T. Pihlajamäki, "Directional audio coding-perception-based reproduction of spatial sound," 01 2009.

[38] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, pp. 16–27, 01 2013.

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