SPHERICAL HARMONICS BASED ACOUSTIC SCENE ANALYSIS FOR OBJECT-BASED AUDIO

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

BY

MERT BURKAY ÇÖTELİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
INFORMATION SYSTEMS

JANUARY 2021

Approval of the thesis:

SPHERICAL HARMONICS BASED ACOUSTIC SCENE ANALYSIS FOR OBJECT-BASED AUDIO

submitted by **MERT BURKAY ÇÖTELİ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems Department, Middle East Technical University** by,

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

Prof. Dr. Sevgi Özkan Yıldırım
Head of Department, **Information Systems** _____

Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu
Supervisor, **Modelling and Simulation, METU** _____

Date:

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: Mert Burkay Çöteli

Signature :

ABSTRACT

SPHERICAL HARMONICS BASED ACOUSTIC SCENE ANALYSIS FOR OBJECT-BASED AUDIO

Çöteli, Mert Burkay
Ph.D., Department of Information Systems
Supervisor: Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu

January 2021, 192 pages

Object-based audio relies on elemental audio signals from individual sound sources and their associated metadata to be reconstructed at the listener side. While defining audio objects in a production setting is straightforward, it is not trivial to extract audio objects from more realistic recording scenarios such as concerts. Thus, existing object-based audio standards also define scene-based formats alongside object-based representations that provide immersive audio, but without the flexibility provided by object-based audio. Presently, there is no reliable approach to transcode from scene-based format to object-based format. This thesis aims to develop acoustic scene analysis techniques to extract the directions of arrival of active sources and separate them from scene-based audio representations. Two DOA estimation methods and three source separation methods that use signals from rigid spherical microphone arrays are proposed for this purpose. The proposed methods allow analyzing scenes comprising multiple coherent or nearly coherent sources in highly reverberant and non-reverberant environments. We describe the algorithms, assess their performance objectively and subjectively and analyse their computational requirements.

Keywords: object-based audio, spherical harmonics, rigid spherical microphone arrays, direction of arrival estimation, source separation

ÖZ

NESNE TEMELLİ SES İÇİN KÜRESEL HARMONİK TABANLI AKUSTİK SAHNE ANALİZİ

Çöteli, Mert Burkay
Doktora, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Hüseyin Hacıhabiboğlu

Ocak 2021 , 192 sayfa

Nesne temelli ses, ayrı ses kaynaklarından gelen temel ses sinyallerine ve dinleyici tarafında yeniden yapılandırılacak ilgili meta verilerine dayanır. Bir prodüksiyon ortamında ses nesnelerini tanımlamak kolay olsa da, konserler gibi daha gerçekçi kayıt senaryolarından ses nesnelerini çıkarmak kolay bir işlem değildir. Bu nedenle, mevcut nesne temelli ses standartları, nesne temelli sesin sağladığı esneklik olmadan, sürükleyici ses sağlayan nesne temelli temsillerin yanı sıra sahne temelli formatları da tanımlar. Halihazırda, sahne temelli biçimden nesne temelli biçime dönüştürme için güvenilir bir yaklaşım yoktur. Bu tez, aktif kaynakların varış yönlerini çıkarmak ve bunları sahne temelli ses sunumlarından ayırmak için akustik sahne analizi teknikleri geliştirmeyi amaçlamaktadır. Bu amaçla, kapalı küresel mikrofon dizilerinden gelen sinyalleri kullanan iki varış yönü tahmin yöntemi ve üç kaynak ayırma yöntemi önerilmiştir. Önerilen yöntemler, yüksek derecede yankılanan ve yankılanmayan ortamlarda çok sayıda tutarlı veya neredeyse benzer kaynaklar içeren sahnelerin analiz edilmesine izin verir. Algoritmaları açıklamanın yanı sıra, performanslarını nesnel ve öznel olarak değerlendirip, hesaplama gereksinimleri de analiz edilmektedir.

Anahtar Kelimeler: nesne tabanlı ses, küresel harmonikler, kapalı küresel mikrofon dizileri, varış yönü tahmini, kaynak ayrıştırma

To my little daughter, wife, and my family

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Dr. Hüseyin Hacıhabiboğlu for his understanding, patience and supervision throughout this thesis. This thesis would not have been completed without his realistic, encouraging and constructive guidance. Many thanks also to Dr. Temel Engin Tuncer and Dr. Banu Günel Kılıç for their guidance during my Phd thesis. The work reported in this thesis is supported by the Turkish Scientific and Technological Research Council (TÜBİTAK) Research Grants 113E513 "Spatial Audio Reproduction Using Analysis-based Synthesis Methods" and 119E254 "Audio Signal Processing for Six Degrees of Freedom (6DOF) Immersive Media".

I would like to thank to all my friends and colleagues from ASELSAN for their understanding and continuous support during my thesis. Special thanks to the members of Spatial Audio Research Group (SPARG) for their support in the MUSHRA listening tests.

Finally, my deep and sincere gratitude to my family for their continuous and unparalleled love, help and support. I am grateful to my wife Süheyla and little daughter Beril for always being there for me with their love. I am forever indebted to my parents and brother for giving me the opportunities and experiences that have made me who I am. They selflessly encouraged me to explore new directions in life and seek my own destiny. This journey would not have been possible if not for them, and I dedicate this milestone to them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF SYMBOLS

c	Speed of sound
T_a	Ambient temperature
ρ_0	Ambient density
w	Radial frequency component
f	Frequency
ν	Particle velocity
p	Acoustic pressure
$\mathbf{I}$	Acoustic intensity
x	Position
$\mathbf{x}$	Position vector
k	Wave number
$\mathbf{k}$	Wave vector
r	Radius
θ	Inclination angle
ϕ	Azimuth angle
n	Spherical harmonics order
m	Spherical harmonics degree
$Y_n^m(\theta, \phi)$	Spherical harmonics
$j_n(kr)$	Spherical Bessel function
N	Maximum spherical harmonics order
$P_n^m(\cos(\theta))$	Legendre function
Θ	Angle between two vectors
a_{nm}	Spherical harmonic coefficients
Q	Number of microphones

w_q	Quadrature weights of microphones
$b_n(kr)$	Mode strength coefficient
τ	Time index for the time-frequency bin
$\varkappa$	Frequency index for the time-frequency bin
$\mathbf{y_N}$	Steered response function vector
P_N	Steered response power
K	HEALPix resolution order
$K_{\max}$	HEALPix maximum resolution order
κ	Spread parameter of von Mises function
ϱ	Dilation parameter of Kent distribution
B	Complement of residual energy ratio for RENT
η	Noise signal
Φ	Dictionary comprising Legendre kernels
Γ	Legendre kernel function
$\beth$	Short-time Fourier transform block size
$\beth_0$	Short-time Fourier transform overlap size
Υ	Analysis interval
M	Number of steering directions
D	Number of Legendre kernels in the dictionary
S	Number of sources
$\mathbf{\Phi}$	Power spectral density
$\mathbf{R}$	Covariance matrix
$\mathbf{\Sigma}$	Eigenvalue matrix
$\mathbf{U}$	Eigenvector matrix
$\dagger$	Pseudoinverse operator
$\mathcal{S}_{\text{RENT}}$	Single source time-frequency bins
$\aleph_{\text{RENT}}$	Noise/diffuse time-frequency bins
$\mathcal{V}_{\text{RENT}}$	Valid time-frequency bins

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Immersive audio is a set of techniques used to mimic the auditory scenes we hear in real life. It provides a "life-like" sound experience to the end-users, unlike the traditional methods. As this new audio experience surrounding the listener, creates a sensation over the audience for source arrival directions by creating credible auditory imitation. Disruptive innovations in recording, encoding, and transcoding between immersive audio formats gain importance for the industry thanks to ever increasing capacity of communication networks.

Object-based audio is the first step towards representing an audio scene that is agnostic to the configuration of the loudspeakers [8] [9]. Instead of focusing on the chain's reproduction end, object-based audio representation aims to demonstrate discrete sound sources in the auditory scene. These discrete sound sources, also called objects, are identified by their positions and the sound signals they emit. A renderer at the playback end of the chain is tasked with recreating the object at the specified position. Being at the playback location, the renderer can be made aware of the geometry of the loudspeaker rig and therefore can generate the appropriate signals for the speakers to create the spatial localization of the sound object. Object-based audio can overcome user interaction problems with the audio renderers using the metadata and channel-based recorded objects.

As the production pipelines and transmission infrastructures for object-based audio are already in place, the high-quality capture of audio objects is still an active field of research. More specifically, transcoding methods between scene and object-based

formats are needed for extracting audio objects and their metadata containing source direction of arrivals (DOA), gain, and semantic information from scene-based recordings. Transcoding between scene-based to object-based representations can be formulated as a source separation problem. Automatic generation of metadata that accompanies the audio objects is in part of a DOA estimation problem.

An automatic transcoding process from scene-based to object-based representation is a highly demanded solution for delivering real recorded audio scenes in object-based formats to the end-user experiences. This would help the potential target devices to create object-based audio rendering from the scene-based recordings. MPEG-H 3D Audio is a new standard for spatial audio coding developed by the ISO/IEC Moving Picture Experts Group (MPEG) [10]. MPEG-H 3D audio implementations should fulfill low latency, high quality, and localizable audio requirements, and the quality of the sound after decoding should scale up with the bitstream transferring rate. Although low computational cost is in the secondary requirement group, the decoding phase's operations should be computationally efficient for real-time operations. This standard also provisions universal representation of encoded 3D sound in channel-based, scene-based, and object-based formats. While the former two are part of the standard mainly for backward compatibility reasons, the latter format allowing unprecedented flexibility while rendering spatial sound is the main novelty. This format consists of many channels to be mixed for creating the sensation of the scene-based audio using the speakers. According to the bandwidth limitations for transmission, Higher Order Ambisonics (HOA) is the technique for storing, transmitting, and reproducing the sound field with the higher-order degree of spatial resolution. Representation of the recorded acoustic field in HOA also simplifies the scene representations. Production and reproduction phases of the object-based, scene-based, and channel-based audio formats with the MPEG-H referenced implementation are given in Fig. 1.1.

Spherical microphone arrays are useful tools for scene-based audio recordings due to their advantages for providing a multi-channel full azimuth and elevation coverage in the recording case for real-life conditions [11]. Besides, spherical harmonic decomposition (SHD) is an encoding method using the microphone arrays to represent the scene in spherical harmonic coefficients. Scene-based audio renderers are already available in the recent electronic device market to reconstruct the sound field

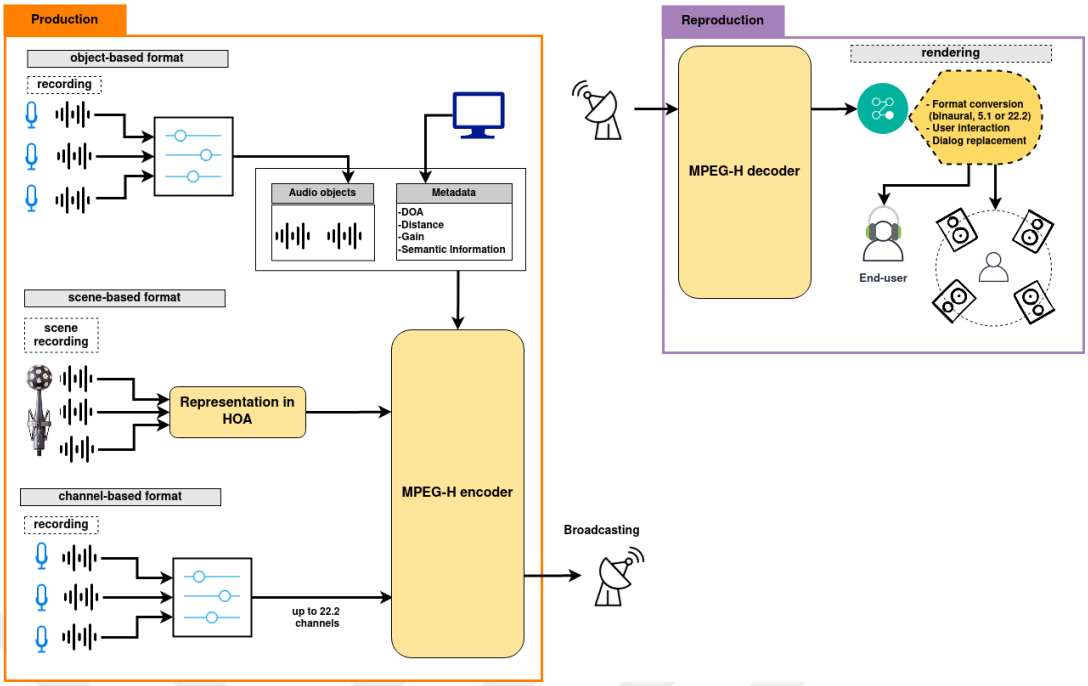


Figure 1.1: Production and reproduction phases of the 3D audio with the MPEG-H referenced implementation.

physically with HOA using the speakers. However, an efficient solution to extract audio objects from HOA is still needed. It is an important research field to make this transition via acoustic scene decomposition.

As the auditory perception of the human being is 3D in real life, estimating the DOA for multiple sound sources is a fundamental step for acoustic scene analysis. Many application areas, such as robot audition and object-based audio (OBA) broadcast, require computationally efficient DOA estimation to allow real-time operation. On the other hand, the scene would comprise several similar acoustic sources located in different positions with different sound pressure levels. It is not easy to separate the sources having similar frequency features from only their spectrograms. Separation of these sources requires grouping frequency components over the time-frequency domain according to their direction of arrivals. Therefore, understanding the scene with the DOA of the active sources on behalf of separating them can be the main objective for the auditory scene decomposition.

1.2 Proposed Methods and Models

According to the problem definition, this thesis's primary motivation is developing an acoustic scene analyzer to extract audio objects and their metadata using scene-based recordings. While this analyzer can be placed at the production phase in the implementation, it is also needed for the end-user devices to cover the cases where the object-based format is not available. The corresponding tool's position in the referenced MPEG-H implementation is given in Fig. 1.2. The pink box has a transcoding objective between scene-based to object-based formats.

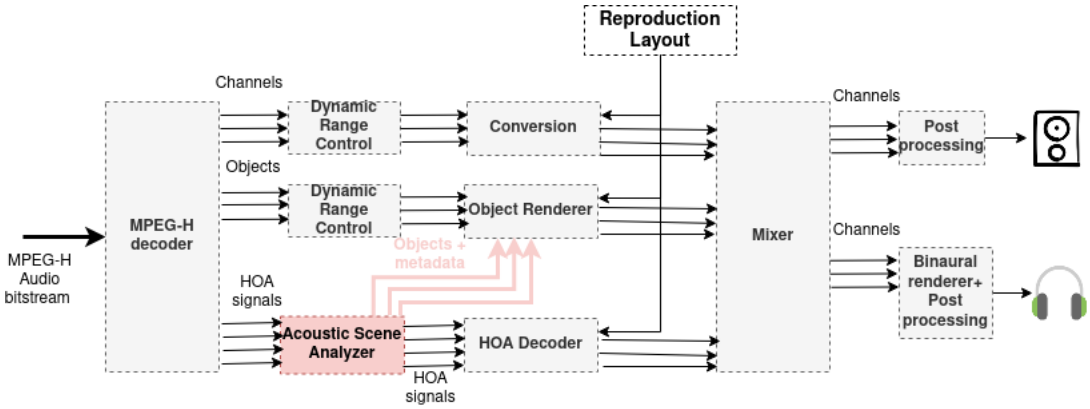


Figure 1.2: Acoustic scene analyzer designed to use after MPEG-H decoder.

The methods developed in this thesis are presented in two parts as multiple source DOA estimation and source separation. Two different DOA estimation methods are proposed to extract the spatial information of the acoustic scene. One of the methods, Hierarchical Grid Refinement (HiGRID), has matching accuracy and lower computational cost in comparison with the Steered Response Power (SRP) based methods using higher-order spherical harmonics. The other one, Residual Energy Test (RENT), operates at a lower computational cost and underlies the source separation techniques proposed in this thesis.

In this study, three novel source separation methods are also proposed to extract the original signals while the other sources' interference and reverberation are available. The first one is an Orthogonal Matching Pursuit (OMP) based acoustic source separation technique that uses a fixed spatial filter for source separation. The next one is an adaptive source separation technique, which models the spatial distribution of

the acoustic sources in each time interval, and uses an adaptive filter that separates each desired source. The last one is a post-processing Wiener filter calculated using the signal and noise covariance matrices extracted from the recordings. This post-processing enhancement technique can be used after beamforming and other source separation techniques.

1.3 Contributions and Novelties

The main contributions are as follows:

- An accurate DOA estimation technique was developed and published in the "IEEE/ACM Transactions on Audio, Speech, and Language Processing". The study's title is "Multiple Sound Source Localization With Steered Response Power Density and Hierarchical Grid Refinement". It provides a novel approach to decrease computational cost compared with the Steered Response Power (SRP) based approaches.
- A method was proposed for the detection of time-frequency bins that have single-source contributions. This method can also be used as a DOA estimation technique if clustering is applied as a post-processing stage on the DOA histogram. The study was presented at the "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019". The study title is "Multiple Sound Source Localization with Rigid Spherical Microphone Arrays via Residual Energy Test". Residual Energy Test that relies on a sparse plane wave decomposition is defined in this publication.
- An effective acoustic source separation technique is proposed to extract the original signals from the multichannel recordings over the reverberant environment. The study is presented in the "International Workshop on Acoustic Signal Enhancement (IWAENC), 2018" conference. The study title is "Acoustic Source Separation Using Rigid Spherical Microphone Arrays Via Spatially Weighted Orthogonal Matching Pursuit."
- Another contribution is the acoustic source separation via Residual Energy Test (RENT). This method extracts the sources' spatial distributions around the rigid

sphere via fitting the data into the Kent distribution. It is used as an adaptive spatial filter instead of using a fixed spatial filter.

- A novel approach is proposed to design an adaptive Wiener post-filter via RENT. This filter is designed as a weighting mask with the computation of active source and noise power spectral density (PSD). It is available as a post-processing stage for the available source separation techniques via weighting the acoustic sources in each time-frequency bin.

1.4 The Outline of the Thesis

The thesis is divided into six chapters. After this introduction and problem formulation, Chapter 2 surveys the primary background of acoustics, spherical harmonics, and rigid spherical microphone arrays (RSMA). This part also comprises the literature about acoustic source DOA estimation and separation. In Chapter 3, the novel methods are introduced and expressed in problem formulations. The main goal here is to introduce the theoretical framework for DOA estimation and acoustic source separation in highly reverberant environments. In Chapter 4, evaluation of the corresponding methods will be presented. The evaluation setups contain controlled real emulation scenarios and as well as real recordings. Statistical analysis methods are also applied to the randomly selected scenarios to follow the covariance effects over the separation algorithms. This chapter also introduces the evaluation of the proposed methods to state of the art, and then proposed approaches are subjectively evaluated in a multiple stimulus, hidden reference, and anchor (MUSHRA) test. The computational cost between the methods is also investigated. We discuss the methods developed during this work with their key findings and limitations for different cases in Chapter 5. Finally, in Chapter 6, we summarize the thesis's significant findings and propose issues inviting future research.

CHAPTER 2

BACKGROUND

2.1 Sound Wave, Pressure, Intensity relations

The speed of sound depends on the type and temperature of the medium in which propagation takes place. This medium is the air in the room acoustics. The speed of sound changes according to the temperature of the surrounding air medium and is given in (2.1), where T_a is the ambient temperature [12].

$$c = (331.4 + 0.6T_a) \tag{2.1}$$

Sound pressure is the measurement of the local pressure deviation from the ambient atmospheric pressure. It changes within time, speed, and the medium's viscosity. It is a scalar quantity and can be measured using a microphone at any position inside the air medium.

$$\rho_0 c \frac{\partial \nu(x,\omega)}{\partial t} = -\nabla p(x,\omega) \tag{2.2}$$

The law of conservation of momentum leads to the relation between pressure (p) and particle velocity (ν) as shown in the equation (2.2), where the radial frequency component is represented with ω and ambient density as ρ_0 [12].

While sound pressure is a scalar component without any information about the particles' oscillatory directions, the acoustic intensity is used as a vectorial quantity that denotes the acoustical energy's flow direction. It is mainly used for the identification of radiation characteristics of sound sources. At a given time instant and position, $I_s(x,t)$ as the instantaneous acoustic intensity can be calculated with the multiplication of the pressure and particle velocity.

$$I_s(x,t) = p(x,t)\nu^*(x,t) \tag{2.3}$$

Monochromatic plane wave refers to a plane wave consisting of a single frequency and propagates in a direction across the free medium. The summation of the multiplication between pressure and the complex conjugate of particle velocity over the period $T = 1/f$ denotes complex intensity for monochromatic sound field [13].

$$I_s(x) = \frac{1}{2T} \int_{t_0}^{t_0+T} p(x,t)\nu^*(x,t)dt \tag{2.4}$$

Particle velocity measurement is not as straightforward as measuring sound pressure and requires the usage of special probes. There are two different types of such probes [14] [15]: (1) p-p arrays which use several pressure-sensitive elements (i.e., omnidirectional microphones) to approximate the particle velocity component with the help of microphone positions and (2) p-u arrays that directly measure the axial components of particle velocity and the pressure values at the positions.

2.2 Acoustical Field Descriptions

Acoustical fields can be represented as a combination of simpler building blocks such as plane waves. The fundamental acoustics concepts used in this thesis are explained in this section.

2.2.1 Plane waves

In acoustical free field, pressure $p(\omega, \mathbf{x}, t)$ is a function of both space (position) ($\mathbf{x}$), frequency (w) and time (t). Sound pressure satisfies the homogeneous acoustic wave equation and it is expressed as [16] :

$$\nabla^2_{\mathbf{x}} p(\omega, \mathbf{x}, t) - \frac{1}{c^2} \frac{\partial^2 p(\omega, \mathbf{x}, t)}{\partial^2 t} = 0 \tag{2.5}$$

For a fixed frequency, the solution of the differential equation is:

$$p(\omega, \mathbf{x}, t) = p(\omega, \mathbf{x}, t)e^{i\omega t}, \tag{2.6}$$

where ω is the radial frequency defined as $\omega = 2\pi f$.

By expressing the wave equation $p(\omega, \mathbf{x}, t)$ with the notation of $p(k, \mathbf{x}, t)$ where k is the wavenumber in radians per meter defined as $k = \frac{\omega}{c}$, this expression is called the

Helmholtz equation [17] and is given as:

$$\nabla^2_{\mathbf{x}} p(k, \mathbf{x}, t) + k^2 p(k, \mathbf{x}, t) = 0 \tag{2.7}$$

The plane wave's sound pressure definition given for a monochromatic plane wave which satisfies the equation (2.7) is given.

$$p(\mathbf{k}, \mathbf{x}, t) = A e^{i\mathbf{k}^\mathbf{T}\mathbf{x}} e^{i\omega t} \tag{2.8}$$

In the equation (2.8), $\mathbf{k}$ is the wave vector defined in Cartesian coordinates as $\mathbf{k} = k_x \mathbf{u}_x + k_y \mathbf{u}_y + k_z \mathbf{u}_z$ gives the direction of the plane wave propagation in the free medium and $\mathbf{x}$ is the position vector. By expressing the unit amplitude acoustic wave equation in the spherical coordinates, the resulting plane wave will be presented in the equation (2.9) [18].

$$p(k, r, t) = 4\pi i^n j_n(kr) Y_n^m(\theta, \phi) e^{i\omega t}, \tag{2.9}$$

where $j_n(kr)$ is the spherical Bessel function of the first kind [19] and $Y_n^m(\theta, \phi)$ is the spherical harmonics with order n and degree m [20].

2.2.2 Spherical harmonics

Spherical harmonics are a group of functions used to solve partial differential equations in different disciplines. While a complete set of orthonormal basis can be formed with the spherical harmonics, functions defined on a sphere can be represented as a sum of spherical harmonics [20]. The spherical harmonic function of order n and degree m is given in the following equation.

$$Y_n^m(\theta, \phi) = \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \tag{2.10}$$

where $0 \leq \theta < \pi$ is the inclination angle defined from the positive z-axis, $0 \leq \phi < 2\pi$ is the azimuth angle defined from the positive x-axis, and $P_n^m(\cos\theta)$ is the associated Legendre function of order n and degree m [21], where $n \in \mathbb{Z}^+$ and, $m \in [-n, n]$. Representation of spherical harmonic functions calculated in 3D view for different m and n are represented in Fig. 2.1.

Properties of the spherical harmonic function: Some of the important properties of the spherical harmonic function are given below [20]:

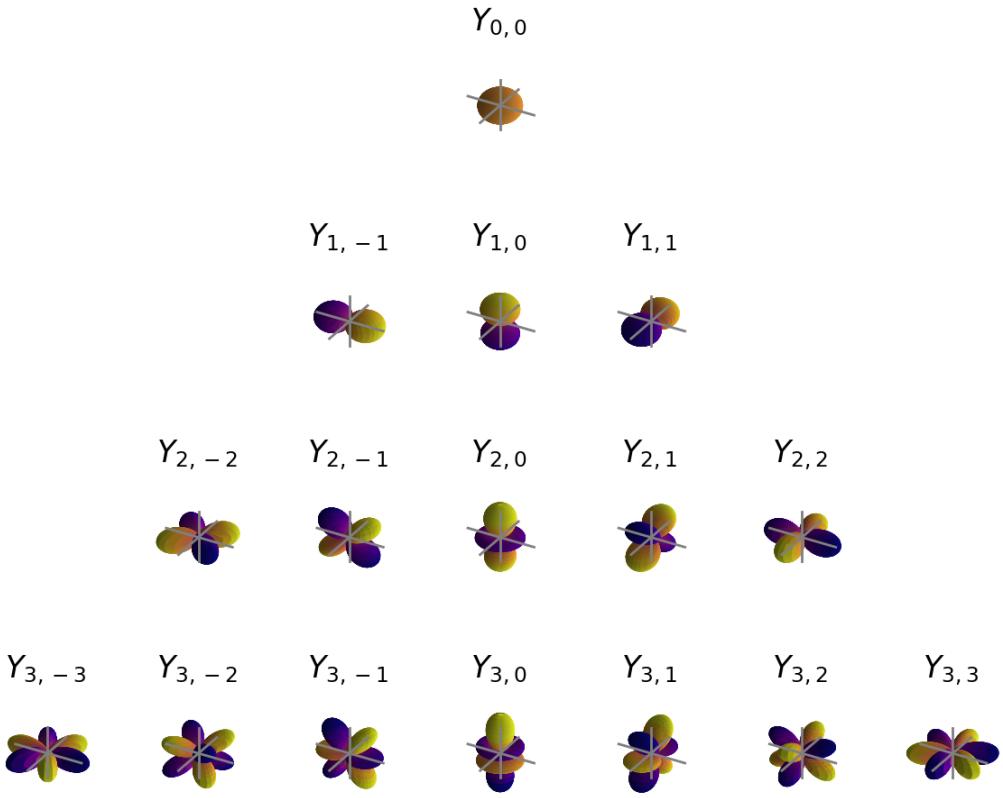


Figure 2.1: Spherical harmonic function Y_{nm} for different n and m.

- Spherical harmonic function is a complex-valued function which contains the term $e^{im\phi}$. The original function's complex conjugate can be presented as :

$$[Y_n^m(\theta,\phi)]^* = (-1)^m[Y_n^{-m}(\theta,\phi)] \tag{2.11}$$

- Degree, $m \in \mathbb{Z}$ of a spherical harmonic function is limited by its order n.

$$m \in [-n, n] \tag{2.12}$$

- Completeness of spherical harmonics:

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}[Y_n^m(\theta,\phi)]^* Y_n^m(\theta',\phi') = \delta(\cos\theta - \cos\theta')\delta(\phi-\phi'), \tag{2.13}$$

where $\delta(.)$ is the delta function which is equal to 1 (impulse), when its argument is equal to 0.

- Spherical harmonics addition theorem:

$$\sum_{m=-n}^{n} [Y_n^m(\theta,\phi)]^* Y_n^m(\theta',\phi') = \frac{2n+1}{4\pi} P_n(\cos\Theta), \tag{2.14}$$

where $\cos\Theta = \cos\theta\cos\theta' + \cos(\phi-\phi')\sin\theta\sin\theta'$ is the angle between the vectors having angles of (θ,ϕ) and (θ',ϕ') , $P_n(\cos\Theta)$ is the Legendre polynomial.

2.2.3 Spherical harmonic decomposition of a monochromatic plane wave

Similar to complex exponentials forming a basis for expanding arbitrary smooth functions in Fourier analysis, a linear combination of plane waves can be used to expand arbitrary pressure fields into a finite number of components. A practical method to achieve this decomposition is via spherical harmonics. Therefore, representing a single plane wave in the spherical harmonics domain is an essential starting point for analyzing the acoustic scene.

Sound pressure measurement according to the origin for a monochromatic plane wave with unit amplitude and propagating from the direction of (θ_k,ϕ_k) is expressed as [16]:

$$p(k,r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{m=n} 4\pi i^n j_n(kr)[Y_n^m(\theta_k,\phi_k)]^* Y_n^m(\theta,\phi), \tag{2.15}$$

where θ,ϕ is the steering direction. According to spherical harmonics' completeness property, a unit amplitude is obtained where the steering direction over the sphere and propagating direction of the plane wave are equal. By using (2.14) in the derivation of (2.15), the final expression contains Legendre polynomials that are dependent on the angle difference between the steering and the plane wave propagating directions.

$$p(k,r,\Theta) = \sum_{n=0}^{\infty} i^n j_n(kr)(2n+1)P_n(\cos\Theta), \tag{2.16}$$

where Θ is the angle between (θ_k,ϕ_k), (θ,ϕ) and represented as $\cos\Theta = \cos\theta_k\cos\theta + \cos(\phi_k-\phi)\sin\theta_k\sin\theta$.

As a general form obtained within the limitation of the spherical harmonics order N,

equation (2.16) is simplified with the following expression.

$$\sum_{n=0}^{N}(2n+1)P_n(\cos\Theta) = \frac{2N+1}{4\pi}\frac{P_{N+1}(cos(\Theta)) - P_N(cos(\Theta))}{P_1(cos(\Theta)) - P_0(cos(\Theta))} \tag{2.17}$$

The resulting plane wave pressure distribution with the dependence over the order limitation N and Θ is revealed in Fig. 2.2. We will define this distribution as a Legendre kernel.

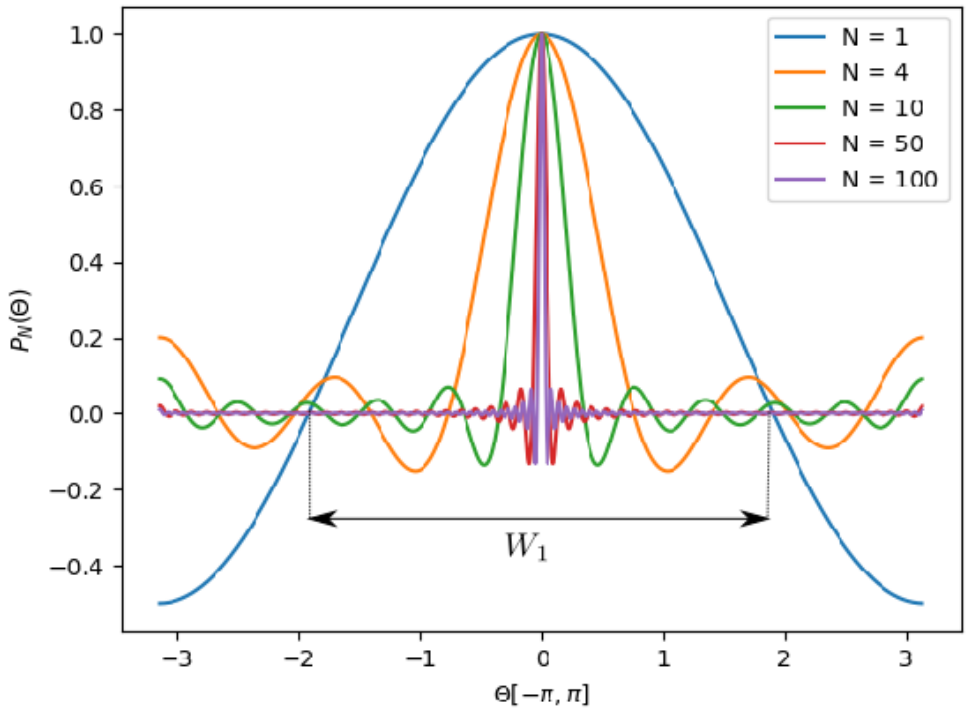


Figure 2.2: Legendre kernel distributions for different N. W_1 is the zero crossing beam width of order $N = 1$.

2.2.4 Composition of an arbitrary sound field with monochromatic plane waves over rigid or open surfaces

As spherical harmonics represent the sound field around the sphere according to directional features, the sound field can be composed of plane waves having amplitudes of $a(k, \theta_k, \phi_k)$. The sound pressure on the sphere from the direction (θ, ϕ) can be

calculated using (2.18) [16].

$$p(k, r, \theta, \phi) = \int_0^{2\pi} \int_0^{\pi} a(k, \theta_k, \phi_k) e^{i\mathbf{k}^\mathbf{T}\mathbf{r}} \sin\theta d\theta d\phi, \tag{2.18}$$

where (r, θ, ϕ) is the positional vector expressed in the spherical coordinate system.

As a general form with spherical harmonic functions, the sound field can be represented as a linear combination of spherical harmonic basis vectors. The resulting formula is the representation of plane wave composition for the spherical harmonics domain.

$$p(k, r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{nm}(k) Y_n^m(\theta, \phi) \tag{2.19}$$

For plane waves having unit amplitudes as in equation (2.15), the spherical harmonic coefficient $a_{nm}(k)$ can be extracted as within the following equation.

$$a_{nm}(k) = 4\pi i^n j_n(kr)[Y_n^m(\theta_k, \phi_k)]^*, \tag{2.20}$$

where $j_n(kr)$ is the first order spherical Bessel function by assuming the medium is open where there is no scattering surface, and r is the distance from the origin. When a scatter such as a rigid sphere is present in the scene, both incident and scattered fields will be present. This case will be discussed in the section about rigid spherical microphone arrays.

2.3 Spherical Microphone Array Design Parameters

This section outlines some of the basic concepts related to microphone arrays having a spherical structure. Theoretical analysis of sound fields typically considers a continuous spherical aperture in continuous measurements. However, this condition is not feasible for real-life conditions where data bandwidth and physical limitations require sampling the sound field at a finite number of points.

Positioning a limited number of microphones over the sphere would result in a minimum spatial resolution requirement for splitting the plane waves. The equation (2.17) in a limited order (N) defines a plane wave representation over the sphere. As the plane wave distribution according to the angle Θ is given in Fig. 2.2, main lobe width Θ_N (the widest beam in the positive axis) is extracted in the Fig. 2.3 for different

spherical harmonics order. Rayleigh condition, which is the minimum angle difference constraint for multiple sources to display them separately, is estimated using the formula $\Theta_N \approx \frac{\pi}{N}$ [22]. The estimated function and zero crossings width $(W/2)$ of the plane wave according to the Fig. 2.2 are displayed in Fig. 2.3 for different truncation errors.

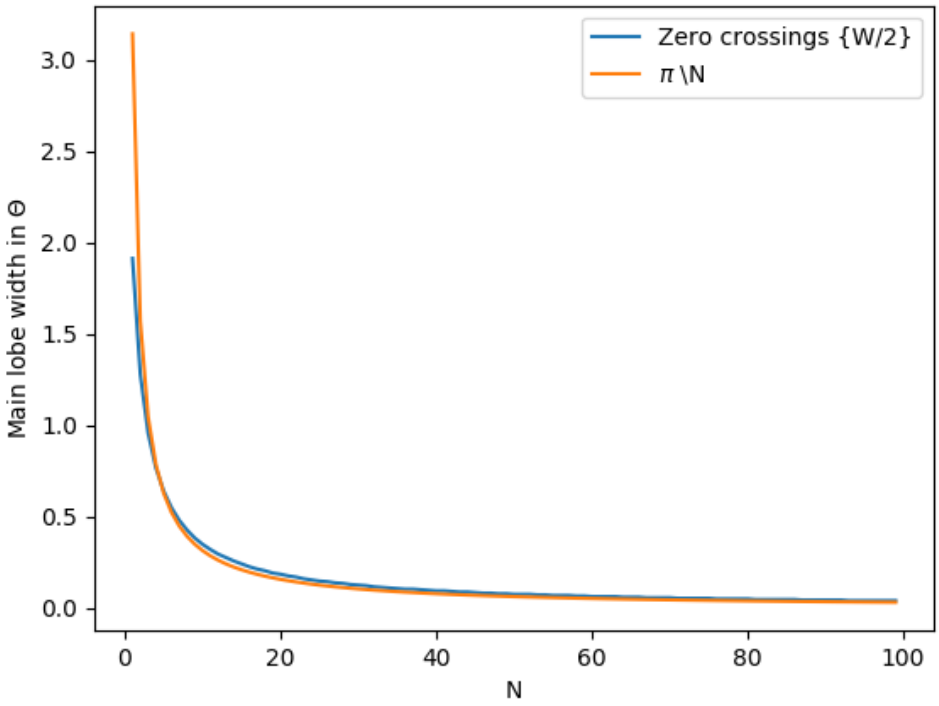


Figure 2.3: Main lobe width Θ_N according to the Fig. 2.2 in radians with respect to the order N

Spherical Harmonic Decomposition (SHD) over the spherical microphone arrays : Although the knowledge of the pressure distribution at infinite small slices over a sphere provides accurate observations for the sound field analysis, this is impossible due to practical limitations that no such continuous surface sensor exists. This problem makes it necessary to sample the spherical surface at a finite number of points by allowing a band-limited reconstruction of the sound field. Sampling is directly related to the numerical integration of the pressure distribution over the sphere [22].

Estimating the pressure distribution on and around a sphere depends on the sampling configurations such as the microphones' angles and their represented areas as quadrature weights on the sphere. Several sampling methods exist on the sphere such as equal-angle, Gaussian, Lebedev, uniform, and nearly uniform sampling schemes [16]. The spatial information accuracy in azimuth and elevation can change with respect to the positioning of microphones.

As the knowledge of microphone positions on the sphere is crucial for spatial information acquisition, open spherical microphone arrays such as p-p arrays use microphone sampling positions to calculate the particle velocity at the center [15]. It is then used to calculate the acoustical intensity estimations for the direction of arrival (DOA). Another method used for open or rigid spherical microphone arrays to represent the sound field around the sphere as a combination of multipoles is the spherical harmonic decomposition (SHD), also known as Spherical Fourier Transform. The relation between the maximum feasible order (N) of the SHD and the number of sampling points (Q) on the sphere is defined with the following definition [16].

$$Q \geq (N+1)^2 \tag{2.21}$$

For example, if a maximum order of spherical harmonics is selected to be 4 that needs to be decomposed from the sampling microphones over the sphere, the number of discrete microphones should be at least 25.

Quadrature methods are used to approximate the integral of a function over a given domain by sampling it with a finite number of sampling points. In general, spherical quadrature is defined in the equation (2.22) [11].

$$\int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) \sin(\theta) d\theta d\phi \approx \sum_{q=1}^{Q} w_q f(\theta_q, \phi_q), \tag{2.22}$$

where w_q is the quadrature weight, Q is the number of sampling points, θ_q and ϕ_q are the elevation and azimuth angles of the sampling positions in spherical coordinates according to the microphone (q).

Spherical Fourier Transform (SFT) is used to express the function defined over a sphere using periodic spherical harmonic functions on the sphere. A sound field can be described using the spherical harmonic coefficients. If the continuous func-

tional form of the pressure distribution is known, SHD can be calculated using equation (2.23) [11].

$$\mathbf{a_{nm}} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi)[Y_n^m(\theta, \phi)]^* \sin(\theta) d\theta d\phi \tag{2.23}$$

By discretizing the continuous pressure distribution over the sphere, SHD is calculated using an appropriate quadrature scheme with a finite number of sampling points. It is expressed with the following formula [11]:

$$\mathbf{a_{nm}} = \sum_{q=1}^{Q} f(\theta_q, \phi_q)[Y_n^m(\theta_q, \phi_q)]^* w_q, \tag{2.24}$$

where $f(\theta_q, \phi_q)$ is the pressure measured by the microphone at the sampling position, w_q is the quadrature weight related to the sampling scheme, and (θ_q, ϕ_q) are the sampling directions in spherical coordinates.

For SHD over spherical microphone arrays, three parameters are used in the computations: (1) the pressure measured by each microphone, (2) the microphones' locations over the sphere, and (3) the quadrature weights related to the area represented by the sampling microphone. As examples, mostly used four types of spherical quadratures: equal-angle, Gaussian, uniform, and the truncated icosahedron (Eigenmike em32 [23]), are presented in detail.

1. Equal-Angle Sampling: [16] Over the sphere, azimuth and elevation specify the direction of a point. Azimuth is represented as $\phi \in [0, 2\pi)$ and elevation is represented as $\theta \in [0, \pi)$. Equal-angle sampling results in splitting the angles into $4(A+1)^2$ slices for azimuth and elevation. The resulting angle differences between each sampling position will be equal to each other. Sampling position angles are calculated with the following expressions.

$$\theta_q = (q + \frac{1}{2})\frac{\pi}{2A+2}, \quad q = 0, ..., 2A+1 \tag{2.25}$$

$$\phi_q = q\frac{2\pi}{2A+2}, \quad q = 0, ..., 2A+1 \tag{2.26}$$

The quadrature weights for each microphone position are calculated as in the

following formula.

$$w_q = \frac{2\pi}{(A+1)^2}\sin(\theta_q)\sum_{q'=0}^{A}\frac{1}{2q'+1}\sin([2q'+1]\theta_q), \quad q = 0, ..., 2A+1, \tag{2.27}$$

which allows a maximum order $N \leq 2A+1$ in SHD.

2. Gaussian Sampling [16]: The Gaussian sampling is the generation of a sampling scheme with equal angle slices in azimuth and elevation separately. In this scheme, $2(A+1)$ points are selected in azimuth, and $A+1$ points are selected in elevation. The angles are defined with the following expressions.

$$\theta_q = (\frac{\pi}{A+1}q), \quad q = 0, ..., A \tag{2.28}$$

$$\phi_q = (\frac{\pi}{A+1}q), \quad q = 0, ..., 2A+1 \tag{2.29}$$

Then, the quadrature weights for each microphone position are calculated.

$$w_q = \frac{\pi}{(A+1)}\frac{2(1-\cos^2\theta_q)}{(A+2)^2 P_{A+2}^2(\cos\theta_q)}, \quad q = 0, ..., A, \tag{2.30}$$

where $P_n(.)$ is the Legendre polynomial. Total number of sampling points will be $2(A+1)^2$. The maximum order for SHD is given as $N \leq \sqrt{2}(A+1) - 1$ using the equation (2.21).

3. Uniform Sampling [16]: The main criterion for this type of scheme is having equal quadrature weights for each sampling point. Quadrature weights are placed according to the number of sampling points (Q) using the formula in the finite approximation of the surface integral.

$$\int_0^{2\pi}\int_0^{\pi} f(\theta,\phi)\sin(\theta)d\theta d\phi \approx \frac{4\pi}{Q}\sum_{q=1}^{Q} f(\theta_q,\phi_q) \tag{2.31}$$

There are 5 different uniform sampling schemes. These are tetrahedron (4 vertices), cube (8 vertices), octahedron (6 vertices), dodecahedron (20 vertices) and icosahedron (12 vertices).

4. Truncated Icosahedron [24] : Eigenmike em32 is a popular spherical microphone array that is used for sound field analysis [24]. It uses a truncated icosahedral arrangement of microphones over a rigid sphere. Each vertex on

a truncated icosahedron has a neighborhood, with the other vertexes having 5 or 6 other vertices. Truncated Icosahedron has a total number of 32 sampling points. This type of scheme assumes that the quadratic weights are equal as in equal area sampling, despite small area differences.

2.4 Rigid Spherical Microphone Arrays

For a rigid spherical microphone array with Q microphone elements positioned at (θ_q, ϕ_q), the spherical harmonic components up to a maximum degree of $n = N$ can be calculated using numerical quadrature.

$$\mathbf{a_{nm}} = \sum_{q=1}^{Q} w_q p(\theta_q, \phi_q)[Y_n^m(\theta_q, \phi_q)]^*, \tag{2.32}$$

where w_q are used as quadrature weights associated with the sampling scheme and $p(\theta_q, \phi_q)$ are the pressure measurements acquired from the microphone q positioned at the location θ_q, ϕ_q.

A bandlimited approximation to the pressure around rigid sphere of radius r_a is presented with the equation (2.33).

$$p(k, \theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} a_{nm}(k) Y_n^m(\theta, \phi), \tag{2.33}$$

where $k = 2\pi f/c$ is the wavenumber and c is the speed of sound. The spherical harmonics due to a unit plane wave incident from the direction (θ_s, ϕ_s) and with a frequency f, can be written as a truncated series for $r \geq r_a$ [22].

$$a_{nm}(k) = 4\pi i^n b_n(kr_a)[Y_n^m(\theta_s, \phi_s)]^* \tag{2.34}$$

with

$$b_n(kr) = j_n(kr) - \frac{j_n^{'}(kr_a)}{h_n^{(2)'}(kr_a)} h_n^{(2)}(kr), \tag{2.35}$$

where $j_n(\cdot)$, $h_n^{(2)}(\cdot)$, and $h_n^{(2)'}(\cdot)$ are the spherical Bessel function of the first kind, spherical Hankel function of the second kind and its derivative with respect to its argument, respectively. This function denotes that the spherical harmonic expression can be decoupled into the acoustic field's directional and frequency components. The

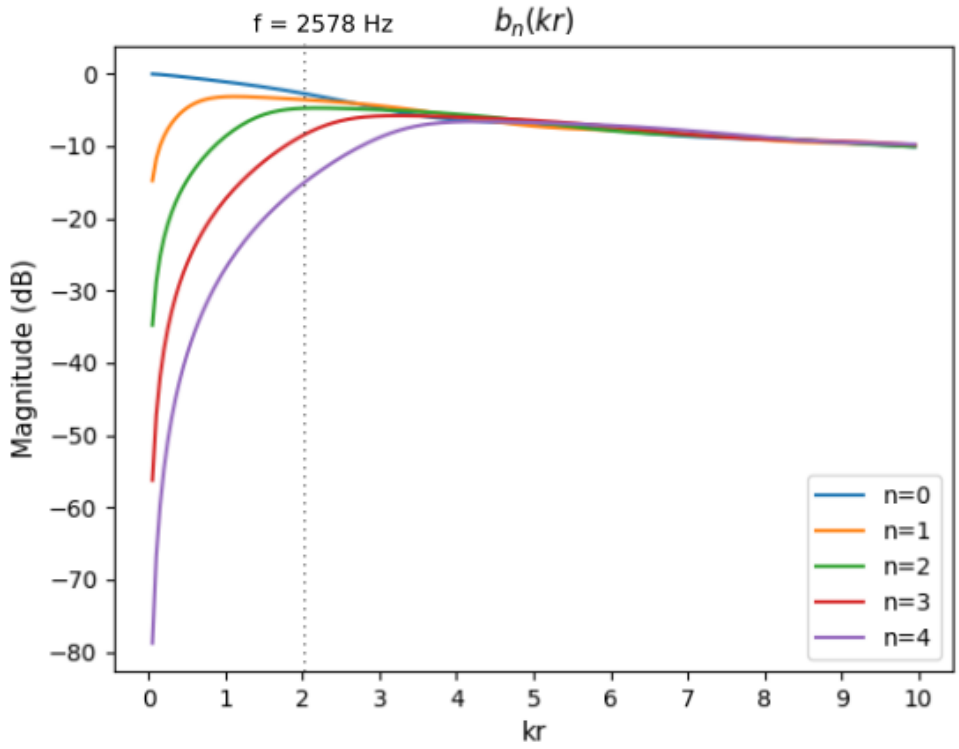


Figure 2.4: $b_n(kr)$ for different kr and n. $f = 2578$ Hz is calculated for $r = 4.2$ cm.

frequency response due to an incident plane wave on a scattered field, $b_n(kr)$ is named as mode strength and presented in Fig. 2.4 for different values of n. Notice that this response needs to be equalized to obtain only the sound field's direction-related components as decoupling from the frequency components.

Steered Response Functional: The maximum order of spherical harmonic coefficients is limited according to the number of sampling points for the RSMA. Therefore, only a spatially low-pass approximation can be obtained. It may be denoted by the spherical harmonic addition theorem [22]. The order limitation (N) results in

$$y_N(k,\theta,\phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{a_{nm}(k)}{4\pi i^n b_n(kr_a)} Y_n^m(\theta,\phi). \tag{2.36}$$

The resulting functional, $y_N(k,\theta,\phi)$ is called the *Steered Response Functional* (SRF). The steered response function's spatial resolution depends on the maximum order limit for the SHD coefficients. SRF can also be defined as maximum directivity beam

factor (MaxDF) steered in the direction (θ, ϕ). The selected beams have maximum directivity in the selected steering direction, and they are called the *regular beam pattern* and have resulted in the maximum suppression of the side lobes [22]. SRF as $y_N(k, \Omega)$ can be expressed in matrix formulation with $\Omega = (\theta, \phi)$. According to Fig. 2.4, it should be noted that using higher-order spherical harmonics in the SRF computations for low-frequency regions $(kr \ll 2)$ results in amplification of sensor noise. Therefore, the highest order N for the SRF should be adaptive with respect to the working frequency region.

- SRF can be calculated with the following matrix multiplication.

$$\mathbf{y_N} = \widetilde{\mathbf{Y}}\mathbf{B^{-1}}\mathbf{Y^H}\mathbf{W}\mathbf{p}, \tag{2.37}$$

  where $\widetilde{\mathbf{Y}}$ is a matrix of $M \times (N+1)^2$, in M steering directions.$\{\Omega_m\}_{m=1\cdots M}$ with rows given as:

$$\widetilde{\mathbf{Y}} = [Y_0^0(\Omega_m), Y_1^{-1}(\Omega_m), Y_1^0(\Omega_m), ..., Y_N^N(\Omega_m)] \tag{2.38}$$

- $\mathbf{B}$ is the acoustic reflecting and scattering component as $(N+1)^2 \times (N+1)^2$ matrix containing diagonal elements of $4\pi i^n b_n(kr)$.

- $\mathbf{p} = [p_1\ p_2\ ...\ p_Q]^T$ as the $Q \times 1$ vector containing the pressure for each microphone,

- $\mathbf{W} = diag\{w_q\}_{\{q=1...Q\}}$ as the diagonal matrix of quadrature weights with $Q \times Q$ vector,

- $\mathbf{Y^H}$ as the $(N+1)^2 \times Q$ matrix with the q-th column given by:

$$\mathbf{Y^H} = [Y_0^0(\Omega_q)^*, Y_1^{-1}(\Omega_q)^*, Y_1^0(\Omega_q)^*, ..., Y_N^N(\Omega_q)^*]^T \tag{2.39}$$

- $\mathbf{A} = [a_{00}(k),\ a_{1-1}(k),\ a_{10}(k),\ a_{11}(k),\ ...a_{NN}(k)\ ]^T$ containing the SHD coefficients in a $(N+1)^2 \times 1$ matrix,

$$\mathbf{A} = \mathbf{Y^H}\mathbf{W}\mathbf{p}. \tag{2.40}$$

- Resulting SRF ($\mathbf{y_N}$) vector is calculated in the designated steering directions. $\mathbf{y_N}$ is in the dimensions of $M \times 1$.

Steered Response Power (SRP) $P_N(k, \theta, \phi)$ is another property to measure the power in the specified steering directions. SRP can be denoted with the following equation.

$$P_N(k, \theta, \phi) = \left| \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{a_{nm}(k)}{4\pi i^n b_n(kr_a)} Y_n^m(\theta, \phi) \right|^2 \tag{2.41}$$

2.5 Acoustic Source DOA estimation methods

The direction of arrival estimation is an essential step for the sound field analysis. Signals captured by arrays containing multiple microphones are needed to develop DOA estimation techniques since a single pressure microphone can not carry the spatial information pertaining to the acoustic scene. In this subsection, DOA estimation methods are grouped into four categories as intensity, steered response power, subspace, and sparse recovery based methods. These works are related to the array signal processing literature.

2.5.1 Intensity based methods

The real part of the acoustic intensity vectors represents the direction of flow of the energy resulting from acoustic sources. It is possible to extract the propagation direction and magnitude of the plane waves using active intensity vectors.

Although RSMAs are not designed to measure acoustic intensity, several methods inspired by the energetic analysis of sound fields have been proposed. One of the methods using this approach is called Pseudo Intensity Vector (PIV). Zeroth and first-order spherical harmonics are used to represent approximations of intensity vectors [25] as given in the following expression:

$$I(k) = \frac{1}{2} \Re\{\tilde{a}_{00}(k)^* d(k)\}, \tag{2.42}$$

where $\tilde{a}_{00}(k) = \frac{a_{00}(k)}{b_0(k)}$ is the zeroth order spherical harmonics divided by the mode strength component and $\mathbf{d} = [D_x(k), D_y(k), D_z(k)]^T$ is a vector that mimics particle velocity in x, y and z directions. Particle velocity vectors in the directions of Ω_a are

given via first order spherical harmonics represented as in the equation (2.43).

$$D_a(k) = \sum_{m=-1}^{1} \frac{Y_{1m}(\Omega_a)}{b_1(k)} a_{1m}(k), \tag{2.43}$$

where Ω_a are steering directions of $\Omega_x = (\pi/2, \pi)$, $\Omega_y = (\pi/2, -\pi/2)$ and $\Omega_z = (\pi, 0)$ for the unit vectors in x, y and z planes.

The PIV is calculated for each time-frequency bin by noting the intensity vector directions in azimuth and elevation as local DOA estimates, and they are formed into a global DOA histogram. The global histogram is a spatially divided map into $(\Delta\theta, \Delta\phi)$ slices for azimuth and elevation. Searching for the bins containing S maxima over the global map is used for the DOA estimation for a known number, S, of sources [25]. However, it is not applicable in highly reverberant environments or when coherent sources are present. The number of local maxima that occur in the global map significantly increases due to reverberation. Labeling the number of S local maxima points does not give accurate DOA estimations if there is no elimination step as the time-frequency (TF) bin selection. TF bin selection methods are explained in the next part.

Subspace PIV is another intensity vector-based method for DOA estimation by extracting the components that represent the signal subspace [26]. Spherical harmonic coefficients representing the signal subspace is used in the PIV DOA estimation. This method firstly calculates the covariance matrix at the computational bin $(\tau, \varkappa)$ as :

$$\tilde{\mathbf{R}}_{\mathbf{a}}(\tau, \varkappa) = E\{\mathbf{a_{nm}}(\tau, \varkappa)\mathbf{a_{nm}}(\tau, \varkappa)^H\}, \tag{2.44}$$

where $\mathbf{a_{nm}}$ is a vector in $(N+1)^2 \times 1$ dimensions. Covariance matrix is divided into signal and noise subspaces using singular value decomposition (SVD) as in equation (2.45).

$$\tilde{\mathbf{R}}_{\mathbf{a}} = \mathbf{U_s}\mathbf{\Sigma_s}\mathbf{U_s^H} + \mathbf{U_n}\mathbf{\Sigma_n}\mathbf{U_n^H} \tag{2.45}$$

The actual spherical harmonics are replaced with the lower-order spherical harmonics in the signal subspace ($\mathbf{U_s} = [\tilde{a}_{00}, \tilde{a}_{1-1}, \tilde{a}_{11}, \tilde{a}_{01}, ..., \tilde{a}_{nm}]^T$) for the PIV computations.

Spatial resolution over the sphere depends on the maximum order of spherical harmonics. Since PIV uses zeroth and first-order, more accurate results can be obtained using higher-order spherical harmonic coefficients. Augmented intensity vec-

tors (AIV) method is used to do a grid search using higher order of spherical harmonics over the region defined by the PIV vector's DOA estimates [27] [28]. The noise-free signal equations are revealed and used for minimizing the cost function. The cost function for the AIV is defined within the following equation.

$$\tilde{C}(\tau, \varkappa) = \sum_{n=0}^{N} \sum_{m=-n}^{n} |a_{nm}(\tau, \varkappa) - \sqrt{4\pi} a_{00}(\tau, \varkappa) Y_{nm}^{*}(\Omega)|^2, \tag{2.46}$$

where $\sqrt{4\pi} a_{00}(\tau, \varkappa) Y_{nm}^{*}(\Omega)$ is used for presenting the pressure which does not depend on the plane wave propagating direction. Minimizing the cost function in a search window centered at the DOA estimate (Ω_{PIV}) and within the direction interval $\Delta\Omega_{PIV} = (\Delta\theta_{PIV}, \Delta\phi_{PIV})$ would result in the adjusted DOA estimate.

$$\Omega_{AIV} = \arg\min_{\Omega} \{\tilde{C}(\tau, \varkappa)\}, \Omega \in \Omega_{PIV} \tag{2.47}$$

These estimates are then formed in a global histogram, and the maximum-valued indexes of this DOA histogram are identified as the source DOAs. AIV provides more accurate results when compared with PIV.

Time-frequency bin selection methods: A preliminary step is required to detect the time-frequency (TF) bins with single-source contributions [29] [30]. Most DOA estimation algorithms can localize multiple simultaneous sources when DOA estimates are based on TF bins that contain energy substantially from single sources. These estimates will be more reliable. Direct Path Dominance (DPD) Test [31], and Single Source Zones (SSZ) detection [32] are thresholding based methods used for pre-elimination.

DPD Test relies on the SVD of the spatial correlation matrix over a time-frequency region $(\tau \pm J_\tau, \varkappa \pm J_\varkappa)$. The sample spatial correlation matrix is given as:

$$\tilde{R}_a(\tau, \varkappa) = \frac{1}{J} \sum_{j_\tau=-J_\tau}^{j_\tau=J_\tau} \sum_{j_\varkappa=-J_\varkappa}^{j_\varkappa=J_\varkappa} a_{nm}(\tau + j_\tau, k + j_\varkappa) \times a_{nm}(\tau + j_\tau, \varkappa + j_\varkappa)^H. \tag{2.48}$$

SVD is the projection of the spatial correlation matrix to the one dimensional signal subspace. Singular values are obtained from the diagonal entities of the matrix (Σ).

$$\tilde{\mathbf{R}}_{\mathbf{a}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{\mathbf{H}} = [\mathbf{U_s}\mathbf{U_n}] \begin{bmatrix} \Sigma_s & 0 \\ 0 & \Sigma_n \end{bmatrix} [\mathbf{U_s}\mathbf{U_n}]^{\mathbf{H}} \tag{2.49}$$

The ratio between the two largest singular values ($\sum_{s1} / \sum_{s2}$) is compared with a pre-defined threshold (σ). If this ratio of the TF bin $(\tau, \varkappa)$ positioned at the center is greater than the selected threshold, this bin is labeled to have single source contribution. The condition results in an effective rank of 1 for the spatial correlation matrix in the specified TF bin.

DPD Test is used as a preliminary step for the PIV method in reverberant environments. PIV calculates the average sum for the intensity vectors over the identified region.

$$\tilde{I}(\tau_i, \varkappa_i) = \frac{1}{J} \sum_{j_\tau=-J_\tau}^{j_\tau=J_\tau} \sum_{j_\varkappa=-J_\varkappa}^{j_\varkappa=J_\varkappa} I(\tau_i + j_\tau, \varkappa_i + j_\varkappa), \tag{2.50}$$

where i is the TF bin index which passes the DPD Test and $J = (2J_r + 1)(2J_\varkappa + 1)$. This intensity vector is used to find the local DOA estimates referenced by the center TF bin $(\tau_i, \varkappa_i)$. This method is named DPD-PIV and provides DOA estimation more accurately compared with the method PIV without DPD.

The incremental study uses the center TF bin referenced by the DPD Test in the computation of the intensity vectors. That modification also reveals more accurate results than DPD-PIV [29]. This method is named DPD-SS-PIV, and it is used as a single intensity vector computation instead of averaging the intensity vectors over the adjacent TF bins.

SSZ detection is another technique to find TF bins that have single-source contributions [32]. SSZ is checked over K frequency-adjacent TF bins to find whether it satisfies the following criterion or not.

$$p(\tilde{\tau, K}) = \frac{1}{L} \sum_{i,j} \frac{R_{i,j}(\tau, K)}{\sqrt{R_{i,i}(\tau, K) R_{j,j}(\tau, K)}} \geq 1 - \epsilon, \tag{2.51}$$

where $R_{i,j}(\tau, K) = \sum_{\varkappa \in K} |X_i(\tau, \varkappa) X_j(\tau, \varkappa)|$ is the cross-correlation between the adjacent microphone signals (i, j) in time (τ) and frequency $(\varkappa)$, ϵ is a pre-defined threshold and L is the number of the possible combinations for the adjacent microphones. If this criterion is satisfied, the method assumes that only one source is dominant over K frequency-adjacent bins.

The elimination of TF bins in a pre-processing stage reduces the total cost of DOA estimation methods because it blocks a significant number of computations at each

TF bin. Pavlidi et al. use intensity vectors over the TF bins, which passes the SSZ criterion to form the DOA histogram [30]. Computation of the intensity vectors for each TF bin passing the SSZ test gives local findings. The DOA map can be extracted for the specified time interval by joining local DOA estimates into a global histogram. Then, the source count and DOA estimations for the time interval are extracted using a Gaussian function smoothing over the DOA map $y(\theta_i, \phi_j)$ such that:

$$y_s(\theta, \phi) = \sum_i \sum_j y(\theta_i, \phi_j) w_A(\theta - \theta_i, \phi - \phi_j), \tag{2.52}$$

where $w_A = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{\theta^2+\phi^2}{\sigma^2}}$. The location (θ, ϕ) having the maximum value over the function y_s is detected as the first DOA. Then, the contribution of the maximum-valued index is subtracted from the DOA histogram. It continues up to having a flat histogram, which does not have a maximum value. The source count is regarded as the number of iterations.

Pavlidi et al. also propose an improvement over the local DOA (θ, ϕ) estimations via beamforming in a number of directions $(2e + 1)^2$ over the adjacent spherical sector $(\theta \pm e\Delta\theta, \phi \pm e\Delta\phi)$ using higher-order of spherical harmonics [33]. The main objective is searching for an index with a greater beamforming magnitude than the output collected from the beamforming over the intensity vector's DOA. Before inserting the local DOA estimates to the global DOA histogram map, intensity vector estimates are updated with more precise results using beamforming in narrower slice directions.

2.5.2 Steered Response Power (SRP) based methods

Beamforming using RSMAs involves summing the linear combination of eigenbeams (i.e., spherical harmonic components) and is given as :

$$y(\tau, \varkappa, \Omega_a) = \sum_{n=0}^{N} \frac{1}{b_n(kr)} \sum_{m=-n}^{n} Y_{nm}(\Omega_a) \underbrace{\sum_{q=1}^{Q} w_q p_q(\tau, \varkappa) Y_{nm}^*(\Omega_q)}_{\text{SHD} : a_{nm}(\tau, \varkappa)}, \tag{2.53}$$

where $p_q(\tau, \varkappa)$ are the microphone signals after time-frequency decomposition, w_q is the quadrature weight associated to each microphone (q), Ω_q are the microphone positions, and Ω_a is the steering direction.

SRP is defined as the output power of a maximum directivity factor steered beamformer corresponding to different steering directions. SRP map obtained for the TF bin $(\tau, \varkappa)$ is given with the expression.

$$P(\tau, \varkappa, \Omega_a) = |y(\tau, \varkappa, \Omega_a)|^2, \Omega_a \in ([0, \pi), [0, 2\pi)) \tag{2.54}$$

DOA estimation via maximum-valued SRP finding involves steering the maximum directivity factor beams in all possible directions. This condition results in a high computational cost that is not feasible for many applications. The spherical arrays' physical design parameters, such as microphone positioning, directly affect the beamformer output's accuracy [34]. Spatial aliasing is another factor that degrades the DOA estimation accuracy, and it puts a constraint such as a minimum directional difference between the acoustic sources to detect their DOA [22].

The main idea behind SRP is beamforming in the pre-defined directions on the sphere and searching for the direction having the maximum power.

$$\Omega = \arg\max |y(\tau, k, \Omega_a)|^2 \tag{2.55}$$

This computation can be handled over all TF bins or TF bins selected from a preprocessing stage to decrease the computational cost. These selected TF bins are only used for the maximum directivity power search. Sec. 2.5.1 describes the TF bin selection methods as DPD and SSZ in detail.

For multiple coherent sources, the SRP functional will have multiple peaks corresponding to respective source DOAs. These can be accurately identified by using a dense search grid with the desired resolution. However, this brute-force approach is not usually suitable for practical applications due to its high computational cost. The spatial detail that SRP can resolve is approximately π/N also known as Rayleigh condition [22]. In other words, only sources having angle differences more than π/N between their DOAs can be discriminated as separated sources.

SRP can also be defined using the phase transform as SRP-PHAT [35]. As the DOA estimation of a single source is determined by searching the local or global peaks in the SRP, phase transformed SRP is defined as :

$$P_n(x) = \int_{nT}^{(n+1)T} \left| \sum_{l=1}^{M} w_l m_l(t - \tau(\mathbf{x}, \mathbf{x_l})) \right|^2 dt, \tag{2.56}$$

where w_l is the weight, m_l is the original microphone signal and $\tau(\mathbf{x}, \mathbf{x_l})$ is the time delay of the source from $\mathbf{x}$ according to microphone position $\mathbf{x_l}$. The cross-correlation of the microphones for all possible pairs are used to determine the SRP. The frequency-domain representation can be seen from the following equation (2.57). This function represents a formulation to steer along with the position $\mathbf{x}$. The objective is to find one or more maxima that identify the source DOAs over the SRP.

$$P_n(\mathbf{x}) = \sum_{k=1}^{M} \sum_{l=1}^{M} \int_{-\infty}^{\infty} W_k(w) W_l^*(w) M_k(w) M_l^*(w) e^{jw(\tau(\mathbf{x},\mathbf{x_k}) - \tau(\mathbf{x},\mathbf{x_l}))} dw \quad (2.57)$$

In order to avoid selecting the same microphones for the SRP-PHAT, the equation is corrected as:

$$P_n(\mathbf{x})' = \sum_{k=1}^{M} \sum_{l=k+1}^{M} \int_{-\infty}^{\infty} \Psi_{kl}(w) M_k(w) M_l^*(w) e^{jw(\tau(\mathbf{x},\mathbf{x_k}) - \tau(\mathbf{x},\mathbf{x_l}))} dw, \quad (2.58)$$

where $\Psi_{kl}(w) = \frac{1}{|M_k(w) M_l^*(w)|}$ is the inverse of the magnitudes of the frequency components for normalizing the SRP-PHAT. Searching for the maxima points over the full medium coordinates is not a computationally efficient way. Therefore, Course to Fine Region Contraction (CFRC) is the proposed technique to volumetrically localize sources in mediums such as a room or a concert hall by starting from larger grids to smaller volumes iteratively [35]. This method is preferred to decrease the higher computational cost, which occurs when steering smaller subvolumes for the whole region.

While the main lobe obtained through maximum directivity factor (MaxDF) beamforming does not result in sharp DOA estimations, Chu et al. propose a method via deconvolution of MaxDF with what they term as the CLEAN-SC filter to obtain narrower beamwidths and more accurate DOA estimations [36] [37]. This method also eliminates the diffuse fields and back lobes in the opposite directions occurring for the MaxDF map.

2.5.3 Subspace-based methods

Subspace-based methods deal with noise and signal subspaces in the cross-spectrum matrix. Classical spectrum estimation methods such as MUSIC and ESPRIT [38] [39],

have been adapted to use via spherical microphone arrays with a formulation within the spherical harmonic decomposition framework [3].

The received signal or the pressure on the microphone positioned over the spherical array is represented with a standard signal representation.

$$p_q(k) = \sum_{d=1}^{D} v_{qd}(k) s_d(k), \tag{2.59}$$

where $s_d(k)$ is the complex magnitude of the plane waves and $v_{qd}(k) = e^{-j\mathbf{k_d^T r_q}}$ is the phase component of a plane wave with a wave vector $\mathbf{k_d}$ and at a microphone position $\mathbf{r_q}$ over the sphere. The pressure signal vector is given in a compact model.

$$\mathbf{p_q} = \mathbf{v_q s}, \tag{2.60}$$

where $\mathbf{v_q} = [v_{q1}(k), v_{q2}(k), \cdots, v_{qD}(k)]$ and $\mathbf{s} = [s_1(k), s_2(k), \cdots, s_D(k)]^T$. In the condition of noise inserted to the spherical array signal model, the model for each microphone (q) becomes as :

$$\mathbf{p_q} = \mathbf{v_q s} + \eta_q \tag{2.61}$$

The array's cross spectrum matrix can be formulated with the following expression using the parameter $\mathbf{p}(k) = [p_1(k)...p_Q(k)]^T$, $\mathbf{V}(k) = [\mathbf{v_1}(k)...\mathbf{v_Q}(k)]^T$, $\eta(k) = [\eta_1(k)...\eta_Q(k)]^T$ and assuming that the signals and noise are uncorrelated.

$$E\{\mathbf{pp^H}\} = \mathbf{V\tilde{S}V^H} + \tilde{\mathbf{S}}_\eta, \tag{2.62}$$

where $\tilde{\mathbf{S}} = E\{\mathbf{ss^H}\}$ and $\tilde{\mathbf{S}}_\eta = E\{\eta\eta^\mathbf{H}\}$. The spatial correlation matrix of the microphones' pressure recordings over the array is in the form of $Q \times Q$, where Q is the number of microphones.

While the spatial correlation matrix is symmetric from left to right, SVD representation of cross-spectrum matrix in the signal and noise subspaces is given within the following expression.

$$E\{\mathbf{pp^H}\} = \mathbf{U\Sigma U^H} = [\mathbf{U_s U_n}] \begin{bmatrix} \mathbf{\Sigma_s} & 0 \\ 0 & \mathbf{\Sigma_n} \end{bmatrix} [\mathbf{U_s U_n}]^H, \tag{2.63}$$

where $\mathbf{U_s}$ is the signal subspace in the form of $Q \times S$ and $\mathbf{U_n}$ is the noise subspace in the form of $Q \times (Q - S)$, where S is the source count parameter. The number

of $(Q - S)$ minimum-valued singular values over the matrix $\mathbf{\Sigma}$ represents the noise. The unknown parameter input for the subspace-based methods is the source count to separate the signal and noise subspaces.

By using the noise subspace matrix $\mathbf{U_n}$, a MUSIC DOA map is obtained as steering in all possible plane wave directions [38].

$$P_{MUSIC} = \frac{1}{\mathbf{a}\mathbf{U_n}\mathbf{U_n^H}\mathbf{a}^*}, \tag{2.64}$$

where $\mathbf{a}$ is the steering vector from the possible plane wave directions to the microphone positioning direction in the dimensions of $M \times Q$ and M is the number of steering direction count. MUSIC provides a much sharper DOA spectrum for monochromatic sources when source count is a priori known.

Minimum Variance Distortionless Response (MVDR) is another technique to minimize the contribution of noise and other acoustic sources coming from any other direction than Ψ_m by setting a fixed gain on the direction of Ψ_m. MVDR spectrum is denoted within the following expression.

$$P_{MVDR} = \frac{1}{\mathbf{a}\zeta^{-\mathbf{1}}\mathbf{a}^*}, \tag{2.65}$$

where $\zeta = E\{\mathbf{p}\mathbf{p}^\mathbf{H}\}$ is the cross-spectrum matrix of the pressure measurements. The maximum entropy (ME) map given as another subspace method is calculated using the vector $\mathbf{c}$, which is the first column of $\zeta^{-\mathbf{1}}$ and in the form of $Q \times 1$ is seen as [40]:

$$P_{ME} = \frac{1}{\mathbf{a}\mathbf{c}\mathbf{c}^\mathbf{H}\mathbf{a}^*} \tag{2.66}$$

The next step is searching the S maximum values from the subspace-based maps and determining the DOA estimates using the indexes of the maxima from the possible plane wave propagating directions in azimuth and elevation.

Khaykin et al. express the spatial correlation matrix in the spherical harmonic domain by dividing with the mode strength [3]. In the spherical harmonics representation, the signal model is defined as in the following equation.

$$\mathbf{a_{nm}}(k) = \mathbf{Y}^\mathbf{H}(\Psi)\mathbf{s}(k) + \mathbf{z_{nm}}(k), \tag{2.67}$$

where $\mathbf{z_{nm}}(k) = \mathbf{B^{-1}}(kr)\mathbf{Y^H}(\Psi)\eta(k)$ is the parameter related to the array's mode strength and noise. $\Psi = (\theta, \phi)$ is the steering direction in azimuth and elevation. The spatial correlation matrix is calculated as :

$$E\{\mathbf{a_{nm}}\mathbf{a_{nm}^H}\} = \mathbf{Y^H}\tilde{\mathbf{S}}\mathbf{Y} + \tilde{\mathbf{Z}}, \tag{2.68}$$

where $\tilde{\mathbf{S}} = E\{\mathbf{ss^H}\}$ and $\tilde{\mathbf{Z}} = E\{\mathbf{z_{nm}}\mathbf{z_{nm}^H}\} = \mathbf{B^{-1}}(\mathbf{B^{-1}})^{\mathbf{H}}\sigma^2$. The spatial correlation matrix is decoupled from the frequency via multiplying both sides with the mode strength $\mathbf{B^{-1}}$. The average of the spatial correlation matrix over K frequency adjacent bins are used in the calculations.

$$\tilde{\mathbf{S}}_{\mathbf{SHD}}(k) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{S_{SHD}}(k), \tag{2.69}$$

where $\mathbf{S_{SHD}} = E\{\mathbf{a_{nm}}\mathbf{a_{nm}^H}\}$ is the spatial correlation of the spherical harmonics in the form of $(N+1)^2 \times (N+1)^2$ and N is the maximum order limitation. Decoupled spherical harmonic coefficients are used to calculate the cross-spectrum matrix. After obtaining MVDR spectrum, maximum values for the possible plane wave propagating directions (Ψ_0) over the spectrum define the estimated DOAs.

$$P_{MVDR} = \frac{1}{\mathbf{Y}(\mathbf{\Psi_0})\tilde{\mathbf{S}}_{\mathbf{SHD}}^{\mathbf{-1}}\mathbf{Y}(\mathbf{\Psi_0})^{\mathbf{H}}} \tag{2.70}$$

MUSIC spectrum using SVD over the cross-spectrum matrix $\mathbf{S_{SHD}}$ after noise subspace detection $\tilde{\mathbf{Z}} = E\{\mathbf{z_{nm}}\mathbf{z_{nm}^H}\}$ is observed to search over the possible plane wave directions (Ψ_0). MUSIC spectrum is defined as :

$$P_{MUSIC} = \frac{1}{\mathbf{Y}(\mathbf{\Psi_0})\tilde{\mathbf{Z}}\mathbf{Y}(\mathbf{\Psi_0})^{\mathbf{H}}}. \tag{2.71}$$

Although subspace-based methods output accurate DOA estimations seen in Fig. 2.5, their performance suffers where coherent sources occur. For such conditions, they require source count as an input to select the signal and noise subspaces before seeking maxima in the resulting DOA spectrum. Coherent sources occurring for the same spectrum can be split according to the source count parameter. If the number of sources occurring in the same TF bin is known, the source counts D can be used to decide for partitioning the cross-correlation matrix into signal and noise subspaces [41].

Kumar et al. propose another method using near field MVDR beamforming instead of far-field MVDR as an improvement to subspace-based methods in near field source

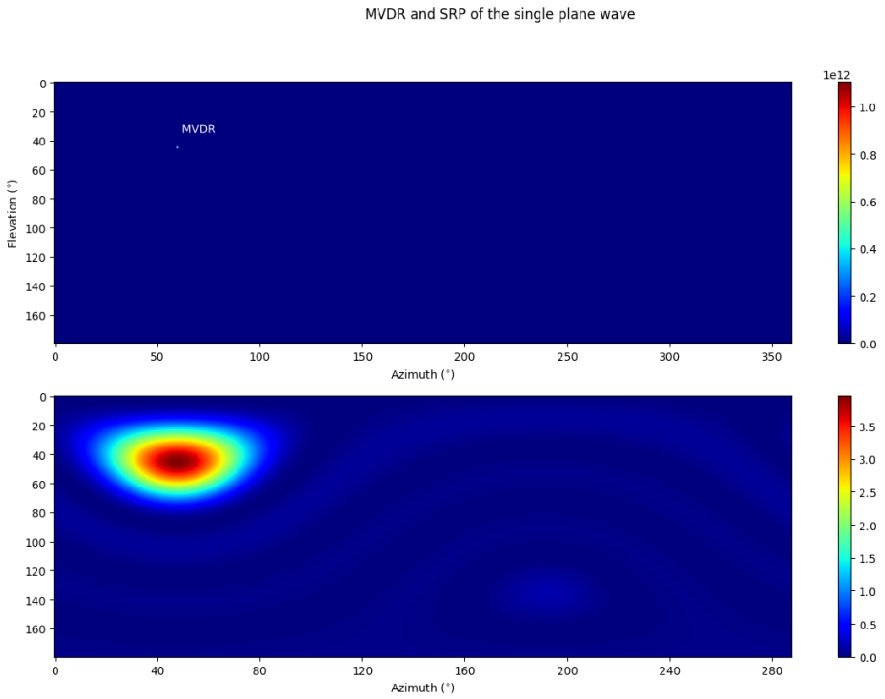


Figure 2.5: Steering in 1° azimuth and elevation for MVDR [3] and SRP of a plane wave propagating in directions of $(45^\circ, 60^\circ)$ in elevation and azimuth.

localization [42]. Steering MVDR weights are given as :

$$\mathbf{w_{nm}} = \frac{\tilde{\mathbf{S}}_{\mathbf{SHD}}^{-\mathbf{1}}\mathbf{B_{r_s}}\mathbf{Y^H}(\mathbf{\Psi_s})}{\mathbf{Y}(\mathbf{\Psi_s})\mathbf{B}_{\mathbf{r_s}}^{\mathbf{H}}\tilde{\mathbf{S}}_{\mathbf{SHD}}^{-\mathbf{1}}\mathbf{B_{r_s}}\mathbf{Y^H}(\mathbf{\Psi_s})}, \tag{2.72}$$

where $\mathbf{B_{r_s}}$ is the mode strength calculated for the radius r_s and Ψ_s is the angle of microphones. MVDR spectrum steering in the directions of Ψ_l is given in the following expression.

$$WG = |\mathbf{w}_{\mathbf{nm}}^{\mathbf{H}}\mathbf{B_{r_l}}\mathbf{Y^H}(\mathbf{\Psi_l})| \tag{2.73}$$

MVDR method is categorized as a benchmark to control the DOA estimation accuracy of the proposed methods in the non-reverberant environments. However, the performance of MVDR beamforming relies on accurate spatial covariance matrix estimation. Accurate DOA estimation can not be easily obtainable for the cases having coherent sources or strong reverberation. The technique's performance degrades in determining the source DOAs accurately for such cases. Optimization of the MVDR spatial covariance matrix as diagonal loading and analyzing it in short time intervals are also studied in the literature [43] [44].

2.5.4 Sparse recovery based methods

Sparse representation aims to express the signals in small number of coefficients. It is significant for data-intensive applications such as machine learning or communication channel estimation to decrease the computational cost because the amount of data rapidly increases using large number of sensors. If the signal is collected from a sensor, it would be possible to have the sensor emitting already compressed information or represented in sparse before transmission. For example, HOA is a form of the acoustic scene sparse representation. The sparse recovery literature for the signal model in the spatial domain deals with the location-based dictionary items. In this subsection, sparse recovery based techniques for the DOA estimation will be issued.

The signal model for the approximation using a sparse set of atoms from a redundant dictionary $\mathbf{X} \in \Re^{k \times m}$ is given with the equation (2.74).

$$\mathbf{y} = \mathbf{X}\mathbf{A} + \epsilon, \tag{2.74}$$

where A is an $m \times n$ matrix with fewer columns than rows ($n << m$) and ϵ is the residual error that is ideally orthogonal to $\mathbf{X}$. The main goal is to approximate l_p norm to find a vector $\widehat{\mathbf{x}}$ such that the l_p approximation error $\|\mathbf{x} - \widehat{\mathbf{x}}\|_p$ is close to $\min_{\mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|$, where $\mathbf{x}'$ covers all vectors over k non-zero terms in the dictionary $\mathbf{X}$ [45].

Xu et al. propose a method to reduce computational cost for the estimation of DOA using sparse representation of an array cross correlation vectors (ACCV) for linear microphone arrays [46]. The signal model $y(t)$ is given in equation (2.74) with an overcomplete dictionary $[x(\theta_1), x(\theta_2), ..., x(\theta_k)]$ and their coefficients $[a_1(t), a_2(t), ..., a_k(t)]$. The steering vector defined for the far-field source k is given as:

$$x(\theta_k) = [1, v_k, ..., v_k^{M-1}]^T, \tag{2.75}$$

where $v_k = e^{-2j\pi(d/\lambda)\sin(\theta_k)}$, d is the sensor spacing, λ is the signal wavelength and M is the number of sensors. The covariance matrix is calculated using the microphone measurements according to the formula $\mathbf{R} = E\{y(t)y(t)^H\}_{p,q}$, where $p, q \in [1, 2, ..., M]$. The dictionary matrix ($\bar{\mathbf{X}}$) defined for each steering vector is also given as:

$$\bar{\mathbf{X}}(\theta_k) = [v_k^{-(M-1)}, v_k^{-M}, ..., v_k^{-1}, v_k^1, ..., v_k^M, v_k^{M-1}]^T. \tag{2.76}$$

The resulting equation for the estimation of the signal powers $\bar{\mathbf{A}}$ is denoted as $\mathbf{R} = \bar{\mathbf{X}}\bar{\mathbf{A}} + \bar{\epsilon}$. The measurement of sparsity is the non-zero elements in the sparse vector estimation (i.e., the l_0- norm), which is also identified as the source count and sparsity of the solution in a convex optimization problem to find the best fitting dictionary items. The identified angles (θ_k) for the best fitting dictionary items are identified as the DOA estimations.

Another study proposes a sensor array design to obtain DOA estimation via sparse recovery techniques. The study proposes a sparse separate nested pressure vector sensor array (SSN-VSA) via integrating traditional linear acoustic vector sensor array with velocity sensor array [47]. Joint Orthogonal Matching Pursuit (J-OMP) is used to obtain the sparse vector coefficients from the array measurement [48]. Tan et al. also propose two methods, as the greedy method (J-OMP) and sparse relaxation method (Joint LASSO) for the sparse recovery via generating co-prime arrays consisting of two linear arrays with N sensors via spacing Md distances between sensors and $2M$ sensors via spacing Nd distances [49]. This work defines a co-prime array to increase the degrees of freedom from $M + N$ to MN when compared with the classical linear arrays. The methods proposed also outperform the source localization performance when compared with MUSIC.

Sparse representations have also been used with spherical microphone arrays within the SHD framework. A direct approach involves using the spherical harmonic functions evaluated with an overcomplete set of steering directions [50].

$$\mathbf{X_{dict}}(\theta_k, \phi_k) = [Y_0^0(\theta_k, \phi_k), Y_{-1}^1(\theta_k, \phi_k), ..., Y_N^N(\theta_k, \phi_k)]^H, \tag{2.77}$$

where θ_k, ϕ_k is the corresponding angle for the plane wave k and the sparse recovery's reformulation is $\mathbf{Y} = \mathbf{X_{dict}}\mathbf{A} + \eta$. The solution is handled iteratively via re-weighted least square (IRLS) where the source directions are selected as the maximum values of the vector $\mathbf{A}$ via l_0-norm minimizing the number of possible sources, the proposed method solves the equation for each time interval iteratively. The solution would degrade in lower frequencies due to the absence of noise deamplification factor as the mode strength coefficient $\mathbf{B_n}(kr)$, which changes according to the spherical array type and frequency band. Frequency domain sparse recovery analysis also outperforms the DOA estimation for the lower frequencies with the spherical microphone

arrays [51] [52] [53]. The main differences are selecting dictionaries using the frequency mode strength and recovery techniques as convex optimization. The signal model in spherical harmonics is expressed in the following equation.

$$\mathbf{a_{nm}} = \mathbf{Y^H A} + \mathbf{z_{nm}}, \tag{2.78}$$

where $\mathbf{Y^H} = [Y_0^0(\Psi_v), Y_{-1}^1(\Psi_v) \dots Y_N^N(\Psi_v)]^H$ is the spherical harmonic functions for the dictionary item's corresponding direction Ψ_v, $v \in [1 \cdots k]$ is the plane wavenumber, $\Psi_v = (\theta_v, \phi_v)$, and $\mathbf{z_{nm}} = \mathbf{B_n^{-1}}(kr)\eta_{\mathbf{nm}}$ is the noise in spherical harmonics. $\mathbf{a_{nm}}$ is also the spherical harmonic coefficients of the acoustic scene resulting from the spherical harmonic decomposition. Minimizing the error ($\mathbf{z_{nm}}$) results in the best fit for a selected azimuth and elevation Ψ_v, defined as the DOA estimates [52].

2.6 Acoustic Source Separation Techniques

In this review, we will classify existing source separation methods into three categories: underdetermined case, beamforming-based, and sparse recovery-based approaches. Underdetermined cases cover the conditions having more sources than the number of microphone recordings. In this case, it is not mostly possible to capture spatial information with a limited number of microphones. Therefore, stochastic methods such as Independent Component Analysis (ICA) and Non-Negative Matrix Factorization (NMF) are used to check stochastic properties of sound sources [54] [55]. Stochastic methods are recursive processes to minimize the cost function between the signals' features to be separated and measured mixtures. Techniques in stochastic approaches require features such as special frequency spectrogram of pure acoustic signals associated with an object initially. For example, training NMF vector sets with the previously identified source features would improve the source separation performance in separated source quality as Signal to Distortion Ratio (SDR) [56]. By using DOA statistics and updating the beamforming-based NMF vector sets of the active sources for each iteration would also result in better source separation quality [54]. These stochastic techniques would require source features or spatial information to guide the process in the most effective way in contrast with classical methods.

Beamforming based approaches are divided into two categories as fixed [57] or signal-

dependent [58] [59]. Fixed beamformers require spatial information about sources in the mixture to filter out noise and reverberation. They require spatial information for each time interval iteratively or before the computation if the sources are stationary. Yan et al. propose an optimal modal beamforming technique that satisfies constraints to get better beamformer outputs and minimum distortion in the beamforming direction [58]. It optimizes side-lobe levels and white noise gain constraints. Shabtai et al. also provide a mathematical framework for noise suppression and dereverberation in the spherical harmonics domain [59].

Maximum Directivity Factor (MaxDF) beamforming is the most well-known source separation and enhancement technique that would get source direction as input and suppress interference from other sources and noise [60]. On the other hand, some of the other beamforming methods do not explicitly require knowledge of DOA estimates [61] [62] [63]. Their objective is noise reduction while handling the problem of signal distortion. In spherical harmonics, frequency decoupling is also a technique to eliminate the dependence of beamforming to the frequency [22].

Designing a post-filter after beamforming would also improve source separation [64]. Estimated power spectral densities of the active sources can be used to design a post-filter that can be applied to the beamformer's output to enhance the desired signal [65]. A similar technique is applied over RSMA by estimating the active sources' PSD values to improve source separation performance [66].

Sparse recovery based methods use initial template estimations or extract the possible dictionary atoms from spatial or spectral features. K-SVD is a method that is applicable when the number of sources to be extracted are known for multichannel sparse mixtures [67] [68]. It learns the dictionaries to represent acoustic sources in terms of atoms from input mixtures. In the spherical harmonics domain, an online dictionary learning method can reduce the cost of source separation in comparison with using an initial over-complete dictionary [69]. Kalkur et al. propose a technique for joint source DOA estimation and separation using a sparsity-based method [70]. The study uses an overcomplete steering vector representation of plane waves as dictionary atoms. These vectors are used for estimating the DOAs and then source magnitudes.

Another study uses prior information of the intensity vector statistics to separate the convolutive mixtures recorded with a coincident microphone array. Von Mises functions centered at the DOA estimates are tuned with a spread parameter according to the statistics for beamforming [71]. A spatial windowing function is also used to filter acoustics components. Previously defined reference signals are used as templates to separate source signals in a convolutive mixture [72]. This method reveals an NMF-based technique to decompose the measurement matrix's spectral power according to the reference signals. State of the art source separation methods will now be explained in more detail.

2.6.1 Underdetermined Case Source Separation Techniques

Independent Component Analysis (ICA) aims to separate individual components based on the assumptions of statistical independence and non-Gaussianity [73]. Otherwise, this method fails in the separation tasks. Single-channel ICA (SCICA) technique is applicable in underdetermined cases, such as those having signals from only one sensor, although multiple sources can be separated [74]. The linear statistical model for the mixture is given as :

$$\mathbf{x} = \sum_i a_i s_i = \mathbf{as}, \tag{2.79}$$

where $\mathbf{a} = [a_1, ..., a_N]$ and $\mathbf{s} = [s_1, ..., s_N]^T$. If the coefficient vector $\mathbf{a}$ is invertible, source vector $\mathbf{s}$ can be obtained using the formula $\mathbf{s} = \mathbf{a^{-1}x}$. The constraint on building the linear signal model is assuming that the source signals are statistically independent. If a single channel data is separated into blocks with a length of N, SCICA is applied for each block to separate the signals.

$$x_s^{(i)}(nN - k + 1) = A_{(k,i)}^{-1} \sum_{j=1}^{N} x(nN - j + 1), \tag{2.80}$$

where i is the source index, j is the block index, k is the sampled index in the block, and N is the number of samples in the block. The validation of the source separation results can already be checked with the error $\epsilon = \mathbf{x} - \sum_{i=1}^{N} \mathbf{x_s}^{(i)}$. Perfect decomposition results in zero residual error.

In the case where the single-channel mixture comprises coherent sources, individual sources can be separated over the mixture's magnitude spectrogram using NMF-based

methods. SCICA cannot be applied for this condition because the main constraint was the statistical independence between the signals to apply SCICA. NMF is a stochastic method to derive two non-negative matrices ($W \in \Re_{\geq 0}^{m \times k}$,$H \in \Re_{\geq 0}^{k \times n}$) from an original matrix ($X \in \Re_{\geq 0}^{m \times n}$) by minimizing the cost function $|\mathbf{X} - \mathbf{WH}|$ in an iterative objective [75].

In the context of NMF, $\mathbf{W}$ is template vector, and $\mathbf{H}$ is the vector's corresponding magnitudes. Although the decomposition process of $\mathbf{X} \approx \mathbf{WH}$ can be initiated with randomly selected vectors and coefficients, templates can also be changed in time intervals with a set of vectors to improve its convergence time and performance as supervised NMF [56].

In the context of an underdetermined case where there are more sources than the number of sensors, deep learning techniques are the popular approaches to solve these problems. However, these techniques have to be trained with the possible ground truth before the decomposition. Seetharaman et al. propose an unsupervised technique to provide a confidence measure to assess the training data. The confidence measure matrix is calculated using the sources' spatial information via calculating the Inter-channel phase difference (IPD) [76]. The main assumption is grouping the time-frequency features based on the calculated IPD and using them in the training phase. Xiao et al. propose another method for speech separation using deep learning, which is trained with the previously identified ground truth [77]. This method works over a single channel without the acoustic sources' spatial information, but it depends on training with the carefully selected ground truth to output clear results.

The source separation techniques used over underdetermined case recordings obtain good performance if the number of sources and semantic information of the desired signals to be separated are known. However, it is not easy to separate the signals from the real-world scenarios without any preliminary information, such as the spatial information. Microphone array signal processing techniques also provide accurate spatial information of the active sources from the environment to generate adaptive filters based on the DOA of the sources. In the next sections, the techniques that use the spatial information or get the DOA before the source separation computations will be investigated.

2.6.2 Beamforming based methods

Beamforming can be categorized into two types: fixed and signal-dependent. Fixed beamformers steer in a fixed direction, and their steering weights are optimized according to the white noise gain and main lobe widths. On the other hand, signal-dependent beamforming makes an optimization belonging to the noise and signal characteristics and the DOA statistics. This section will be introduced in two parts as fixed and signal-dependent beamformers.

2.6.2.1 Fixed beamformers

Delay and sum beamformer is the simple fixed beamformer whose weights are delays resulting from plane wave arriving with respect to the spherical array sensors. This type of beamforming also outputs maximum white noise gain (WNG) and provides maximum robustness to noise. Note that this type of beamforming is applicable only if the plane wave is propagating in the free field such as open array configurations [16]. The implementation over the open spherical array using the delay parameter defined for the plane wave can be expressed as :

$$w^*(k,r) = e^{-i\mathbf{k_l}\mathbf{r}}, \tag{2.81}$$

where $\mathbf{k_l}$ is the propagating wave vector and $\mathbf{r}$ is the representation of the spherical array in Cartesian coordinates $[r_x, r_y, r_z]$. This equation can be written in the spherical harmonics function as:

$$w_{nm}(k) = 4\pi i^n j_n(kr) \left[Y_n^m(\theta_l, \phi_l)\right]^*, \tag{2.82}$$

where θ_l, ϕ_l is the plane wave propagating directions. The axis-symmetric beamforming weights defined for the delay and sum beamforming is expressed as [78]:

$$d_n(k) = \left|4\pi i^n j_n(kr)\right|^2. \tag{2.83}$$

The resulting beamforming formula defined with the beamforming weights are given according to the formula :

$$y(t,\Omega_a) = \sum_{n=0}^{N} \left|b_n(kr)\right|^2 \sum_{m=-n}^{n} Y_{nm}(\Omega_a) \underbrace{\sum_{q=1}^{Q} w_q p_q(t) Y_{nm}^*(\Omega_q)}_{\text{SHD}\,:\,a_{nm}(\tau,k)}, \tag{2.84}$$

where $b_n(kr) = 4\pi i^n j_n(kr)$ is the open array dependent mode strength component calculated according to the wavenumber and distance.

Rafaely et al. present the spherical beamforming using plane-wave decomposition and express the beamforming equation by separating its direction and frequency-dependent parts [22]. This method is fixed beamforming that depends on the spherical harmonics order limitation (N) of the computation, the steering direction (Ω_a), and the array geometry. More specifically, the output of this fixed beamformer (also known as maximum directivity factor beamformer [maxDF]) is given in the time-frequency domain as:

$$y(\tau, \varkappa, \Omega_a) = \sum_{n=0}^{N} \frac{1}{b_n(kr)} \sum_{m=-n}^{n} Y_{nm}(\Omega_a) \underbrace{\sum_{q=1}^{Q} w_q p_q(\tau, \varkappa) Y_{nm}^*(\Omega_q)}_{\text{SHD}: a_{nm}(\tau, \varkappa)} . \tag{2.85}$$

This relation can also be expressed in the time domain with the following equation without sensor noise deamplification term $b_n(kr)$ which is dependent to frequency.

$$y(t, \Omega_a) = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_{nm}(\Omega_a) \underbrace{\sum_{q=1}^{Q} w_q p_q(t) Y_{nm}^*(\Omega_q)}_{\text{SHD}: a_{nm}(t)} \tag{2.86}$$

Without the establishment of the wavenumber k, this situation can cause the amplification of sensor noise in lower frequencies given in Fig. 2.4. If the microphones positioned around the array have higher sensor noise, noise amplification can be observed for the beamforming calculated without the mode strength coefficient.

2.6.2.2 Signal-dependent beamformers

A tradeoff beamformer is proposed to reduce the noise and minimize the speech distortion [61] [79] by using the signal model in the spherical harmonics domain. The signal model over the microphones can be expressed in the frequency domain as (2.88).

$$\mathbf{p_q}(k) = \mathbf{p_q}(k)\mathbf{s}(k) + \eta_{\mathbf{q}}(k) \tag{2.87}$$

$$= \mathbf{x_q}(k) + \eta_{\mathbf{q}}(k) \tag{2.88}$$

The signal model in the spherical harmonics domain after dividing with the mode strength is expressed as :

$$\mathbf{a_{nm}}(k) = \mathbf{B^{-1}}(k)[\mathbf{V_{nm}}(k)\mathbf{S}(k) + \eta_{\mathbf{nm}}(k)] \tag{2.89}$$

$$= \tilde{\mathbf{x}}_{\mathbf{nm}}(k) + \tilde{\eta}_{\mathbf{nm}}(k) \tag{2.90}$$

The desired signal vector, $\mathbf{S}(k)$, presented in spherical harmonics is divided with the zeroth-order and degree of the spherical harmonics $X_{00}(k)$, which is direction-independent. This signal $X_{00}(k)$ is used as a reference to reveal the tradeoff beamformer. Beamforming is expressed with the term $X_{00}(k)$ in the equation (2.91).

$$\mathbf{a_{nm}} = \tilde{\mathbf{d}}X_{00} + \tilde{\eta}_{\mathbf{nm}}, \tag{2.91}$$

where $\tilde{\mathbf{d}} = [1, \frac{X_{1(-1)}(k)}{X_{00}(k)}, ..., \frac{X_{NN}(k)}{X_{00}(k)}]$ comprises eigenbeams normalized with respect to the $(n, m) = (0, 0)$ component. Tradeoff beamformer is multiplication of the eigenbeams with a complex magnitude vector ($\mathbf{h^H}$) and summing over the results.

$$\mathbf{Z} = \mathbf{h^H}\tilde{\mathbf{d}}X_{00} + \mathbf{h^H}\tilde{\eta}_{\mathbf{nm}} \tag{2.92}$$

The output power of the beamformer is denoted with the equation by assuming the noise and signal are uncorrelated and $\mathbf{\Phi_Z} = E\{\mathbf{ZZ^H}\}$.

$$\mathbf{\Phi_Z} = |X_{00}|^2|\mathbf{h^H}\tilde{\mathbf{d}}|^\mathbf{2} + \mathbf{h^H}\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}\mathbf{h} \tag{2.93}$$

Tradeoff beamformer ideally deals with two limitations, as the noise reduction factor and speech distortion index. The relations between noise and the desired signal are given with the following constraints.

$$\varrho_{nr} = \frac{|X_{00}|^2}{\mathbf{h^H}\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}\mathbf{h}} \tag{2.94}$$

$$\xi_{sd} = \frac{|\mathbf{h^H}\tilde{\mathbf{d}}X_{00} - X_{00}|^2}{|X_{00}|^2} = |\mathbf{h^H}\tilde{\mathbf{d}} - 1|^2 \tag{2.95}$$

By satisfying the following relations, the tradeoff beamformer is designed to filter out the noise and minimize the speech distortion.

$$\min |\mathbf{h^H}\tilde{\mathbf{d}} - 1|^2 \; s.t. \; \mathbf{h^H}\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}\mathbf{h} = \varrho_{\mathbf{nr}}^{-\mathbf{1}}|X_{00}|^2 \; ; \varrho_{nr}^{-1} < 1 \tag{2.96}$$

Here, the resulting beamformer by optimizing the cost function is given with the equation (2.97). $\nu = 0$ results in a MVDR beamformer and $\nu = 1$ results in a multi-channel Wiener filter.

$$h_\nu = \frac{|X_{00}|^2\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}^{-\mathbf{1}}\tilde{\mathbf{d}}}{\nu + |X_{00}|^2\tilde{\mathbf{d}}^{\mathbf{H}}\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}^{-\mathbf{1}}\tilde{\mathbf{d}}} \tag{2.97}$$

The unknown parameters of the tradeoff filter are noise PSD matrix $\mathbf{\Phi}_{\tilde{\eta}_{\mathbf{nm}}}$ and desired signal spherical harmonics coefficient ratio vector $\tilde{\mathbf{d}}$. Desired Speech Presence Probability (DSPP) was proposed to calculate the unknown parameters using DOA-based statistics and speech presence probability (SPP) estimation [63] [62]. These parameters were used to calculate the noise PSD.

Another outperforming approach to fixed spherical beamformer is the Linearly Constrained Minimum Variance (LCMV) method for noise reduction [60]. The noise and desired signal are assumed to be uncorrelated because noise is assumed to be stationary. PSD of the noise can be estimated if there is not any source to be located. This work deals with designing a beamformer to minimize the noise power such that:

$$\min\{\mathbf{h^H \Phi}_\eta \mathbf{h}\}\ s.t.\ \mathbf{h^H a_{nm}} = \delta, \tag{2.98}$$

where $\mathbf{h^H}$ provides normalization of the spherical harmonic coefficients without distorting the desired signal. This filter is designed to reduce noise without introducing distortion. The resulting LCMV filter is shown to be :

$$\mathbf{h} = \mathbf{\Phi}_\eta^{-1}\mathbf{a_{nm}}(\mathbf{a_{nm}^H \Phi}_\eta^{-1}\mathbf{a_{nm}})^{-1}\delta. \tag{2.99}$$

The signal characteristics can also be identified using the PSD of an audio signal. Suppressing the undesired acoustic sources, background noise, and late reverberation from the acoustic environment recordings can be aimed via the PSD estimation. While there are proposed techniques used in the source separation, post-filter usage can enhance the desired signal by improving the separation performance. PSD estimation of the desired source, reverberation, and noise are used to design a Wiener post-filter, placed as a multiplication term at the end of the conventional beamforming. Yamamoto et al. propose a method for estimating the direct source PSD to design a Wiener post-filter [64] realized as a time-frequency mask given as:

$$W(\tau, \varkappa) = \frac{\Phi_{DS}}{\Phi_{DS} + \Phi_R}, \tag{2.100}$$

where Φ_{DS} is the estimated PSD of the desired source and Φ_R is the estimated PSD of the reverberation and noise at the time-frequency bin.

The article's method assumes that the DOA of the desired source is known [64]. The main idea is to define several steering directions, including the desired source's DOA.

Then, the MaxDF is applied for each steering direction, and a magnitude $S^l(\tau, \varkappa)$ is obtained for each direction. The resulting formula is given as in the equation (2.101).

$$W(\tau, \varkappa) = \frac{|S^l(\tau, \varkappa)|^2}{|S^l(\tau, \varkappa)|^2 + \frac{1}{L-1}|\sum_{l=2}^{L} S^l(\tau, \varkappa)|^2}, \tag{2.101}$$

where $l = 1$ is the desired source's beamforming output directions and outputs obtained for other steering directions are assumed as the reverberation and noise.

Noise reduction and dereverberation are obtained by multiplying the desired source's beamforming output as in the following equation.

$$Z(\tau, \varkappa) = S^l(\tau, \varkappa)W(\tau, \varkappa) \tag{2.102}$$

There are other methods for the PSD estimation to build a Wiener post-filter for the separation of each active source [66]. The spatial correlation matrix between the spherical harmonic coefficients of a sound field can also be used as the starting point for PSD computation. The results are evaluated for the maximum directivity beamformer, and separation outputs are improved with the Wiener post-filter defined for each acoustic source. The details of this study are omitted from this review, and the interested reader is referred to [66] for a discussion.

As an alternative, Multichannel Wiener Filtering (MWF) approach uses long short-term memory (LSTM) Recurrent Neural Network for the noise, and desired signal PSD estimations [80]. Perotin et al. used 16 spatially distributed RIR recordings measured in a room with $T_{60} = 270$ ms and convolved them with the target speech, which is picked as a 10 s duration including babble noise of 20 dB SNR and these convolved audio files are inserted to the data set for training. The output of the network is the estimated covariance matrices of the desired speech $s_w(\tau, \varkappa)^2$ and noise $n_w(\tau, \varkappa)^2$. By using the diagonal entries of these matrices, the PSD of the target masks as a Wiener post-filter were also calculated using the equation (2.100).

The acoustic signals recorded in the room, which has higher reverberation, can be split into two objects, such as direct and diffuse fields. Acoustic signal enhancement is the method to separate the diffuse field from the original sound recordings to obtain dry and perceptually clear speech and music signals. Hu et al. propose an estimation

method for clean spherical harmonic coefficients [81].

$$\beta_{nm}(k) \approx \frac{S_{a_{nm}a_{00}}(k)}{S_{a_{00}a_{00}}(k)}, \tag{2.103}$$

where $\beta_{nm}(k)$ is estimated through the PSD equations with respect to the direction-independent spherical harmonics. The PSD formulas for the numerator and denominator can be calculated as $S_{a_{nm}a_{00}}(k) = E\{a_{nm}(k)a_{00}^*(k)\}$ and $S_{a_{00}a_{00}}(k) = E\{a_{00}(k)a_{00}^*(k)\}$. Then, the clean spherical harmonic coefficients are estimated with the formula $y_{nm}(k) = \beta_{nm}(k)y_{00}(k)$. These coefficients can be used as enhanced acoustic signal coefficients for the source separation. MaxDF using the enhanced coefficients in the DOA direction outputs better source separation results.

2.6.3 Sparse recovery based methods

Dictionary-based sparse recovery methods are used for the representation of the signal with the sparse mixtures. The main idea behind these techniques is defining an over-complete dictionary and expressing the mixture using the atoms in the dictionary. The algorithm employing an overcomplete dictionary would degrade in the computational efficiency. Therefore, online dictionary learning-based methods are also used to decrease the cost for the algorithms [69]. The mixture model for the sparse recovery is given in the equation (2.74). Noise can be regarded as the residual error that can not be efficiently represented. The dictionary to employ in sparse representations depends on the domain of operation, e.g., time, frequency, space, etc.

A group of methods use SHD in spatial domain for signal extraction [70] [69]. Abolghasemi et al. also use a time-domain representation of the signals, and the K-SVD method is used to extract the selected dictionary coefficients [67]. Frequency domain vector matching in underdetermined mixtures is also investigated as an alternative in the article [68].

The pressure over the spherical microphone array (SMA) consisting Q microphones and D dictionary atoms would be expressed as in the following equation with $\mathbf{p}(k) = [p_1(k), ..., p_Q(k)]^T$, $\mathbf{v}(k, \Psi) = [v(k, \Psi_{1,1}), ..., v(k, \Psi_{Q,D})]$ and $\mathbf{s}(k) = [s_1(k), ..., s_D(k)]^T$.

$$\mathbf{p} = \mathbf{V}\mathbf{s} + \eta, \tag{2.104}$$

where $\mathbf{V}$ is a dictionary matrix comprising steering vectors from the microphone positions to the possible point source positions. The steering vector is expressed with the following equation.

$$\mathbf{v}(k, \Psi_{q,d}) = e^{\mathbf{k_d^T r_q}} \mapsto \mathbf{Y}(\Omega_q)\mathbf{B}(kr)\mathbf{Y^H}(\Psi_d), \tag{2.105}$$

where $\mathbf{r_q}$ is microphone positioning vector with respect to the spherical microphone array center and $\mathbf{k_d}$ is the wave vector defined as $\mathbf{k_d} = [k_d \cos(\phi_d) \sin(\theta_d), k_d \sin(\phi_d) \sin(\theta_d), k_d \cos(\theta_d)]$ where $k_d = w_d/c$ is the wavenumber. The resulting acoustic scene equation with the microphones' pressure representation is given in spatial domain as [82]:

$$\mathbf{p_q}(k) = \mathbf{Y}(\Omega_q)\mathbf{B}(kr)\mathbf{Y^H}(\Psi_d)\mathbf{s}(k) + \eta_{\mathbf{q}}(k). \tag{2.106}$$

After application of SHD over the measured microphone signals and dividing the coefficients with the mode strength are followed by multiplying with $\mathbf{B^{-1}}(kr)\mathbf{Y^H}(\Omega_q)$ from the left side. Then, the result is denoted as in the equation (2.107).

$$\mathbf{a_{nm}} = \mathbf{Y_d^H S} + \underbrace{\mathbf{B^{-1}Y_q^H}\eta}_{\text{spherical harmonic noise } z(k)}, \tag{2.107}$$

where the matrices are in the form of $\mathbf{a_{nm}} \in \mathbb{C}^{(N+1)^2 \times 1}$, $\mathbf{B^{-1}} \in \mathbb{C}^{(N+1)^2 \times (N+1)^2}$, $\mathbf{S} \in \mathbb{C}^{D \times 1}$, $\mathbf{Y_d^H} \in \mathbb{C}^{(N+1)^2 \times D}$, and $\mathbf{Y_q^H} \in \mathbb{C}^{(N+1)^2 \times Q}$. The over-complete dictionary definitions comprise $\mathbf{Y_d^H}$ which is calculated in accordance with the possible number of atoms (D). In an iterative objective, the best fitting dictionary definitions as the number of sources S are found up to the minimizing the residual error using a convex or non-convex optimization method.

Expressing an acoustic field in terms of plane waves makes the sparse recovery easier in the analysis. This makes dictionaries comprising spherical harmonics a good choice if the sparse representation is to be pursued in the SHD domain [70]. The dictionary atoms for this specific case are :

$$\Phi = [Y^H(\Psi_1), Y^H(\Psi_2), ..., Y^H(\Psi_d)]^T, \tag{2.108}$$

where $Y(\Psi_d) = [Y_0^0(\Psi_d), Y_1^{-1}(\Psi_d), ..., Y_N^N(\Psi_d)]$ is the row of Φ, the number of columns is much more greater than the number of available source count S. The constraints given to calculate the localization feature of the desired source (s) is given

within the following equation.

$$\min ||\mathbf{a_{nm}} - \Phi\mathbf{Rs}||_2 + \lambda||\mathbf{R}||_1 \; s.t. \; \mathbf{R^T R} = I, \tag{2.109}$$

where $\mathbf{R}$ is the real magnitude matrix representing the selection gains of the dictionary atoms. The solution as the localization feature ($\mathbf{A_n} = \Phi\mathbf{R}$) and the source strength $\mathbf{s}$ to the non-convex optimization problem can be tackled using the Bregman iteration [83]. The desired source magnitude is also obtained using the pseudo inverse formula of the localization feature as:

$$\mathbf{s} = (\mathbf{A_n^H A_n})^{-1}\mathbf{A_n^H a_{nm}}, \tag{2.110}$$

which provides a solution that is optimal in the least square sense. The source strengths are associated with the separated sources based on the localization features.

While source separation using an overcomplete dictionary is computationally costly, using the DOA estimation parameters to generate a more compact dictionary was also proposed as a possible solution [69]. The method learns to create a dictionary from localization statistics gradually. The algorithm starts with an empty dictionary, and it tries to fill the dictionary according to the DOA estimates. In the case of heavily underdetermined multichannel sound mixtures, a phase optimized K-SVD and Orthogonal Matching Pursuit (OMP) based optimization methods are proposed [68]. These methods examine two conditions whether there is an initial dictionary or not.

2.7 Conclusions

This chapter reviewed the theory of spherical microphone arrays, including the array design, spherical harmonic decomposition, and steered response functions. These were the necessary foundations of the methods proposed in this thesis. The chapter also reviewed the state of the art methods related to DOA estimation that can be classified into four main categories as intensity-based, subspace-based, SRP, and sparse recovery based techniques. State of the art source separation techniques related to the spherical harmonics are also included in this chapter, such as Wiener-Post filtering, sparse recovery, and beamforming based approaches. In the following chapter, the DOA estimation and source separation methods developed during the research carried out within this thesis's context are described.

CHAPTER 3

DESIGN OF THE STUDY

Object-based audio comprises metadata and channel recordings to present the immersive audio for the end-users. Generation of metadata, which contains DOA, gain and semantic information of the sources, requires a complex architecture containing several techniques, and it is a field of research in immersive audio. Therefore, useful tools are required for the creation of metadata automatically. This chapter expresses this thesis's main contributions by providing practical DOA estimation and source separation tools to transcode between scene-based to object-based audio. While the analysis investigated in this chapter starts from SHD through rigid spherical microphone array recordings, the algorithms would also operate via taking the spherical harmonics as input. Firstly, the DOA estimation techniques will be discussed, and then, the algorithms developed for the sound source separation are introduced.

3.1 DOA Estimation Techniques

Source DOAs are required to create the metadata and obtain better performances for the source separation in highly reverberant environments. While DOA estimation algorithms are also available using the spherical harmonics, accurate, computationally efficient, and robust to noise algorithms are demanded for object identification, tracking, and source separation. Thus, this section contains two different source DOA estimation methods named HiGRID and RENT. As HiGRID decreases the computational cost compared to SRP-based DOA estimation methods and provides accurate DOA estimations, RENT is a sparse recovery-based algorithm that identifies the time-frequency bins having single source contributions as well as it provides DOA

estimation with the histogram clustering as an output.

3.1.1 Spatial Entropy based Hierarchical Grid Refinement (HiGRID)

Steered Response Power Density (SRPD): Maximum power detection over the SRP defined in the steering directions would fail to discover local peaks if the search grid has a low resolution. Therefore, a more appropriate functional named Steered Response Power Density (SRPD) is defined for an arbitrary, bounded surface element, $\mathcal{S}_i$, over the unit sphere:

$$\mathcal{P}_i(k) = \frac{1}{R_i} \int_{\mathcal{S}_i} |y_N(k, \theta_i, \phi_i)|^2 d\mathcal{S}_i, \tag{3.1}$$

where $R_i \triangleq R(\mathcal{S}_i)$ is the area of the surface element and $y_N(k, \theta_i, \phi_i)$ is the steered response function decoupled from the frequency as described in equation (2.36). SRPD and SRP can be related with the following equation (3.2).

$$\lim_{R_i \to 0} \mathcal{P}_i(k) = |y_N(k, \theta_i, \phi_i)|^2, \tag{3.2}$$

where (θ_i, ϕ_i) is the center of the differential region, $\mathcal{S}_i$.

By using the spherical harmonic coefficients and mode strength terms given in (2.53), SPRD can be expressed as:

$$\mathcal{P}_i(k) = \sum_{n,m,n',m'} \frac{a_{nm}(k) a^*_{n'm'}(k)}{b_n(kr) b^*_{n'}(kr)} Q^{m,m'}_{n,n'}(\mathcal{S}_i), \tag{3.3}$$

where,

$$Q^{n',m'}_{n,m}(\mathcal{S}_i) = \frac{1}{(4\pi)^2 R_i} \int_{\mathcal{S}_i} Y^m_n(\theta, \phi) \left[Y^{m'}_{n'}(\theta, \phi) \right]^* d\mathcal{S}_i. \tag{3.4}$$

From the resulting equation $\mathcal{P}_i(k)$, SRPD can be split into the frequency and direction-dependent terms via SHD. This representation allows isolating the integration terms $Q^{n',m'}_{n,m}(\mathcal{S}_i)$ that do not depend on the SHD coefficients. It is possible to express the sum in (3.3) as the grand sum of the matrix:

$$\mathbf{H}_i = \mathbf{A} \circ \mathbf{Q}_i, \tag{3.5}$$

where $\circ$ represents the Hadamard (i.e. element-wise) product, $\mathbf{A} = \mathbf{a}\mathbf{a}^H$ is an $(N +$

$1)^2 \times (N+1)^2$ matrix with

$$\mathbf{a} = \left[\frac{a_{00}(k)}{4\pi b_0(kr)}, \frac{a_{1-1}(k)}{4\pi b_1(kr)}, \cdots, \frac{a_{NN}(k)}{4\pi b_N(kr)} \right]^T \tag{3.6}$$

and the $(N+1)^2 \times (N+1)^2$ cross spatial density matrix $\mathbf{Q}_i$ is given as:

$$\mathbf{Q}_i = \begin{bmatrix} Q_{0,0}^{0,0}(\mathcal{S}_i) & Q_{0,0}^{1,-1}(\mathcal{S}_i) & \cdots & Q_{0,0}^{N,N}(\mathcal{S}_i) \\ Q_{1,-1}^{0,0}(\mathcal{S}_i) & Q_{1,-1}^{1,-1}(\mathcal{S}_i) & \cdots & Q_{1,-1}^{N,N}(\mathcal{S}_i) \\ \vdots & \vdots & \ddots & \vdots \\ Q_{N,N}^{0,0}(\mathcal{S}_i) & Q_{N,N}^{1,-1}(\mathcal{S}_i) & \cdots & Q_{N,N}^{N,N}(\mathcal{S}_i) \end{bmatrix}, \tag{3.7}$$

where both $\mathbf{A}$ and $\mathbf{Q}_i$ are Hermitian. The grand sum of $\mathbf{H}_i$ can be represented as:

$$\mathcal{P}_i = \mathbf{e}^T \mathbf{H}_i \mathbf{e} = \mathbf{e}^T \left(\mathbf{A} \circ \mathbf{Q}_i \right) \mathbf{e}, \tag{3.8}$$

where $\mathbf{e}$ is a column vector of ones, and i is the region where SRPD is computed. This expression can be simplified by employing an identity of the Hadamard product [84] and the SVD of the cross spatial density matrix ($\mathbf{Q}_i^T$) such that:

$$\mathcal{P}_i = \mathrm{tr}(\mathbf{A}\mathbf{Q}_i^T) = \mathrm{tr}(\mathbf{A}\mathbf{U}_i\mathbf{\Sigma_i}\mathbf{U}_i^H);\ s.t. \mathbf{Q}_i^T = \mathbf{U}_i\mathbf{\Sigma_i}\mathbf{U}_i^H, \tag{3.9}$$

where $\mathrm{tr}(\cdot)$ is the trace operator to sum the main diagonal entries of the matrix. Here, the columns of $\mathbf{U}_i$ are the eigenvectors, and the diagonal matrix $\mathbf{\Sigma}_i$ contains the eigenvalues $\lambda_{i,m}$ of $\mathbf{Q}_i^T$. By using the cyclic permutation invariance property of the trace operator [84], SRPD is expressed as:

$$\begin{aligned} \mathcal{P}_i &= \mathrm{tr}(\mathbf{U}_i^H \mathbf{A} \mathbf{U}_i \mathbf{\Sigma}_i) && (3.10) \\ &= \mathbf{e}^T \left(\mathbf{U}_i^H \mathbf{A} \mathbf{U}_i \circ \mathbf{\Sigma}_i \right) \mathbf{e}. && (3.11) \end{aligned}$$

Since the eigenvalue matrix $\mathbf{\Sigma}_i$ is diagonal, the resulting Hadamard product contains a sum of the diagonal elements of $\mathbf{U}_i^H \mathbf{A} \mathbf{U}_i$ weighted by the eigenvalues, $\lambda_{i,m}$ of $\mathbf{Q}_i^T$. Substituting the trace operator with the summation, we obtain :

$$\mathcal{P}_i = \sum_m \lambda_{i,m} \left[\mathbf{U}_i^H \mathbf{A} \mathbf{U}_i \right]_{m,m}, \tag{3.12}$$

or in a more compact form as:

$$\mathcal{P}_i = \| \mathbf{U}_i^H \mathbf{a} \|_{\mathbf{\Sigma}_i}^2, \tag{3.13}$$

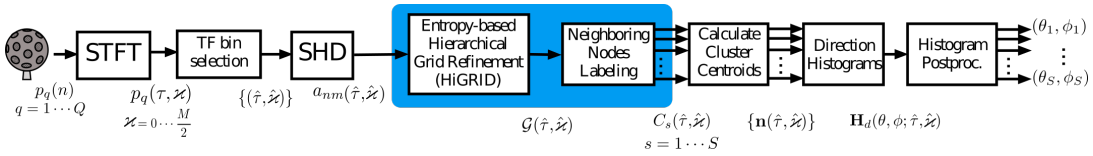


Figure 3.1: The flow diagram of the proposed algorithm. The core part is highlighted with a blue box.©2018 IEEE [1]

where $\|\mathbf{x}\|_{\mathbf{\Sigma}} = (\mathbf{x}^H\mathbf{\Sigma}\mathbf{x})^{1/2}$ represents the weighted norm. Since the eigenvalue matrix $\mathbf{\Sigma}$ only depends on the cross spatial density matrix Q_i^T, it can be used to eliminate the computational cost related to eigendecomposition and stored offline. As a consequence, SRPD can be calculated using the eigenvalues and the eigenvectors of the cross spatial density $\mathbf{Q}_i$, once the SHD coefficients are available.

The exposition above assumes that all the eigenvalues and eigenvectors of the spatial density matrix can be calculated and stored offline. However, most of the eigenvalues will be very small, depending on the area where the SRPD is calculated. Defining an energy ratio ensures that fewer eigenvalues and eigenvectors can be used in order to reduce the computational cost resulting from the calculation of SRPD.

$$E_{i,M} \triangleq \sum_{m=0}^{M} \lambda_{i,m}^2 \Big/ \sum_{m=0}^{(N+1)^2} \lambda_{i,m}^2 \tag{3.14}$$

The largest $M < (N+1)^2$ eigenvalues and the corresponding eigenvectors can be selected, for which this ratio is greater than a given threshold close to unity. This way, the computational cost can be substantially reduced without any significant effect over the calculated SRPD for the corresponding area. In this study, the largest M eigenvalues and eigenvectors satisfying the condition $E_{i,M} \geq 0.99$ are used.

Spatial Entropy Based Grid Refinement: Searching a fine grid on the sphere to identify maxima is a process that dramatically increases the computational cost. One of the main contributions of this thesis is an entropy-based, data-driven, and high precision DOA estimation method. The proposed approach involves defining data-dependent search grids using spatial entropy instead of a full grid search in finer azimuth and elevation resolutions. The general purpose is to detect local DOA estimates from the selected time-frequency bins and establish each local DOA estimates

to a DOA histogram map. Clustering of DOA estimates and defining each cluster's mean index from the histogram output the final DOA estimates for the corresponding time interval.

For this purpose, short-time Fourier transform (STFT) is applied for each microphone channel first. This process is followed by a time-frequency bin selection criteria that choose bins likely to contain one or more strong direct path components. In this context, the algorithm only uses the time-frequency (TF) bins $(\tau, \varkappa)$ that correspond to onsets. SHD is applied for the selected SHD coefficients in the corresponding TF bins to find spherical harmonic coefficients, $a_{nm}(\tau, \varkappa)$. The method that we call hierarchical grid refinement (HiGRID) is applied for forming adaptive SRPD maps for the selected TF bins. The output of HiGRID for each time-frequency bin is a multiresolution SRPD map formed according to a spatial entropy criterion. Segmentation and labeling using a variant of connected components labeling (CCL) of this map provide the source count and the mean index of each cluster for local DOAs. A global histogram of the centroids obtained from all the processed TF bins is then used to estimate the global source DOAs. The diagram in Fig. 3.1 shows the algorithm, which consists of 1) pre-processing, 2) SRPD map calculation via grid refinement, 3) local DOA estimation, and 4) post-processing stages.

Pre-processing: STFT, Onset Detection and SHD: Pre-processing steps involve three cases as i) converting the microphone signals from the time domain to the time-frequency domain, ii) selecting the time-frequency bins that would likely contain one or more sources but no reflections or reverberation for local DOA estimation, and iii) calculating the SHD coefficients from the microphone array signals for these bins.

In the first pre-processing stage, a windowed Fourier transform is used to represent the recorded microphone signals $p_q(t)$ in the time-frequency domain, $p_q(\tau, \varkappa)$, where q is the microphone number, and τ , $\varkappa$ are the time and frequency indices, respectively.

The second pre-processing stage involves selecting bins where the proposed algorithm will be applied to find local DOAs. Based on the presupposition that onsets contain strong direct-path components, TF bins that correspond to onsets are identified as good candidates. For this purpose, a spectrum-based onset detection algorithm

can be applied to the average sum of all microphone elements given as:

$$p_o(\tau, \varkappa) \approx \frac{1}{Q} \sum_{q=1}^{Q} p_q(\tau, \varkappa), \tag{3.15}$$

to obtain the time indices, $\{\hat{\tau}\}$, at which the onsets occur. This is followed by the selection of the bins, $(\hat{\tau}, \hat{\varkappa})$ that have a normalized energy above a prescribed threshold. The preferred algorithm is *SuperFlux* that is robust to short time-frequency modulations [85]. The set of TF bins that are selected by this way will be represented as $\mathcal{T} \triangleq \{(\hat{\tau}, \hat{\varkappa})\}$.

Computation over the TF bins, which do not have any source activity would also be redundant, and this condition significantly increases the computational complexity. Other possible techniques to detect the source activity over the TF bins were also proposed [31] [30] [2].

Finally, SHD coefficients are calculated as in (3.16) such that:

$$a_{nm}(\hat{\tau}, \hat{\varkappa}) = \sum_{q=1}^{Q} w_q p_q(\hat{\tau}, \hat{\varkappa}) [Y_n^m(\theta_q, \phi_q)]^* \tag{3.16}$$

for all the onset detected bins, $(\hat{\tau}, \hat{\varkappa}) \in \mathcal{T}$. These spherical harmonic coefficients are used in the computation of SRPD maps at their respective time-frequency locale.

Hierarchical Grid Refinement (HiGRID): Local peak searching over the sphere requires a suitable grid. Each grid has to be grouped in a hierarchy such that each refinement of the grid element also encompasses the spherical region of the coarser grid. Each grid element can be represented as a node in a tree structure. A feasible grid property is to be approximately, even if not precisely, uniform.

Gaussian sampling (i.e., sampling azimuth and elevation uniformly and separately) increases the computational cost since sampling azimuth uniformly close to the poles of a sphere generates non necessarily fine sampling at these positions. Therefore, Hierarchical Equal Area isoLatitude Pixelisation (HEALPix) [86] grid is used in this study to model the searching regions as it allows an efficient representation of spherical data using quadrilateral, non-overlapping, and refinable grid elements (i.e., pixels) on a sphere. At the resolution level, $K \in \mathbb{N}$, HEALPix provides $12 \cdot 2^{2K}$ equal-area

grid elements. The elements are separated uniformly with an angular resolution (in radians) of:

$$\Theta_{\Delta} = \sqrt{\frac{3}{\pi}} \frac{\pi}{3 \cdot 2^K}, \tag{3.17}$$

and their areas depend on the resolution level such that:

$$A_K = \frac{4\pi R^2}{12 \cdot 2^{2K}}, \tag{3.18}$$

where R is the radius of the sphere [86].

Isolatitude pixelization partitions a grid element, $\mathcal{S}_{K,m}$ at the resolution level K with index m, into four new grid elements neighboring each other, at the higher resolution level, $K+1$, such that:

$$\mathcal{S}_{K,m} = \mathcal{S}_{K+1,4m} \oplus \mathcal{S}_{K+1,4m+1} \oplus \mathcal{S}_{K+1,4m+2} \oplus \mathcal{S}_{K+1,4m+3} \tag{3.19}$$

where $\{\mathcal{S}_{K+1,4m+k}\}$ for $k = 0 \cdots 3$ are the four higher resolution grid elements. This allows for a *quadtree* representation where a spherical distribution can be expressed at different levels of detail for a limited number of sampling points. Since HEALPix provides a representation over the sphere at different sampling levels, it also provides a hierarchy between the pixel indices at different levels.

For the case where multiple coherent sources are present in a TF bin, SRPD will have more than one meaningful local peak that correspond to the DOAs of active sources or their reflections. In order to find these maxima without steering in all directions, it is necessary to identify the possible grid elements that would contain these peaks in a higher representation level of HEALPix. Spatial entropy [87], which is a measure of the spatial disorganization, is used to select pixels to be steered in high resolution, based on the presupposition that the peaks in the SRPD map are band-limited in space, resulting in similar values around them.

In a given iteration, t, the analysis grid (or equivalently the leaf nodes of the corresponding quadtree), $\mathcal{G}_t$ for a time-frequency bin $(\hat{\tau}, \hat{\varkappa})$ consists of the elements (i.e. nodes) $\mathcal{G}_t = \{\mathcal{S}_{K,m}^{(t)}\}$ potentially at different resolution levels, $K \in \{0, \cdots, t-1\}$. The total spatial entropy of the representation is given as:

$$H(\mathcal{G}_t) = -\sum_{\forall \mathcal{S}_{K,m}^{(t)} \in \mathcal{G}_t} \gamma(\mathcal{S}_{K,m}^{(t)}) \log \frac{\gamma(\mathcal{S}_{K,m}^{(t)})}{A(\mathcal{S}_{K,m}^{(t)})} \tag{3.20}$$

and

$$\gamma(\mathcal{S}^{(t)}_{K_{max},M}) = \frac{\mathcal{P}_{K,M}}{\sum_{\forall \mathcal{S}^{(t)}_{K,m} \in \mathcal{G}_t} \mathcal{P}_{K,m}}, \tag{3.21}$$

where $\mathcal{P}_{K,m}$ is the SRPD of the grid element [as in (3.13)] with index m at level K, and $A(\mathcal{S}^{(t)}_{K,m})$ is its area. Here, $0 \leq \gamma(\mathcal{S}^{(t)}_{K,m}) \leq 1$ is the probability of the grid element $\mathcal{S}^{(t)}_{K,m}$ of containing a source, and:

$$\sum_{\forall \mathcal{S}^{(t)}_{K,m} \in \mathcal{G}_t} \gamma(\mathcal{S}^{(t)}_{K,m}) = 1. \tag{3.22}$$

The decision to refine a specific grid element into four leaf nodes is given according to the total spatial entropy change. A candidate grid $\mathcal{G}'_t$ can be obtained from the existing grid $\mathcal{G}_t$ by refining a specific grid element, $\mathcal{S}^{(t)}_{K,m}$ such that:

$$\mathcal{G}'_t = \left[\mathcal{G}_t \cup \mathcal{V}(\mathcal{S}^{(t)}_{K,m})\right] \setminus \{\mathcal{S}^{(t)}_{K,m}\}. \tag{3.23}$$

where $\mathcal{V}(\mathcal{S}^{(t)}_{K,m}) = \{\mathcal{S}_{K+1,4m}, \mathcal{S}_{K+1,4m+1}, \mathcal{S}_{K+1,4m+2}, \mathcal{S}_{K+1,4m+3}\}$ is the set of children nodes in the quadtree representation of the analysis grid.

The proposed algorithm relies on the change in the total spatial entropy of the representation by using the difference between the total spatial entropy before and after the refinement, i.e. :

$$\mathcal{I}(l,m) = H(\mathcal{G}_t) - H(\mathcal{G}'_t). \tag{3.24}$$

This expression is known as *mutual information* or *information gain* in the context of decision trees [88]. When it is negative for a given element on the employed spherical grid as the spatial entropy increases with the children nodes insertion, the grid would not be refined for that locality, since such a refinement will be redundant despite increasing the computational cost. Therefore, if the refinement decreases the total spatial entropy $H(\mathcal{G}'_t)$ resulting that, $\mathcal{I}(l,m) > 0$, the representation is updated with the refined candidate grid, $\mathcal{G}'_t$. Otherwise, the corresponding branch of the quadtree will stay at its present resolution level, and it will be removed from the search path. Algorithm 1 presents the pseudocode for the HiGRID method.

Spatial entropy-based refinement progress of the proposed algorithm is given in Fig. 3.2 for different levels using the SHD coefficients from the sum of three unit amplitude, monochromatic plane waves with the common frequency of $f = 3$ kHz, incident on

Algorithm 1 Hierarchical Grid Refinement (HiGRID)

Input: SHD coefficients, a_{nm} for the time-frequency bin $(\hat{\tau}, \hat{\varkappa})$ selected via onset detection and the maximum refinement level, $K_{\max}$

Output: HEALPix quadtree with the set of leaf nodes, $\mathcal{G} = \{\mathcal{S}_{K,m},\ K = 0..K_{max}\}$ containing the SRPD evaluated at the corresponding pixel, $\mathcal{P}_{K,m}$ $treeLevel \leftarrow 0$ Initialize the quadtree by calculating the SRPD at each leaf node,$\mathcal{P}_{0,m}$

$N \leftarrow \emptyset$

$\mathcal{G} = \{\mathcal{S}_{0,m}\}$

while *treeLevel* $\leq K_{\max}$ **do**

  while $\mathcal{G} \neq \emptyset$ **do**

    $\mathcal{S} \leftarrow$fetchRandomLeafNodeFrom($\mathcal{G}$)

    $C \leftarrow$ childrenOf($\mathcal{S}$)

    Calculate SRPDs for $C = \{C_i, i = 1 \cdots 4\}$ $\mathcal{G}' = [\mathcal{G} \cup C_i] \setminus S$

    $E_{\mathcal{G}} \leftarrow spatialEntropy(\mathcal{G})$

    $E_{\mathcal{G}'} \leftarrow spatialEntropy(\mathcal{G}')$

    if $E_{\mathcal{G}'} < E_{\mathcal{G}}$ **then**

      $N \leftarrow N \cup C$

    else

      $N \leftarrow N \cup \{\mathcal{S}\}$

    end

    $\mathcal{G} \leftarrow \mathcal{G} \setminus \mathcal{S}$

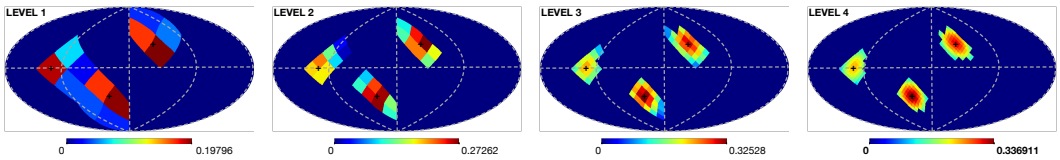
  end

  $\mathcal{G} \leftarrow N$,

  treeLevel $\leftarrow$ treeLevel + 1

end

a spherical aperture of radius 4.2 cm from the directions of $(\phi_1, \theta_1) = (\pi/2, 3\pi/5)$, $(\phi_2, \theta_2) = (2\pi/3, \pi/5)$, and $(\phi_3, \theta_3) = (\pi/3, 9\pi/5)$, respectively. The black cross signs denote the real DOAs of the plane waves. The decomposition at different resolution levels, from 1 through to 4, obtained using HiGRID are shown using the Mollweide projection. It may be observed that the SRPD map contracts in regions corresponding to the DOA of each plane wave. SRPD is calculated at a cumulative total of 33, 72, 165, and 429 directions for levels 1, 2, 3, and 4, respectively. The

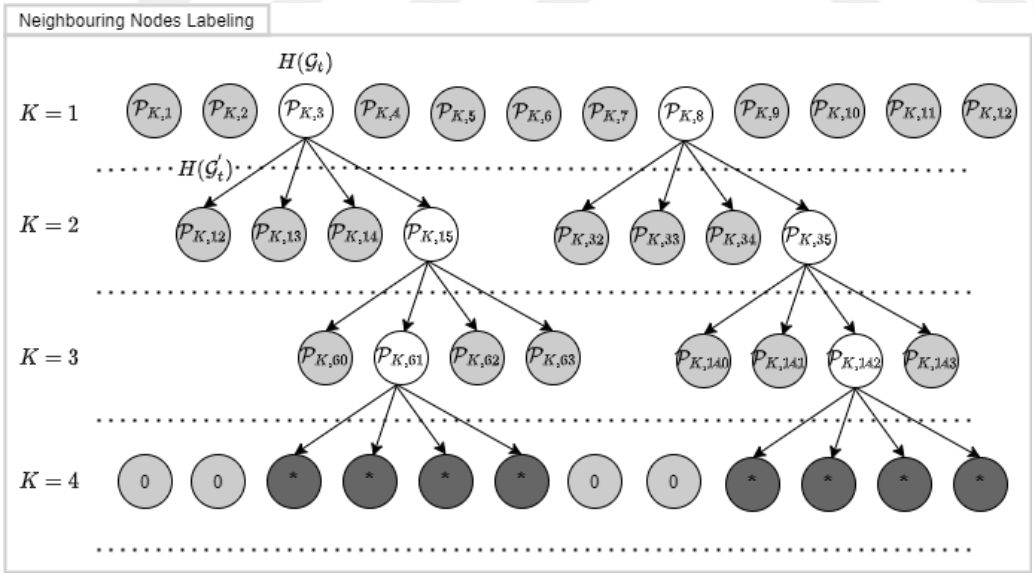


(a) (b) (c) (d)

Figure 3.2: Hiearchical grid refinement (HiGRID) applied on a single time-frequency bin, containing a sum of three unit amplitude, monochromatic plane waves with a frequency of $f = 3$ kHz. The progress of the algorithm at levels 1-4 are shown in (a)-(d), respectively. ©2018 IEEE [1]

same angular resolution can be achieved over a full-resolution grid by scanning 48, 192, 768, and 3072 directions at each of the corresponding levels.

Source Localization using SRPD in multiresolution : SPRD quadtree representation by the refinement of the grids to the deeper levels using mutual information results in a search map that consists of nodes to be visited to find the local maximum



(a)

Figure 3.3: If spatial entropy decreases as $H(\mathcal{G}_t) > H(\mathcal{G}'_t)$, the spatial resolution is incremented ($K = K + 1$) for that node.

points. *Connected Components Labeling* (CCL) [89] is adapted from the computer vision domain to use with the quadtree representation of the grid to identify and label regions of interest based on neighbourship relations between the nodes. From the DOA histogram randomly selected pixel with a magnitude greater than threshold was chosen as the cluster. Then, the other pixels were checked according to the neighbouring relations for the association of the previously defined cluster or creating a new cluster. The algorithm called as *Neighboring Nodes Labeling* (NNL) is shown in Algorithm 2.

While the regions having the local peaks are identified, the center pixels of each region, $\mathbf{n}(\hat{\tau}, \hat{\varkappa})$, is stored as the local DOA estimation for one of the sources in the time-frequency bin. The DOA unit vectors will not necessarily encounter the steering directions of the centers of HEALPix grid elements. Therefore, an average value is computed for the local DOA unit vectors.

Post-processing The post-processing stage uses a 2D direction histogram (with a uniform 1° bin size for elevation θ and azimuth ϕ) obtained by associating unit vectors that present source DOAs, $\{\mathbf{n}(\hat{\tau}, \hat{\varkappa})\}$ to the histogram map. Three operations are used over the global histogram to find the global DOA estimation:

1. The directional histogram may contain some outliers due to reverberation or noise. Indices with single or noisy occurrences are eliminated, and a median filter with a neighborhood size of 3 is applied to the resulting 2D histogram in order to denoise the 2D histogram from these TF bins.

2. The denoised 2D DOA histogram is convolved with a Gaussian filter with a kernel size of 3 to obtain a representation available for segmentation-based clustering.

3. The resulting DOA histogram is then mapped to a HEALPix grid to be clustered with the NNL method described earlier. Finally, the centroids of the identified clusters are registered as source DOAs, and the number of clusters indicates the source count in the acoustic scene.

Algorithm 2 Neighboring Nodes Labelling (NNL)

Input: Set of leaf nodes, $\mathcal{G} = \{\mathcal{S}_{K,m}\ l = 0..K_{max}\}$ of a HEALPix quadtree containing the values of SRPD distribution, $\mathcal{P}_{K,m}$

Output: Sets of nodes C_s where $s = 1 \cdots S$ `// Drop the regions via thresholding`

$THR \leftarrow$calculateThreshold($\{\mathcal{P}_{K,m}\}$)

for $\{\mathcal{S}_{K,m}\} \in \mathcal{G}$ **do**

  if $\mathcal{P}_{K,m} < THR$ **then**

    Remove node $\mathcal{S}_{K,m}$ from $\mathcal{G}$

  end

end

$s \leftarrow 1$ `// Initialise the label count`

$\mathcal{G}_{K_{max}} = \{\mathcal{S}_{K,m} \mid K = K_{\max}\}$

`// Select leaf nodes at the highest resolution level`

while $\mathcal{G}_{K_{max}} \neq \emptyset$ **do**

  stopFlag$\leftarrow$`True`

  $C_s \leftarrow \emptyset$

  `// Initialise cluster s`

  $\mathcal{S} \leftarrow$fetchRandomElementFrom($\mathcal{G}_{K_{max}}$)

  $C_s \leftarrow C_s \cup \{\mathcal{S}\}$

  $\mathcal{G}_{K_{max}} \leftarrow \mathcal{G}_{K_{max}} \setminus \{\mathcal{S}\}$

  while *stopFlag is* `True` **do**

    $N \leftarrow$neighborsOfSet(C_s)

    $D \leftarrow \mathcal{G}_{K_{max}} \cap N$

    if $D \neq \emptyset$ **then**

      $C_s \leftarrow C_s \cup D$

      $\mathcal{G}_{K_{max}} \leftarrow \mathcal{G}_{K_{max}} \setminus D$

    else

      $s \leftarrow s + 1$, stopFlag$\leftarrow$`False`

    end

  end

end

3.1.2 Residual Energy Test (RENT)

Acoustic scene can contain different sources having similar frequency components. Interference from the other sources and reverberation from the environment would result in overlapping of different source features for the same time-frequency bin. A pre-elimination step used prior to HiGRID [85] to decrease the computational cost is onset detection. There are also other methods in the literature to identify the time-frequency bins which have contributions from a single source only [31] [30]. A computationally efficient new time-frequency bin selection and DOA estimation approach is presented in this section. This method consists of two stages: (1) time-frequency bin selection and (2) clustering for DOA estimation of multiple sources. The proposed algorithm is a sparse recovery based algorithm using the Legendre kernels. Firstly, Legendre SRF kernel descriptions and then the involved steps of the algorithm are issued.

Plane Wave Legendre SRF Kernels : A plane wave's pressure distribution over the RSMA can be calculated using the monochromatic plane wave descriptions. Plane wave decomposition provides a flexible tool for analyzing sound fields. It can be expressed using real-valued functions defined on the unit sphere given as [90]:

$$\begin{aligned} \Gamma_N(\Omega_d, \Omega_m) &= \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\Omega_d) Y_n^m(\Omega_m)^* \\ &= \frac{2N+1}{4\pi} \frac{P_{N+1}(\cos\Theta_{d,m}) - P_N(\cos\Theta_{d,m})}{P_1(\cos\Theta_{d,m}) - P_0(\cos\Theta_{d,m})}, \end{aligned} \tag{3.25}$$

where $\Theta_{d,m}$ is the angle between the unit vectors in propagating wave direction Ω_d and steering direction Ω_m. Notice that $\Gamma_N(\Theta_{d,m})$ is rotationally symmetric with respect to the direction, $\Theta_{d,m}$ and that:

$$\lim_{N \to \infty} \Gamma_N(\Theta_{d,m}) = \frac{1}{2\pi} \delta(\cos\Theta_{d,m} - 1), \tag{3.26}$$

where $\delta(\Theta_{d,m})$ is the Dirac delta function on the sphere. Therefore, the function will be called $\Gamma_N(\Omega_d, \Omega_m)$ as a *Legendre kernel* of order N. The spherical sampling scheme used for determining steering directions is Hierarchical Equal Area isoLatitude Pixelation (HEALPix) of a sphere [86] with this thesis. The sphere is hierarchically split into curvilinear quadrilaterals. The lowest resolution partition is comprised

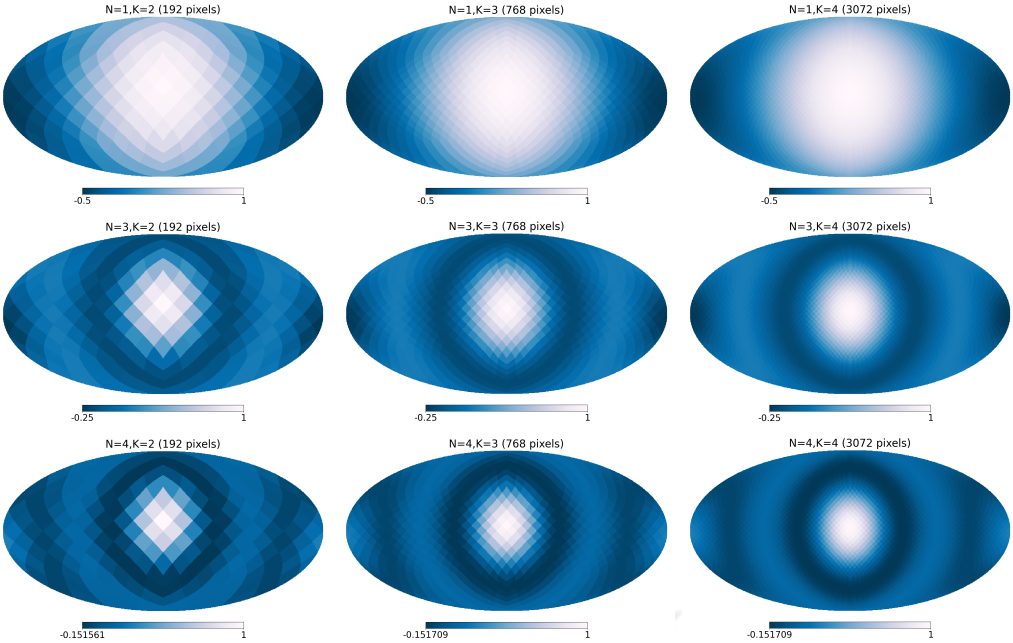


Figure 3.4: Legendre dictionary kernel in the elevation and azimuth direction $(90°, 0°)$ for spherical harmonic order of $N = 1$, $N = 3$, $N = 4$ and HEALPix level of $K = 2$, $K = 3$, $K = 4$, sampling in 192, 768 and 3072 points respectively.

of 12 base pixels. Resolution of the splitting increases by dividing each pixel into four new ones. This provides a hierarchy for sampling the sphere in different tessellation levels. The following properties are defined for the computation of overcomplete dictionaries.

- The dictionary has Legendre kernels centered at $\Omega_d = (\theta_d, \phi_d)$, and sampled at pixel centroids $\Omega_m = (\theta_m, \phi_m)$ of a HEALPix grid [86].
- The order of HEALPix grid at level K, where the number of pixels that completely cover the unit sphere at the same level is $M = 12 \times 2^{2K}$.
- Plane wave Legendre kernel dictionary is given in the equation (3.27) using D kernels. It is noted that M does not have to be equal to $D = 12 \times 2^{2K_l}$.

$$\Phi_{MD} = [\Gamma_N(\Theta_{1,d}), \Gamma_N(\Theta_{2,d}), ..., \Gamma_N(\Theta_{M,d})]^T_{\{d=1\cdots D\}}, \tag{3.27}$$

It is seen in Fig. 3.4 that higher order of spherical harmonics and a higher level of HEALPix grids provide sharper beams and higher resolution. Computational com-

plexity rises in the condition where the order and the level increase. Therefore, the algorithms' parameters have to be adjusted to obtain acceptable results with lower computational complexity and higher accuracy [given in Sec. 4.7.1, 4.7.2].

Notice that the algorithms proposed in this thesis assume that a plane wave decomposition is a feasible approach for representing general sound fields. The Legendre kernel formulation also relies on the plane wave assumption. If sources in the acoustic near field would need to be considered, the formulation would have been fundamentally different, possibly resulting in complex-valued dictionary atoms.

Representation of SRF with Legendre kernels : It may be noted that the steered response function given in (2.53) can be expressed as a linear combination of Legendre kernels such that:

$$y(\overline{\Omega}_m) = \sum_{d=1}^{D} \alpha_d \Gamma_N(\overline{\Omega}_d, \Omega_m). \tag{3.28}$$

Using the expression in (3.28), it is also possible to express the SRF $\mathbf{y} \in \mathbb{C}^{M\times 1}$ sampled at M discrete directions in matrix form as:

$$\mathbf{y} = \mathbf{L}_{MD}\mathbf{a}_D, \tag{3.29}$$

where $\mathbf{L}_{MD} \in \mathbb{R}^{M\times D}$ is a matrix with the rows given as $\mathbf{\Lambda}_m = [\Gamma_N(\overline{\Omega}_1, \Omega_m)), \Gamma_N(\overline{\Omega}_2, \Omega_m), \cdots \Gamma_N(\overline{\Omega}_D, \Omega_m)]$, and $\mathbf{a_D} \in \mathbb{C}^{D\times 1}$ is the column vector containing complex magnitudes. The DOA estimation problem involves the identification of S maxima values in $\mathbf{a}_D$ and the corresponding source directions Ω_d. It is also known that while D can potentially be very large in a reverberant environment, only a few complex coefficients, α_d, corresponding to sound sources will have large magnitudes.

A sparse approximation, $\mathbf{y} = \mathbf{L}_{MG}\mathbf{a}_G$ using a smaller number of G terms is possible. The optimal selection of G and $\{\Omega_g\}$ corresponds to the minimization of the approximation error:

$$\epsilon = \|\mathbf{y} - \mathbf{y}_G\|^2 + T^2 S, \tag{3.30}$$

where T is a Lagrange multiplier that limits the number of terms used in the approximation. While finding the best G-term representation is minimizing the residual error [91], satisfactory solutions can be obtained using non-convex optimization

algorithms. Orthogonal matching pursuit (OMP) [92] can be used to obtain such satisfactory solutions.

Residual energy test (RENT) is proposed as a time-frequency bin selection algorithm to identify the bins that contain only single sources. The algorithm is based on a single-iteration OMP used to select the time-frequency bins and estimate source DOAs. The function over the sphere to be represented in the sparse dictionaries is the SRF. The employed over-complete dictionary consists of Legendre kernels sampled at pixel centroids of a HEALPix grid [86] as their center directions.

The center direction of the Legendre kernels can be sampled from the HEALPix grid of a resolution level, K resulting in an over-complete dictionary consisting of the set of $D = 12 \times 2^{2K}$ atoms $\Phi_{\mathbf{MD}} = \{\mathbf{\Gamma}_d = [\Gamma(\overline{\Omega}_d, \Omega_1), \cdots, \Gamma(\overline{\Omega}_d, \Omega_M)]^T\}_{d=1\cdots D}$ steering in M directions. The dictionary composed of D items that do not depend on the frequency components except for the maximum order selection to present a general sparse dictionary that covers all frequency ranges given as in equation (3.25). Since the frequency decoupled SRF is used for the estimation by dividing it with the $b_n(kr)$ component, the over-complete dictionary has to be limited with the highest spherical harmonics order according to the frequency $kr_a = 2\pi f r_a/c$ [see in Fig. 2.4] by updating N as 1, 2, 3, and 4 for $kr_a < 1.0$, $kr_a < 2.0$, $kr_a < 3.0$, and $3.0 < kr_a$ respectively. Each time-frequency bin is processed via RENT which identifies the time-frequency bins predominantly with a single source contribution and RENT estimates the corresponding local DOAs using the steps to be described below.

1. Obtain the short-time Fourier transform $\{P_q(\tau, \varkappa)\}_{q=1\cdots Q}$ of microphone array signals, where τ is the time index and $\varkappa$ is the frequency index.

2. Obtain the SHD coefficients for each time-frequency bin, $(\tau, \varkappa)$, using (3.31)

$$\mathbf{a_{nm}} = \sum_{q=1}^{N} P_q(\tau, \varkappa)[Y_n^m(\theta_q, \phi_q)]^* w_q \tag{3.31}$$

3. For each time-frequency bin, calculate the SRF vector, $\mathbf{y}(\tau, \varkappa) \in \mathbb{R}^{M\times 1}$ at a finite number of M directions quasi-uniformly sampled on the HEALPix grid at the resolution level K.

4. Find the dictionary atom which best matches the SRF vector :

$$\mathbf{\Gamma}_d = \max_{\mathbf{\Gamma} \in \Phi} \langle \mathbf{\Gamma}, \mathbf{y}(\tau, \varkappa) \rangle \tag{3.32}$$

5. Calculate the residual error vector orthogonal to the SRF vector:

$$\begin{aligned} \mathbf{r}(\tau, \varkappa) &= \mathbf{R}_d \, \mathbf{y}(\tau, \varkappa) \\ &= \left[\mathbf{I} - \mathbf{\Gamma}_d \left(\mathbf{\Gamma}_d^T \mathbf{\Gamma}_d \right)^{-1} \mathbf{\Gamma}_d^T \right] \mathbf{y}(\tau, \varkappa). \end{aligned} \tag{3.33}$$

Note that $\mathbf{R}_d$ can be calculated offline for each dictionary atom in the dictionary and stored for increased computational efficiency.

6. Calculate ones' complement of the ratio between the residual energy and the total energy of the SRF vector, such that:

$$B(\tau, \varkappa) = 1 - \frac{\|\mathbf{r}(\tau, \varkappa)\|^2}{\|\mathbf{y}(\tau, \varkappa)\|^2}, \tag{3.34}$$

where $0 < B(\tau, \varkappa) < 1$. It is estimated that for a single plane wave in acoustic free field $\mathbf{y}(\tau, \varkappa)$ can be represented by a single atom and the energy of the residual vector would be zero resulting in $B(\tau, \varkappa) = 1$ indicating that $B(\tau, \varkappa)$ is close to unity when the time-frequency bin contains a single dominant source.

7. Identify the set of time-frequency bins for which (3.34) is greater than a selected threshold, such that:

$$\mathcal{S}_{\text{RENT}} = \{(\tau, \varkappa) : B(\tau, \varkappa) > THR_{SS}\} \tag{3.35}$$

8. Register the index of dictionary atom, d, identified for each time-frequency bin found in (3.32) as the DOA estimate for that time-frequency bin by incrementing the value of the pixel corresponding to $\{\Omega_d = (\theta_d, \phi_d)\}$. DOA histogram (Ξ) contains D pixel values that are initially set as zero.

After the local DOA estimates are obtained for each time-frequency bin that passes RENT, a global DOA histogram is created whose bin centers are aligned with pixel centroids of the employed HEALPix grid. This global map is used for the clustering and DOA estimates.

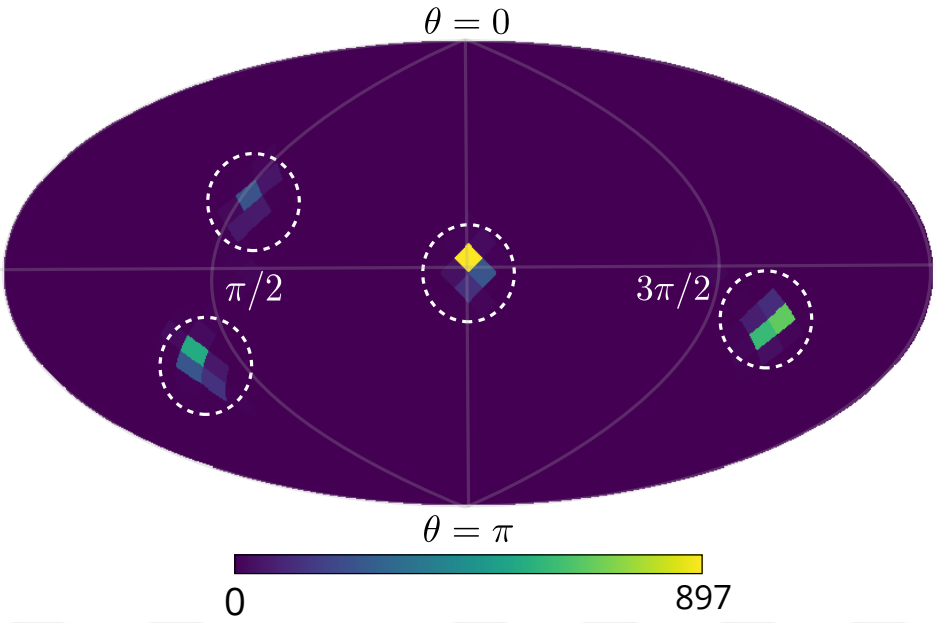


Figure 3.5: DOA histogram of four sound sources located at $(118°, 116°)$, $(68°, 90°)$, $(90°, 0°)$, and $(105°, 243°)$. Mollweide projection is used in the figure.©2019 IEEE [2]

Multiple DOA Estimation via Agglomerative Histogram Clustering : RENT is followed by clustering the resulting global histogram to obtain the DOA estimates. The primary assumption in the histogram clustering approach's development is that sources are spatially separated sufficiently well so that clusters can be identified based on their contiguity. Using the neighbouring information and count of the pixels, and first identifying the cluster centroids starting from the pixels having maximum counts (local maxima points), the connected regions can be separated well for some cases. While the proposed clustering approach is similar to the Neighboring Nodes Labeling (NNL) approach that we presented in [1], it is simple quick sorting for DOA map consisting D elements followed by condition checks for the neighbouring relations with a computational complexity of $O(DlogD) + O(D)$ to find local maxima points.

The global DOA histogram contains the number of occurrences (i.e., counts) that a given pixel on the HEALPix grid was identified as the DOA (see Fig. 3.5). The pixels are first sorted into a list according to frequency of occurence. The neighbouring relations of each pixel is also stored in a table. The pixel with the highest count is

Algorithm 3 Residual Energy Test (RENT)

Input: Over-complete dictionary Φ comprising D kernels in M steering directions, each atom's corresponding residual vector estimator $\mathbf{R}_d$, the recorded microphone signals $p_q(t)$ for $t \in [t_0, \cdots, t_0 + T]$, $\mathcal{S}_{\text{RENT}} = \emptyset$

Output: The number of sources $\mathbf{S} = |\Omega_s|$, corresponding angles $\mathbf{\Omega_s}$ and identified single source contributing time-frequency bins $\mathcal{S}_{\text{RENT}} = \{\hat{\tau}, \hat{\varkappa}\}$

$p_q(t) \rightarrow p_q(\tau, \varkappa)$ // STFT

for *each* $\tau, \varkappa$ **do**

  $p_q(\tau, \varkappa) \rightarrow \mathrm{a}_{nm}(\tau, \varkappa)$ // SHD

  CalculateSRF$(a_{nm}(\tau, \varkappa), M) \rightarrow y(\tau, \varkappa)$

  $\mathbf{\Gamma}_d = \max_{\mathbf{\Gamma} \in \Phi} \langle \mathbf{\Gamma}, \mathbf{y}(\tau, \varkappa) \rangle$ // Find best fit atom m

  $\mathbf{r}(\tau, \varkappa) = \mathbf{R}_d \, \mathbf{y}(\tau, \varkappa)$ // Calculate the residual vector

  CalculateRatio$(\mathbf{r}(\tau, \varkappa), \mathbf{y}(\tau, \varkappa)) \rightarrow B$

  if $B > THR_{SS}$ **then**

    $\mathcal{S}_{\text{RENT}} \cup (\tau, \varkappa) \rightarrow \mathcal{S}_{\text{RENT}}$

    RegisterIndexToHistogram$(d) \rightarrow \Xi$

  end

end

Sorting$(\Xi) \rightarrow \hat{\Xi}$ // Sort the histogram from max.to min. and note the pixel locations

DOAHistogramClustering$(\hat{\Xi}) \rightarrow \mathbf{S}, \mathbf{\Omega_s}$// Source count and DOA est.

{

$\mathbf{\Omega_s}, C = \emptyset$// C indice, $\wp(C)$ is the neighbouring list

Create a cluster for the max. item's indice $\mathbf{\Omega_1} \rightarrow \mathbf{\Omega_s}, C_1 \cup \wp(C_1) \rightarrow C$

for *each* $i \in \{2, |\hat{\Xi}|\}$ **do**

  if $C_i \notin C$ **then**

    $\mathbf{\Omega_i} \cup \mathbf{\Omega_s} \rightarrow \mathbf{\Omega_s}$

  end

  $C_i \cup \wp(C_i) \cup C \rightarrow C$

end

}

removed from the sorted list of selected pixels and is defined as the first cluster. The neighborhood of the cluster is defined as the set of pixels surrounding it. The next pixel in the sorted list is selected and removed from the list. If the pixel is in the neighborhood of the first cluster, it is added to the cluster, and the neighborhood of the cluster is expanded to account for the new pixel. If it is not in the neighborhood, a new cluster is formed. For each pixel from the sorted list, the neighborhoods of all existing clusters are checked, and if the new pixel is not in one of the existing clusters or their neighborhoods, a new cluster is formed. The iteration continues until the list of selected pixels is exhausted. The number of identified clusters and cluster centroids provide both source count and DOAs, respectively. Thresholding with the mean value of the DOA histogram would also eliminate indices having smaller DOA estimate counts occuring from the reverberation and diffuse field.

The RENT's computation is carried out in longer, overlapping intervals as 0.5 s, and the resulting observed source angles in each time interval would be associated with each other. RENT is used as a DOA estimator within the same framework in source separation as will be discussed in the next sections. The performance measurements of RENT in several noisy measurements and controlled setups are also investigated in Chapter 4. The pseudocode is given in Algorithm 3. The computational cost of the algorithm is also presented in Sec. 4.7.2.

3.2 Source Separation Techniques

Object-based audio consists of channel-based audio recordings for distributing the format to the audio renderers with the metadata. While scene-based audio format as higher order ambisonics does not contain separated audio files, array signal processing techniques are required for the acoustic scene decomposition. In this section, the proposed source separation methods are investigated, and they are sparse plane wave decomposition using fixed and adaptive spatial filtering, and Wiener post-filtering.

3.2.1 Source Separation with Sparse Plane Wave Decomposition

The combination of direct and reverberant fields can be expressed as a sum of multiple plane waves, using Legendre kernel, $\Gamma_N(\Omega_d, \Omega_m)$, described earlier such that:

$$y(\Omega_m, k) = \sum_{d=1}^{\infty} \alpha_d(k)\Gamma_N(\Omega_d, \Omega_m), \tag{3.36}$$

where Ω_d corresponds to the incidence direction of the plane wave d and Ω_m represents the steering direction of SRF, the direction dependent term is given in the eq. 2.17 with $\cos(\Theta)$ is replaced $\cos(\Theta_{d,m}) = \cos(\theta_m)\cos(\theta_d) + (\cos(\phi_m) - \cos(\phi_d))$ $\sin(\theta_m)\sin(\theta_d)$. The plane wave definition given for the single plane wave can be shown according to the spherical harmonics property [93] that this function is a spatially bandlimited pulse with its maximum at $\Theta_{d,m} = 0$, (i.e. $\Omega_d = \Omega_m$) is given in eq. 3.25.

The distribution of the Legendre kernel according to the spherical harmonics order limitations (N) over the array may be observed from Fig. 2.2.

An overcomplete dictionary $\Phi_{MD} = \{\Gamma(\Omega_d|\Omega_m)\}_{d\in D}$ comprising a finite number, D, of vectors in $\Re^{M\times 1}$ is used for obtaining the sparse plane wave decomposition (SPWD). Each vector contains Legendre kernel samples centered at Ω_d, and calculated at M steering directions Ω_m , on the unit sphere. Assuming that a good approximation of the SRF can be obtained using a finite number of D terms, for which $\alpha_d(k)$ will be non-zero, a sparse approximation of the steered response functional in (3.36) can be obtained such that:

$$y(\Omega_m, k) = \sum_{d=1}^{D} \alpha_d(k)\Gamma(\Omega_d|\Omega_m). \tag{3.37}$$

The above equation can be expressed in matrix form as:

$$\mathbf{y} = \Phi_{\mathbf{MD}}\mathbf{a}, \tag{3.38}$$

where Φ_{MD} is the $M \times D$ matrix containing D dictionary kernels steering in M locations with the corresponding column $\Gamma_d = [\Gamma(\Omega_d|\Omega_1)\cdots\Gamma(\Omega_d|\Omega_M)]^T$ calculated from the plane wave amplitudes, and $\mathbf{a}$ is a $D \times 1$ vector containing the complex amplitudes corresponding to each dictionary atom. The optimal selection of atoms

and weights corresponds to the minimization of the approximation error, using a finite number, G, of terms such that:

$$\epsilon = \|\mathbf{y} - \mathbf{y}_G\|^2 + T^2 S \tag{3.39}$$

where T is a Lagrange multiplier penalizing the number of terms used in the approximation thereby promoting sparsity. The main objective is to minimize the residual error ϵ via selecting the minimum number of best fitting kernels to the SRF vector.

SPWD also uses OMP over the same kernel dictionary Φ_{MD} as used by RENT. If the dictionary atoms used for OMP are not a perfect orthonormal basis for the SRF, convergence can not be achieved in a finite number of steps. The number of atoms D in the dictionary define the rate and speed of convergence [94]. Therefore, decisions to either limit the residual power ratio or the number of iterations should be taken instead of aiming for the perfect convergence with many dictionary atoms. This decision lowers the computational cost required for the operations. The residual error is initialized with the original SRF vector to be approximated (i.e. $\mathbf{R}_0 = \mathbf{y}$) and each iteration consists of the selection of an atom from the kernel dictionary and the calculation of the residual vector such that:

$$\mathbf{R}_{i+1} = \left[\mathbf{I} - \mathbf{\Gamma}_i \left(\mathbf{\Gamma}_i^T \mathbf{\Gamma}_i\right)^{-1} \mathbf{\Gamma}_i^T\right] \mathbf{R}_i \tag{3.40}$$

with $\mathbf{\Gamma}_i = [\mathbf{\Gamma}_{i-1} \,|\, \Phi_i]$ where:

$$\Phi_i = \underset{\Phi_d \in \Phi_{MD}}{\arg\max} |\Phi_d^T \mathbf{R}_i|, \tag{3.41}$$

is the atom which provides the best fit to the residual error at the current iteration. OMP can be terminated either when the residual norm is below a prescribed threshold or after a predefined number of iterations. Upon the termination of OMP, the complex magnitude vector, $\mathbf{a}$ is obtained as:

$$\mathbf{a}_i = \left(\mathbf{\Gamma}_i^T \mathbf{\Gamma}_i\right)^{-1} \mathbf{\Gamma}_i^T \mathbf{y} = \mathbf{\Gamma}_i^{\dagger} \mathbf{y}, \tag{3.42}$$

where $\mathbf{\Gamma}_i^{\dagger}$ is the pseudoinverse of the corresponding atom, $\mathbf{a}_i$ is the complex magnitude for the corresponding atom i and $\left[\mathbf{I} - \mathbf{\Gamma}_i \mathbf{\Gamma}_i^{\dagger}\right]$ is the residual vector estimator (orthogonal projector on to the kernel $\mathbf{\Gamma}_i^T$) for the atom i from the residual $\mathbf{R}_i$. Since each atom in the selected dictionary matrix corresponds to a specific look direction

Ω_d, it becomes possible to reconstruct a spatially weighted SRF in the desired steering direction, Ω_m, and obtain a virtual beam in that direction with a higher level of interference rejection than that of maximum directivity factor beamforming, such that:

$$x_s(k) = \overline{\Phi}_s \mathbf{W}\mathbf{a}, \tag{3.43}$$

where $\overline{\Phi}_s = [\Gamma(\Omega_s|\Omega_1) \cdots \Gamma(\Omega_s|\Omega_M)]$ is the dictionary item corresponding to the identified source s and its DOA Ω_s and

$$\mathbf{W_d} = [w(\Omega_1|\Omega_d) \cdots w(\Omega_M|\Omega_d)]^T, \tag{3.44}$$

is a real valued matrix of non-negative weights. While it could be possible to select a non-negative window that allows controlling the beamwidth such as the Dolph-Chebyshev window [95], symmetric von Mises function is used for fixed spatial filtering (FSF) to collect the magnitudes in neighboring pixels [96] as:

$$w(\Omega_i|\Omega_j) = \frac{e^{\kappa \cos \Omega_\Delta}}{2\pi I_0(\kappa)}, \tag{3.45}$$

where κ is the concentration parameter, Ω_Δ is the angular separation between the directions Ω_i and Ω_j and $I_0(\cdot)$ is the modified Bessel function of order 0. Note that, if $\kappa = 0$, $\mathbf{W}$ would be an identity matrix, and the result obtained via (3.43) would be equivalent to the output of a summation of maximally directive beamformer for the corresponding pixels.

For the simplest case where a target and interference in the acoustic free field are separated by Ω_Δ, the maximum level of interference suppression, $D(\Omega_\Delta)$, above what is attainable by maximally directive beamforming can be expressed as:

$$D(\Omega_\Delta) = 10 \log_{10} \left(e^{\kappa(\cos \Omega_\Delta - 1)} \right) \tag{3.46}$$

Fig. 3.6 shows $D(\Omega_\Delta)$ for different values of κ as well as the maximally directive beam for $N = 4$. It may be observed that the suppression provided by the employed spatial weighting function is higher than that of a maximally directive beam. It was observed during simulations that selecting a high value for κ may result in undesired artifacts, and a value of $\kappa = 10$ works well in separation performance by creating narrower spatial filters. The effects of κ over the distribution and the separation performance are discussed in Chapter 4 and Appendix A. Since the version of

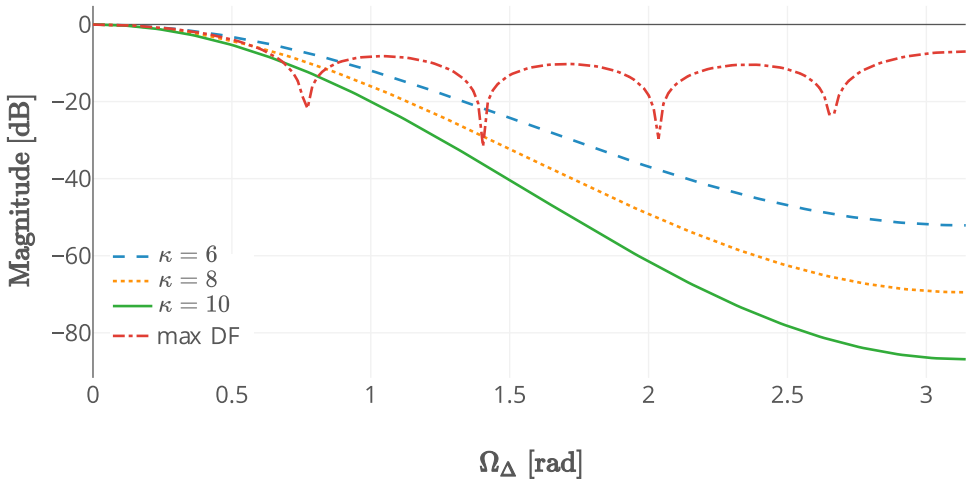


Figure 3.6: Suppression of a single interference that is separated by Ω_Δ from the target afforded by the von Mises function for different values of κ. Maximum directivity factor beam for $N = 4$ is also shown for comparison.©2018 IEEE [4]

the method, thus described uses a fixed spatial filter, we will henceforth call it sparse plane wave decomposition with fixed spatial filtering (SPWD-FSF). The pseudocode for the algorithm is also shown in Algorithm 4.

3.2.2 Sparse Plane Wave Decomposition with Adaptive Spatial Filtering

The spread of source distributions varies according to the distance and direct to reverberant ratio associated with a source, thus using a non-adaptive fixed spatial filter as described would reduce the source separation performance. Therefore, an adaptive spatial filtering design is proposed to solve this problem.

Analysis of the acoustic scene in small time intervals and calculating the DOA estimates via observing bins that have single dominant components would improve the source separation algorithm performance in the defined conditions. RENT algorithm is used to estimate DOA, source count, and directional source distribution over the analysis interval prior to clustering. The obtained source DOA distributions are then used to adapt the spatial filters used in virtual beamforming. The proposed method picks up the source count of the best fitting atoms for the identified frequencies (single source bins). These source count parameters in the global DOA map are used to

generate a Kent probability distribution function. This is done by fitting Kent distributions to clustered DOA distributions via maximum likelihood estimation (MLE). This results in spatial filters that are customized to individual sources.

Algorithm 4 Sparse Plane Wave Decomposition - Fixed Spatial Filtering (SPWD-FSF)

Input: Over-complete dictionary Φ with D atoms in M steering directions, each atom's pseudo inverse matrix $\mathbf{\Gamma}_d^\dagger$ and residual vector estimator $\left[\mathbf{I} - \mathbf{\Gamma}_d \mathbf{\Gamma}_d^\dagger\right]$, the recorded microphone signals $p_q(t)$, the number of sources $\mathbf{S}$ and their corresponding DOAs $\mathbf{\Omega_s}$ for $t \in [t_0, \cdots, t_0 + T]$. The residual vector computation is limited to L_{max} iterations.

Output: Separated sources $x_s(t)$

$p_q(t) \rightarrow p_q(\tau, \varkappa)$ // STFT

FixedSpatialFilters($\mathbf{S}, \Omega_s$)$\rightarrow \mathbf{W}$// von Mises Filters

for *each* $\tau, \varkappa$ **do**

  $p_q(\tau, \varkappa) \rightarrow a_{nm}(\tau, \varkappa)$ // SHD

  CalculateSRF$(a_{nm}(\tau, \varkappa), M) \rightarrow y(\tau, \varkappa)$

  $y(\tau, \varkappa) \rightarrow \mathbf{R}(\tau, \varkappa)$

  $i = 0$

  while *i* < L_{max} **do**

    $\mathbf{\Gamma}_i = \max_{\mathbf{\Gamma} \in \Phi} \langle \mathbf{\Gamma}, \mathbf{R}(\tau, \varkappa) \rangle$ // Find best fit atom i

    $\mathbf{a}_i = \mathbf{\Gamma}_i^\dagger \mathbf{R}(\tau, \varkappa)$ // Calculate the magnitude

    $\mathbf{R}(\tau, \varkappa) = \left[\mathbf{I} - \mathbf{\Gamma}_i \mathbf{\Gamma}_i^\dagger\right] \mathbf{R}(\tau, k)$ // Calculate the residual

    CalculateRatio$(\mathbf{R}(\tau, \varkappa), \mathbf{y}(\tau, \varkappa)) \rightarrow B$

    if $B < THR_{SPWD}$ **then**

      break;

    end

    $i = i + 1$

  end

  $\overline{\Phi}_s \mathbf{W} \mathbf{a} \rightarrow x_s(\tau, \varkappa)$

end

$x_s(\tau, \varkappa) \rightarrow x_s(t)$// ISTFT

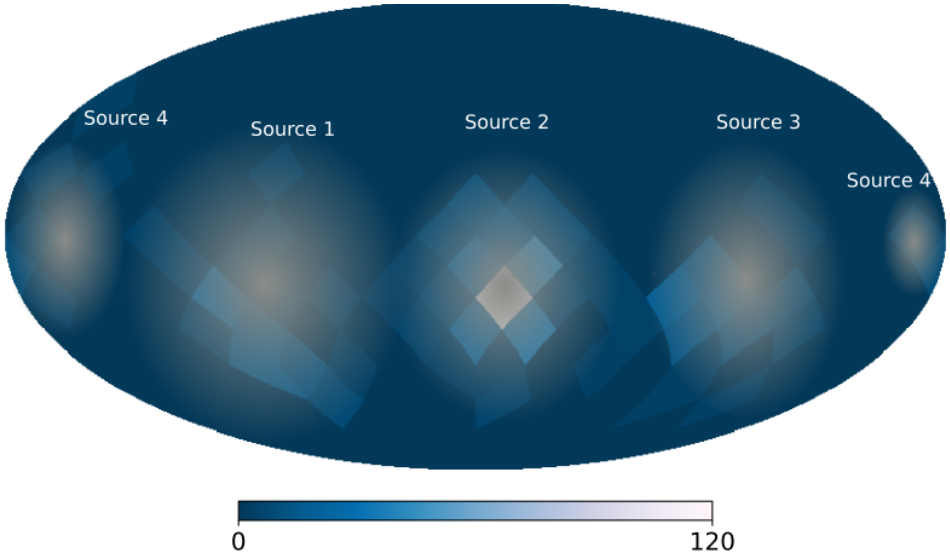


Figure 3.7: Output of RENT for the sources having different distribution characteristics for sources $(100°, 131°)$, $(120°, 0°)$, $(120°, 270°)$,and $(100°, 180°)$ in elevation and azimuth.

As each atom in the selected over-complete dictionary corresponds to a specific look direction Ω_d, spatially weighting a sparse version of SRF in the desired steering directions would result in good source separation performances. The main objective is to obtain a virtual beam in the direction of the source whose DOA is known, with a level of interference rejection that is higher than maximum directivity beamforming. The separation formula after the application of OMP iteratively is given as the equation (3.43). The objective is to replace spatially weighting $\mathbf{W}$ defined as a fixed von Mises function with the adaptive spatial filter. Creating a consistency between the spatial filter and source DOA distributions improve the separation performances to suppress the interference in the cases where sources are approaching to each other.

The separated source signals' quality largely depends on the selected spatial weighting function, especially in terms of interference and reverberation [4]. When a window function with a narrow beamwidth is used for this purpose, sources that are closer to the recording position are separated well, but the distant sources will have artifacts. In contrast, when a window with wide beamwidth is used, source separation performance degrades for closer sources.

The window selection approach we propose is based on the observation that DOA distributions obtained with RENT do not only change for different individual source D/R levels, but are also rotationally non-symmetric. Distributions on the unit sphere without rotational symmetry can be modeled using the Kent distribution [97]. Defining an arbitrary point $\mathbf{x}$ on the unit sphere, the probability density function of the Kent distribution is given as:

$$K(\mathbf{x}|\boldsymbol{\xi}) = \frac{1}{c(\kappa,\beta)}\exp\{\kappa\gamma_1^T\mathbf{x} + \beta[(\gamma_2^T\mathbf{x})^2 - (\gamma_3^T\mathbf{x})^2]\}, \tag{3.47}$$

with $\boldsymbol{\xi} = \{\kappa, \beta, \gamma_1, \gamma_2, \gamma_3\}$ where the parameter κ represents the spread, β represents the ellipticity of the distribution. γ_1 represents the central tendency, γ_2 and γ_3 represent the principal axes of the distribution, respectively, and $c(\kappa,\beta)$ is a normalizing term given as:

$$c(\kappa,\beta) = 2\pi\sum_{j=0}^{\infty}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\beta^{2j}(\frac{1}{2}\kappa)^{-2j-\frac{1}{2}}I_{2j+\frac{1}{2}}(\kappa). \tag{3.48}$$

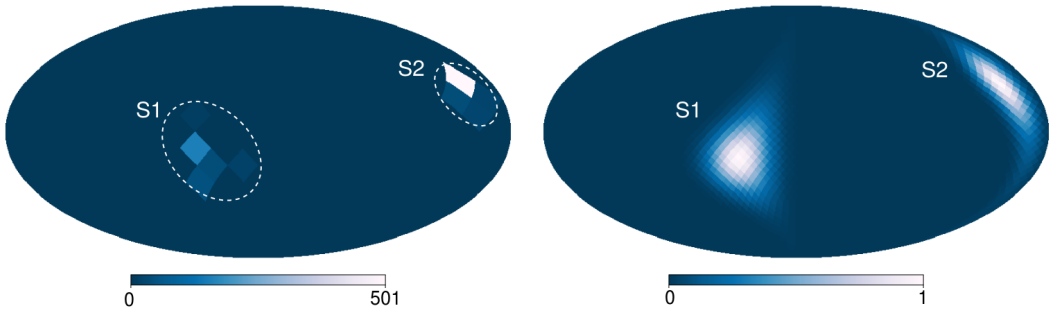
Here, $I_s(\kappa)$ denotes the Bessel function of the first kind, and $\Gamma(.)$ is the gamma function. Kent distribution would result in non-symmetric spatial filters whose spread over the sphere can be controlled via the parameter κ.

For each analysis interval, $\Upsilon_c = [t_0, \cdots, t_0+T]$ and each cluster representing source s, a single Kent distribution, $K_{c,s}(\mathbf{x}|\xi_s)$ is obtained using MLE [98]. The distributions obtained this way are non-normalized and may result in sharp spatial filters which are used as weights, result in audible artifacts. These problems can be mitigated by uniformly dilating the distribution around its mean by multiplying its spread parameter (κ) by a constant ϱ, and then normalizing the resulting distribution such that:

$$K_{c,s}(\mathbf{x}) = \frac{c(\varrho\kappa,\beta)}{\exp(\varrho\kappa)}K(\mathbf{x}|\overline{\boldsymbol{\xi}}_m), \tag{3.49}$$

where $\overline{\boldsymbol{\xi}}_s = \{\varrho\kappa_s, \beta_s, \gamma_{s,1}, \gamma_{s,2}\}$ and ϱ is a dilation factor generating narrower filter while it is greater than 1 (see in Fig. A.4). This distribution is then sampled on the employed spherical grid to obtain $\mathbf{W}_{c,s} = \text{diag}\{K_{c,s}(\mathbf{x}_s)\}_{s=1\cdots S}$ where $\mathbf{x}_s = [\cos\phi_s \sin\theta_s\,, \sin\phi_s \sin\theta_s\,, \cos\theta_s]^T$ is the unit vector in Cartesian coordinates. Fig. 3.8 shows DOA distributions and the spatial weighting functions calculated for two sources.

It should be noted that while OMP is applied on all time-frequency bins, the spatial weighting matrix $\mathbf{W}_{c,s}$ is updated only once for the corresponding analysis interval,



(a) (b)

Figure 3.8: a) DOA distribution for two acoustic sources obtained using RENT, b) Kent distributions fitted to data after clustering with CCL.

Υ_c. If a previously identified source is not active in a given time interval, a previously fitted Kent distribution for the respective source is used for spatial filtering in order to maintain continuity. Each interval is analyzed independently to find the source parameters, and sources obtained in each interval are matched to each other if their DOA estimates are less than Ψ between two consecutive intervals. This method is a simple tracking to capture small movements of the acoustic source around the sphere and provide continuity. The algorithm can also be adapted to moving sources, but this problem is beyond the scope of this thesis. The effects of adaptive spatial filtering with respect to fixed spatial filtering are observed in Chapter 4 and also controlled setups in Appendix A in detail. The method is also described in pseudocode form in Algorithm 5.

3.2.3 Wiener post-filtering with the Residual Energy Test

Wiener post-filtering [79] [61] can be used to improve the performance of source separation and to mitigate the degrading effects of diffuse field and sensor noise. The proposed technique groups the time-frequency bins into three as noise/diffuse, single source and multiple sources using RENT. The SRF was given as $\mathbf{y}(\tau, \varkappa) \in \mathbb{C}^{M\times 1}$ at a finite number of M directions quasi-uniformly sampled on the HEALPix grid at the resolution level K. Time-frequency bins for which there is no dominant source are used for calculating the noise covariance matrix employed in Wiener post-

Algorithm 5 Sparse Plane Wave Decomposition with Adaptive Spatial Filtering (SPWD-ASF)

Input: Over-complete dictionary Φ with D atoms in M steering directions, each atom's pseudo inverse matrix $\mathbf{\Gamma}_d^{\dagger}$ and residual vector estimator $\left[\mathbf{I}-\mathbf{\Gamma}_d\mathbf{\Gamma}_d^{\dagger}\right]$, the recorded microphone signals $p_q(t)$, the number of sources $\mathbf{S}$ and their corresponding DOAs $\mathbf{\Omega_s}$ for $t \in [t_0, \cdots, t_0+T]$. The residual vector computation is limited to L_{max} iterations.Ψ is the angle difference parameter for associating the current DOA estimates to the previously identified ones.

Output: Separated sources $x_s(t)$. `// Computation over each interval`

for *each* Υ_c **do**

  $RENT \rightarrow S, \mathbf{\Omega_{s,c}}, \mathbf{x}_s, y(\tau,\varkappa)$ `// Algorithm 3`

  AssociateDOA($\mathbf{\Omega_{s,c}}, \mathbf{\Omega_{s,c-1}}, \Psi$)

  $\mathbf{x}_s \xrightarrow{\text{MLE}} \mathbf{W}_{c,s}$

  for *each* $\tau, \varkappa$ **do**

    $y(\tau,\varkappa) \rightarrow \mathbf{R}(\tau,\varkappa)$

    i = 0

    while *i < L_{max}* **do**

      $\mathbf{\Gamma}_i = \max_{\mathbf{\Gamma}\in\Phi}\langle \mathbf{\Gamma}, \mathbf{R}(\tau,\varkappa)\rangle$ `// Find best fit atom`

      $\mathbf{a}_i = \mathbf{\Gamma}_i^{\dagger}\mathbf{R}(\tau,\varkappa)$ `// Calculate the magnitude`

      $\mathbf{R}(\tau,\varkappa) = \left[\mathbf{I}-\mathbf{\Gamma}_i\mathbf{\Gamma}_i^{\dagger}\right]\mathbf{R}(\tau,\varkappa)$ `// Calculate the residual`

      CalculateRatio$(\mathbf{R}(\tau,\varkappa), \mathbf{y}(\tau,\varkappa)) \rightarrow B$

      if $B < THR_{SPWD}$ **then**

        break;

      end

      i = i + 1

    end

    $\overline{\Phi}_s\mathbf{W}\mathbf{a} \rightarrow x_s(\tau,\varkappa)$

  end

end

$x_s(\tau,\varkappa) \rightarrow x_s(t)$ `// ISTFT`

filtering. The modified version of RENT to identify the noise/diffuse field is described in pseudocode form in Algorithm 6.

In the ideal case, time-frequency bins that predominantly contain diffuse energy or noise have SRF vectors for which energy is distributed uniformly in all directions. These time-frequency bins are selected as:

$$\aleph_{\text{RENT}} = \{(\hat{\tau}, \hat{\varkappa}) : THR_{df} > B(\tau, \varkappa)\}, \tag{3.50}$$

where $B(\tau, \varkappa)$ is calculated in equation (3.34). Note that it is estimated that these bins $\aleph_{\text{RENT}}$ comprise the noise term $\eta(\tau, \varkappa)$ in the signal definition $\mathbf{y} = \mathbf{La} + \eta$. The noise covariance matrix $\mathbf{\Phi}_\eta = E\{\eta\eta^H\}$ is obtained by averaging the SRF vector of the identified diffuse field time-frequency bins.

$$\mathbf{\Phi}_\eta = \frac{1}{\mathbf{card}(\aleph_{\text{RENT}})} \sum_{(\tau,\varkappa)\in\aleph_{\text{RENT}}} \mathbf{y}(\tau, \varkappa)\mathbf{y}(\tau, \varkappa)^H, \tag{3.51}$$

where $\mathbf{card}(.)$ is the cardinality of the set and noise covariance matrix is updated in each time interval according to RENT update. The covariance matrix for the SRF vector used in the computation of Wiener post-filter is given in the equation (3.52).

$$\begin{aligned} \mathbf{\Phi_y} &= E\{\mathbf{yy^H}\} = E\{(\mathbf{La} + \eta)(\mathbf{La} + \eta)^\mathbf{H}\} \\ &= \mathbf{L}E\{\mathbf{aa^H}\}\mathbf{L^H} + E\{\eta\eta^\mathbf{H}\} \\ &= \mathbf{L\Phi_a L^H} + \mathbf{\Phi}_\eta, \end{aligned} \tag{3.52}$$

where Φ_a is the source covariance matrix with $S \times S$ elements at the given frequency bin, and $\mathbf{L}$ is the identified Legendre kernel matrix with $M \times S$ elements. The least squares solution outputs the corresponding source cross covariance matrix given as:

$$\mathbf{\Phi_a} \approx \mathbf{L}^\dagger(\mathbf{\Phi_y} - \mathbf{\Phi}_\eta)(\mathbf{L}^\dagger)^H, \tag{3.53}$$

where $\mathbf{L}^\dagger = (\mathbf{L^H L})^{-1}\mathbf{L^H}$ is the pseudoinverse of the DOA associated Legendre kernels. Estimates of power spectral densities corresponding to each source can be obtained directly from the terms on the diagonal of the signal covariance matrix. Wiener post-filtering applied on the source, s, involves the application of the following time-frequency mask:

$$\widetilde{x}_s(\tau, \varkappa) = x_s(\tau, \varkappa)\frac{\mathbf{\Phi}_{a,ss}}{tr(\mathbf{\Phi}_a) + tr(\mathbf{\Phi}_\eta)}, \tag{3.54}$$

Algorithm 6 Modified Version of Residual Energy Test (RENT)

Input: Over-complete dictionary Φ comprising D kernels in M steering directions, each atom's corresponding residual vector estimator $\mathbf{R}_d$, the recorded microphone signals $p_q(t)$ for $t \in [t_0, \cdots, t_0 + T]$, $\mathcal{S}_{\text{RENT}} = \emptyset$, $\aleph_{\text{RENT}} = \emptyset$, $\mathcal{V}_{\text{RENT}} = \forall(\tau, \varkappa)$

Output: The number of sources $\mathbf{S} = |\Omega_s|$, corresponding angles $\mathbf{\Omega_s}$ and identified single source contributing time-frequency bins $\mathcal{S}_{\text{RENT}} = \{\hat{\tau}, \hat{\varkappa}\}$, noise/diffuse field bins $\aleph_{\text{RENT}} = \{\hat{\tau}, \hat{\varkappa}\}$, valid source bins $\mathcal{V}_{\text{RENT}} = \{\hat{\tau}, \hat{\varkappa}\}$, and multiple source bins $\mathcal{V}_{\text{RENT}} \setminus \mathcal{S}_{\text{RENT}}$. $\mathbf{\Phi}_\eta$ is also the noise PSD matrix.

$p_q(t) \rightarrow p_q(\tau, \varkappa)$ // STFT

for *each* $\tau, \varkappa$ **do**

  $p_q(\tau, \varkappa) \rightarrow \mathrm{a}_{nm}(\tau, \varkappa)$ // SHD

  CalculateSRF$(a_{nm}(\tau, \varkappa), M) \rightarrow y(\tau, \varkappa)$

  $\mathbf{\Gamma}_d = \max_{\mathbf{\Gamma} \in \mathbf{\Phi}} \langle \mathbf{\Gamma}, \mathbf{y}(\tau, \varkappa) \rangle$ // Find best fit atom d

  $\mathbf{r}(\tau, \varkappa) = \mathbf{R}_d \, \mathbf{y}(\tau, k)$ // Calculate the residual vector

  CalculateRatio$(\mathbf{r}(\tau, \varkappa), \mathbf{y}(\tau, \varkappa)) \rightarrow B$

  if $B > THR_{ss}$ **then**

    $\mathcal{S}_{\text{RENT}} \cup (\tau, \varkappa) \rightarrow \mathcal{S}_{\text{RENT}}$

    RegisterIndexToHistogram$(d) \rightarrow \Xi$

  end

  if $THR_{df} > B$ **then**

    $\aleph_{\text{RENT}} \cup (\tau, \varkappa) \rightarrow \aleph_{\text{RENT}}$

    $\mathcal{V}_{\text{RENT}} \setminus (\tau, \varkappa) \rightarrow \mathcal{V}_{\text{RENT}}$

    $\Phi_\eta(\tau, \varkappa) = E\{y(\tau, \varkappa) y(\tau, \varkappa)^H\}$

  end

end

Averaging $\Phi_\eta, \forall(\tau, \varkappa) \in \aleph_{\text{RENT}} \rightarrow \mathbf{\Phi}_\eta$// Noise covariance matrix

Sorting$(\Xi) \rightarrow \hat{\Xi}$ // Sort the histogram from maximum to minimum and note the pixel locations

DOAHistogramClustering$(\hat{\Xi}) \rightarrow \mathbf{S}, \mathbf{\Omega_s}, \mathbf{x_s}$// Source count, DOA estimation and source distribution

where $\Phi_{a,ss}$ is the s^{th} diagonal element of the signal covariance matrix and tr(.) is the matrix trace operator. The post-filtering approach comprises a weighting mask for the

resulting source separation algorithms such as MaxDF, SPWD-FSF and SPWD-ASF, and the joint source DOA estimation and separation will be issued in the next section.

3.3 Joint Source DOA Estimation and Separation

The proposed technique has three different blocks for the DOA estimation and separation. The blocks working for the current time interval are (i) finding the time-frequency bins which have single-source contributions, or noise/diffuse field contributions using RENT, (ii) source separation using SPWD-based methods as described (see in 3.2.1, 3.2.2) and (iii) weighting of time-frequency bins' complex-valued source coefficients with the PSD weighting matrix estimated from noise/diffuse field bins identified via RENT.

The first step to follow is conversion of microphone signals from the time domain into time-frequency using short-time Fourier transform (STFT) with a block length of $\beth$ and an overlap of $\beth_0$. These sliding window outputs are used as algorithm input blocks iteratively. Dictionary (Φ) comprising D Legendre kernels is calculated in M steering directions for the HEALPix level of K . The number of steering directions M on the sphere is also tuned as the length for the overcomplete dictionary kernel by aiming the best performance. The algorithm operates on each frame independently except for associating the DOA estimates between the consecutive frames. The main components of the proposed method are described below:

(i) The operations over RENT : The distribution of DOA estimates, $\mathbf{x}_{c,s}$, is obtained by aggregating individual DOA estimates in a given time-frequency region $\Upsilon_c = \{(\tau, \varkappa) | \tau \in [\tau_c - \tau_w/2, \tau_c + \tau_w/2]\,,\ \varkappa \in [\varkappa_c - \varkappa_w/2, \varkappa_c + \varkappa_w/2]\}$ where τ_c is the center and τ_w is the length of the time interval, $\varkappa_c$ is the center and $\varkappa_w$ is the range of the frequency interval, respectively. Original usage of RENT allows identifying single-source TF bins. It can also be used to identify bins that predominantly contain noise or diffuse field. This is done by identifying two different thresholds (THR_{ss}, THR_{df}) for grouping single source and noise/diffuse field TF bins. The pseudocode for this version can be followed with the Algorithm 6.

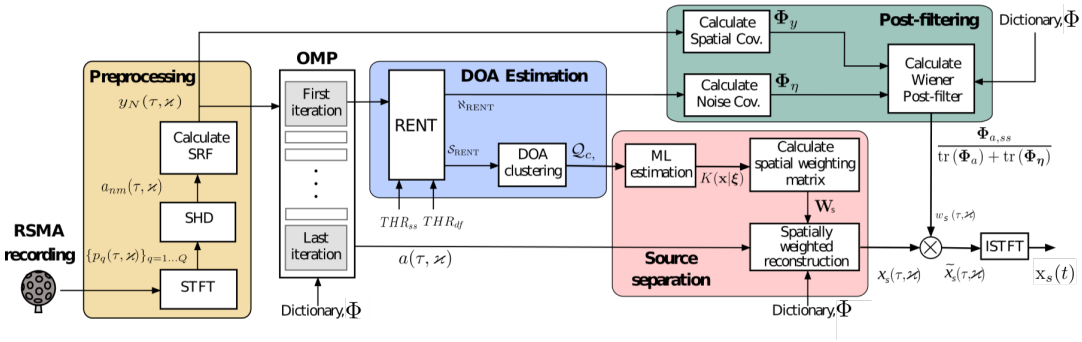
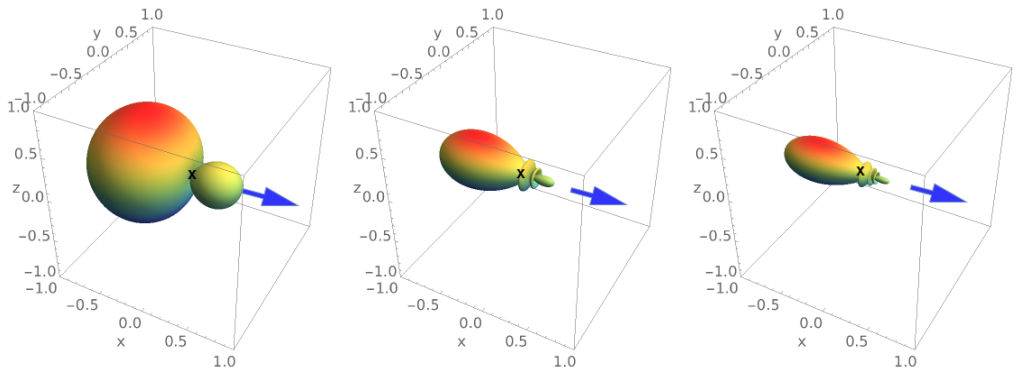


Figure 3.9: Different stages of the proposed joint DOA estimation and source separation method based on sparse plane wave decomposition.

- (a) $B_i \geq THR_{ss} \rightarrow (\tau_i, \varkappa_i) \in \mathcal{S}_{\text{RENT}}$ is a single source contributed TF bin.
- (b) $B_i \geq THR_{df} \rightarrow (\tau_i, \varkappa_i) \in \mathcal{V}_{\text{RENT}}$ is a valid TF bin having sources.
- (c) $THR_{df} > B_i \rightarrow (\tau_i, \varkappa_i) \in \aleph_{\text{RENT}}$ is a non-valid TF bin which does not have any source.

(ii) Source Separation Methods : While RENT provides the initial spatial information for the source separation, a combination of RENT and a source separation technique can be used as a joint separation. The general form of the separation block is reserved as a black box that has inputs $(S, \Omega_s, a_{nm}(\tau, k))$ and outputs of $x_s(\tau, k)$. The placed method selected for the black box is the SPWD-based technique represented in section 3.2.2 for the joint separation. The flow diagram of the algorithm can be observed from the Fig. 3.9.

(iii) Wiener post-filtering (WPF) using Residual Energy Test : Spatial filtering is an effective technique to separate acoustic sources from their spatial information. However, it is impossible to obtain a perfect beam in the spherical array steering directions due to spherical harmonics order limitation. While a spatial filter is designed in the opposite direction of a propagating plane wave, interference coming from the plane wave in the opposite or near directions would already occur in the obtained spatial filter outputs. For example, spatial filters placed to capture the plane wave components in the directions of $\theta, \phi = (90°, 0°)$ would not suppress the full



(a) (b) (c)

Figure 3.10: Monochromatic plane wave distributions calculated in the propagating directions of $\theta, \phi = (90^o, 180^o)$. Spatial filter is placed in the directions of $\theta, \phi = (90^o, 0^o)$ with the arrow looking directions and cross sign is the center of the sphere . From left to right; $N = 1, N = 3$ and $N = 4$.

interference coming from the propagating plane wave components in the directions of $\theta, \phi = (90°, 180°)$. This condition is represented in Fig. 3.10 with the SRP due to a single monochromatic plane wave plotted as a directivity balloon, for different orders N. It may be observed that using a spatial filter in the highest order spherical harmonics would result a sharper beam which would eventually result in better DOA estimation performance, but the higher spherical harmonics order limits the performance. Therefore, PSD-based Wiener post-filtering can be used to improve the separation performance for the acoustic signals when separation performance is limited according to the source positions such as back lobes.

Computational cost would increase with the analysis of the full time-frequency bins. Therefore, calculation and then application of Wiener post-filtering mask only in the valid time-frequency bins $\mathcal{V}_{\text{RENT}}$ decrease the computational cost. $\aleph_{\text{RENT}}$ is only used for the noise cross-covariance matrix calculation.

Inverse short-time Fourier transform (ISTFT) with the same $\beth$ and overlap $\beth_0$ is used to obtain time-series signals for the separated active sound source spectrograms. Each separation would be transferred through a single channel output. The pseudocode for the algorithm is given in Algorithm 8.

Algorithm 7 Single time-frequency bin SPWD

Input: Over-complete dictionary Φ with D kernels in M steering directions, each atom's pseudo inverse matrix $\mathbf{\Gamma}_d^{\dagger}$ and residual vector estimator $\left[\mathbf{I} - \mathbf{\Gamma}_d \mathbf{\Gamma}_d^{\dagger}\right]$, the adapted spatial filter W for each source distribution. The residual vector computation is limited with L_{max} iterations.

Output: Separated sources in time-frequency bin $x_s(\tau, \varkappa)$.

$y(\tau, \varkappa) \rightarrow \mathbf{R}(\tau, \varkappa)$

$i = 0$

while $i < L_{max}$ **do**

  $\mathbf{\Gamma}_i = \max_{\mathbf{\Gamma} \in \Phi} \langle \mathbf{\Gamma}, \mathbf{R}(\tau, \varkappa) \rangle$ `// Find best fit atom d`

  $\mathbf{a}_i = \mathbf{\Gamma}_i^{\dagger} \mathbf{R}(\tau, k)$ `// Calculate the magnitude`

  $\mathbf{R}(\tau, \varkappa) = \left[\mathbf{I} - \mathbf{\Gamma}_i \mathbf{\Gamma}_i^{\dagger}\right] \mathbf{R}(\tau, k)$ `// Calculate the residual vector`

  CalculateRatio$(\mathbf{R}(\tau, \varkappa), \mathbf{y}(\tau, \varkappa)) \rightarrow B$

  if $B < THR_{SPWD}$ **then**

    `// break;`

  end

  i = i + 1

end

$\overline{\Phi}_s \mathbf{W} \mathbf{a} \rightarrow x_s(\tau, \varkappa)$

While the proposed approach is a joint source DOA estimation and separation method, its outputs are audio objects and the associated metadata (except distance). The evaluations in the randomly selected scenarios and controlled setups for the performance measurements are presented in Chapter 4 and Appendix A.

Algorithm 8 Joint Source DOA Estimation and Separation via SPWD-ASF and Wiener post-filtering

Data: Over-complete dictionary Φ with D atoms in M steering directions, the recorded microphone signals $p_q(t)$ for $t \in [t_0, \cdots, t_0 + T]$. Note that $\{\tau, \varkappa\} = \mathcal{S}_{\text{RENT}} \cup \aleph_{\text{RENT}} \cup \mathcal{V}_{\text{RENT}}$

Result: Separated sources $x_s(t)$

$p_q(t) \rightarrow p_q(\tau, \varkappa)$ // STFT

// Computation over each interval

for *each* Υ_c **do**

$RENT \rightarrow \mathbf{S}, \mathbf{\Omega_{c,s}}, \mathbf{x}_{c,s}, y(\tau, \varkappa)$,$\text{S}_{\text{RENT}}, \aleph_{\text{RENT}}, \mathcal{V}_{\text{RENT}}$ // Algorithm 6

$AssociateDOA(\mathbf{\Omega_{s,c}}, \mathbf{\Omega_{s,c-1}}, \Psi)$

$\text{E}\big[y(\tau, \varkappa) y^H(\tau, \varkappa)\big]_{\tau, \varkappa \in \aleph_{\text{RENT}}} \xrightarrow{\text{mean}} \Phi_\eta$

$\mathbf{x}_{c,s} \xrightarrow{\text{MLE}} \mathbf{W}_{c,s}$

for *each* $\tau, \varkappa$ **do**

SourceSeparation($\mathbf{W}, y(\tau, \varkappa)) \rightarrow x_s(\tau, \varkappa)$// Algorithm 7

// To decrease the computational cost, use only valid bins

if $\tau, \varkappa \in \mathcal{V}_{RENT}$ **then**

$\mathbf{\Phi_a} \approx \mathbf{L}^\dagger(\mathbf{\Phi_y} - \mathbf{\Phi}_\eta)(\mathbf{L}^\dagger)^H$;// Source covariance matrix calculation

$\widetilde{x_s}(\tau, \varkappa) = x_s(\tau, \varkappa).w_{a,ss}(\tau, \varkappa)$;// Wiener post-filter

else

$\widetilde{x_s}(\tau, \varkappa) = x_s(\tau, \varkappa)$;

end

end

end

$\widetilde{x_s}(\tau, \varkappa) \rightarrow x_s(t)$// ISTFT

CHAPTER 4

RESULTS AND EVALUATION

This chapter contains the evaluation of the algorithms developed during this thesis. While the evaluations presented in this chapter use simulated, emulated and real recordings made with an Eigenmike em32 microphone array [24], other RSMAs or synthetic Higher-Order Ambisonics recordings can also be used. Various different types of objective and subjective evaluations are also presented.

4.1 Evaluation Metrics

Different evaluation metrics used in to assess the proposed approaches for DOA estimation and source separation are described in this section.

4.1.1 Performance Evaluation of DOA Estimation

The evidence of DOA estimation methods' effectiveness depends on the objective metrics as average or root mean square of DOA estimation error, the average number of identified sources, and source count disparity. It is assumed that the DOA estimation algorithm under test outputs the number of sources, S_{est} in the scene and the unit vectors $\mathbf{u_{est}}$ in the directions of their incidence.

Average DOA estimation error is an indicator of the DOA estimation accuracy. It is the average of the absolute differences between real and predicted DOAs and is given as :

$$\Theta = \frac{1}{S}\sum_{s=1}^{S} \left|\arccos < \mathbf{u}_{s,est}, \mathbf{u}_{s,real} >\right|, \tag{4.1}$$

where S is the total number of the estimated sources and $\mathbf{u}_{s,est}$ is the estimated unit vector and $\mathbf{u}_{s,real}$ is the known DOA unit vector. Note that the association of each predicted DOA to the real ones depend on the criterion as :

$$\Psi > \left|\arccos < \mathbf{u}_{s,est}, \mathbf{u}_{s,real} >\right|, \tag{4.2}$$

where Ψ is used as a threshold to match each estimated DOA to the real ones.

DOA estimation root mean square error (RMSE) is another indicator of the DOA estimation accuracy. While each DOA estimation error directly influences the average DOA estimation error, RMSE is significantly affected with higher-valued errors. It is square root of the average of squared absolute differences between real and predicted DOAs and is given as :

$$\Theta = \sqrt{\frac{1}{S}\sum_{n=1}^{S} \left|\arccos < \mathbf{u}_{s,est}, \mathbf{u}_{s,real} >\right|^2}, \tag{4.3}$$

where S is the total number of the estimated sources and $\mathbf{u}_{s,est}$ is the estimated unit vector and $\mathbf{u}_{s,real}$ is the real value of the DOA.

Average number of source count is the evaluation parameter for the comparison between the number of sources in the scene and the number of estimated sources.

Source count disparity is the metric for the evaluation of the source count estimate's (S_{est}) accuracy. Since the proposed algorithms were developed for direction of arrival estimation, source count disparity (ΔS) indicates a particular method's effectivity in estimating source count. The parameter is denoted with the following expression (4.4).

$$\Delta S = S_{est} - S_{real} \tag{4.4}$$

An algorithm with negative source count disparity is a more serious problem indicating that some of the active sources can not be identified. On the other hand, a vanishing or positive source count disparity means that the algorithm potentially identifies all of the sources and the multipath components.

4.1.2 Source Separation Evaluation Metrics

Source separation evaluation metrics are partitioned into two groups as objective and subjective evaluation. Source separation performance can be measured by comparing the estimated with the reference signals using the methods in the referenced articles [99] [100]. Besides, perceptual evaluation of the source separation would also be conducted via a survey of different individuals using the popular listening test methods, such as ITU-R BS.1534 - MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [101]. MUSHRA is a method for conducting a codec listening test to evaluate the output's perceived quality from lossy audio compression algorithms. As source separation is an operation for transcoding between scene-based to object-based format, we have applied webMUSHRA to score the audible quality of the separated sources [6].

Objective evaluation consists of the separation quality measurements in terms of signal to interference ratio (SIR), signal to distortion (SDR), and signal to artifacts ratio (SAR) [99]. Besides, perception evaluation of speech quality (PESQ) [102] by the individuals is another separation quality metric grouped in the evaluation. In the next subsection, the objective metrics that will be employed in the evaluations will be defined and described.

Objective evaluation metrics : The signal model $\mathbf{X}$ having convolutive mixtures to be separated is denoted within the following expression.

$$\mathbf{X} = \mathbf{A}\mathbf{s_j} + \eta_j, \tag{4.5}$$

where $\mathbf{s_j}$ is the original referenced signal and η_j is the noise corresponding to the sensors or environment. The summation for the source separation outputs satisfies the condition given in equation (4.6).

$$\mathbf{X} = \sum_{j}^{N} \hat{\mathbf{s}}_\mathbf{j} + \hat{\eta}_j, \tag{4.6}$$

where $\hat{\mathbf{s}}_\mathbf{j}$ are the separated signals. The aiming for the specified source separation algorithms is to identify the measurement parameters as SIR, SDR, and SAR via estimating the resulting separated signal. Estimated source separation output can be

decomposed into four terms to evaluate the objective metrics.

$$\hat{\mathbf{s_j}} = \mathbf{s}_{\text{dtarget}} + \mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}}, \tag{4.7}$$

where $\mathbf{s}_{\text{dtarget}} = f(s_j)$ is a modified version of target signal with distortion, $\mathbf{e}_{\text{interf}}$ is the term related to the interference coming from the other sources, $\mathbf{e}_{\text{noise}}$ is the noise and $\mathbf{e}_{\text{artif}}$ is the artifact error terms. Each parameter is estimated using orthogonal projections with the reference signal for the target $\mathbf{s_j}$, separated source outputs $\hat{\mathbf{s_j}}$, and signals for the others $\mathbf{s_{j'}}$ are given as:

$$\mathbf{s}_{\text{dtarget}} = \frac{< \hat{\mathbf{s_j}}, \mathbf{s_j} > \mathbf{s_j}}{\left|\mathbf{s_j}\right|^2} \tag{4.8}$$

By estimating the sources are mutually orthogonal, the following formula is used for the interference term.

$$\mathbf{e}_{\text{interf}} = \sum_{j \neq j'} \frac{< \hat{\mathbf{s_j}}, \mathbf{s_{j'}} > \mathbf{s_{j'}}}{\left|\mathbf{s_{j'}}\right|^2} \tag{4.9}$$

By estimating the noise is mutually orthogonal to each source, the following formula is used for the noise term.

$$\mathbf{e}_{\text{noise}} = \sum_{i=1}^{M} \frac{< \hat{\mathbf{s_j}}, \eta_{\mathbf{i}} > \eta_{\mathbf{i}}}{\left|\eta_i\right|^2} \tag{4.10}$$

$$\mathbf{e}_{\text{artif}} = \hat{\mathbf{s_j}} - \mathbf{s}_{\text{dtarget}} - \mathbf{e}_{\text{interf}} - \mathbf{e}_{\text{noise}} \tag{4.11}$$

According to the terms given in the equations (4.8), (4.9), (4.10), (4.11), the mathematical definitions of the evaluation metrics are given within the following expressions. The Python code used for the objective evaluation is available in [103] [99].

$$SDR = 10 \log_{10} \frac{\left|\mathbf{s}_{\text{dtarget}}\right|^2}{\left|\mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \mathbf{e}_{\text{artif}}\right|^2} \tag{4.12}$$

$$SIR = 10 \log_{10} \frac{\left|\mathbf{s}_{\text{dtarget}}\right|^2}{\left|\mathbf{e}_{\text{interf}}\right|^2} \tag{4.13}$$

$$SAR = 10 \log_{10} \frac{\left|\mathbf{s}_{\text{dtarget}} + \mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}}\right|^2}{\left|\mathbf{e}_{\text{artif}}\right|^2} \tag{4.14}$$

The selection method of target signals (s_j) for the evaluation determines the source separation algorithms' performance. There are two ways to evaluate the corresponding metrics via selecting s_j either from the original anechoic signals or the omnidirectional component of the array recordings. The calculation of the objective metrics defined for this study uses original anechoic signals by positioning them at the

same location as in the multi-source scenario without any other sources. It is compared with the estimated source signals. As an alternative to this condition, original multi-source scene's omni-directional recordings can also be used as the target signals. As this case is a proof for the improvement obtained via the algorithms, it provides better results when it is compared to the case of anechoic signals used.

Perception Evaluation of Speech Quality (PESQ) is an objective score of speech quality that is highly correlated with the mean opinion scores (MOS) obtained in subjective experiments. It is standardized in ITU-T recommendation P.862 [104] as the score of 0.5 to 4.5, while higher scores mean good quality. The main purpose of the PESQ score is the measurement of efficiency in handset telephony and narrowband speech codecs. Rix et al. define the method to be used to estimate PESQ [102]. Hu et al. also define the mathematical expressions of the PESQ score for evaluating signal distortion, noise distortion, and overall speech quality [105]. The Python code used for this evaluation given in the work is available in [106].

Speech Intelligibility (SI) is another measure for the perceptual quality of the speech that can replace subjective listening tests [107]. SI as an objective intelligibility measure results in a high correlation ($\rho = 0.95$) with the intelligibility of both noisy, and TF-weighted noisy speech.

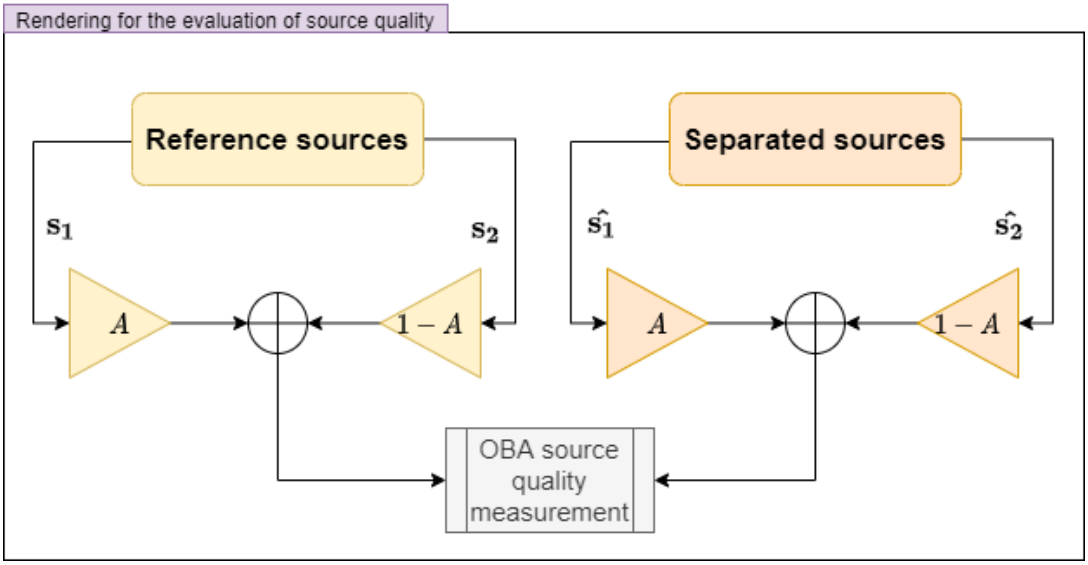


Figure 4.1: Source quality evaluation setup for the object-based audio [5].

Object-based audio source quality is another metric to evaluate the source quality

aiming for better rendering outputs. Representing an acoustic scene in terms of objects using the source separation algorithms is a challenging task due to the trade-off between the separation and source quality. While an algorithm could provide better source separation output via suppressing the interference better, it could also degrade the source quality as measured with the metrics such as PESQ or SDR. The main motivation of source separation for object-based audio is presenting the higher quality results in the reproduction phase. Therefore, rendering the separated channels using a gain parameter A associated to each source would remove the distortion and artifacts at the output. The sum of gain parameters are set to be one for normalization. Liu et al. propose an objective evaluation technique for the rendered outputs to measure the source quality [5]. The evaluation setup is given in Fig. 4.1. We present the results using this approach to provide insight into what would happen in a real-life use-case.

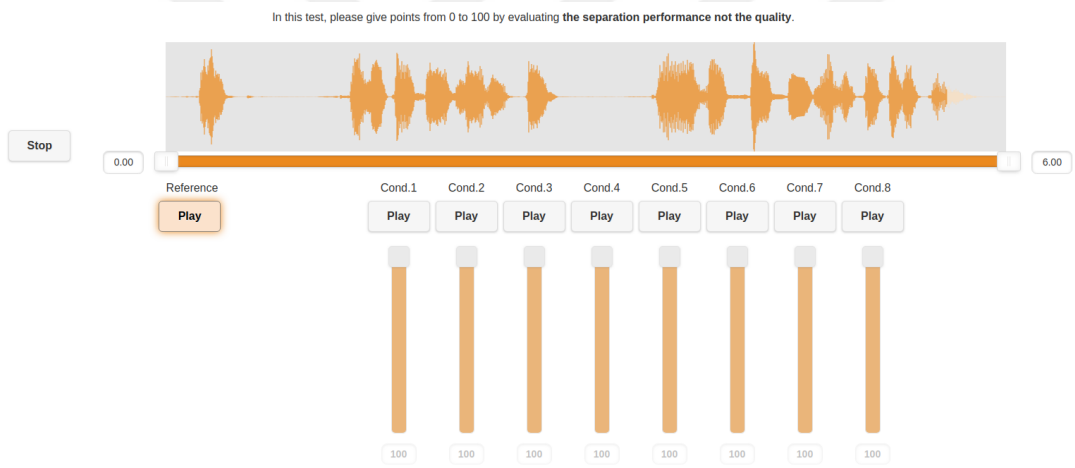


Figure 4.2: webMUSHRA evaluation setup containing randomly placed eight signals as reference, anchor and six source separation algorithm results [6].

Subjective evaluation metrics : Another perceptual study to be performed in the evaluation of separation is a MUSHRA test conducted with several participants using the webMUSHRA tool [6], where the participants scored source separation quality of the selected algorithms in comparison with each other. Reference, anchor, and separated signals are displayed to listen in randomized order at each trial. The anchor is the 10 dB distortion added to the mixture, and the reference is the sound object evaluated at the same position without any interfering sources or reverberation. During this

test, the score given by the participants to the anchor signal should be the lowest, and the score given to the reference signal should be the highest. Any difference from this would indicate that the participant lacks an understanding of the task and provides reason to exclude their scores from the final analysis. The screenshot from the webMUSHRA evaluation setup can be viewed in Fig. 4.2.

4.2 Acoustic Impulse Response (AIR) Datasets used for the evaluation

The emulation-based evaluations employ measured acoustic impulse responses. The dataset used in the evaluations is explained below [7].

4.2.1 METU SPARG AIR dataset collected via Rigid Spherical Microphone Array

An eigenmike em32 [24] array was positioned at a central position of the classroom in the METU Informatics Institute Class 5. The measurements were made in the room having a high reverberation time ($T_{60} \approx 1.12$ s) when it is empty. The room has the dimensions $6.5 \times 8.3 \times 2.9$ m and is approximately rectangular. 240 points on a rectilinear grid of 0.5 m horizontal and 0.3 m vertical resolution surrounding the spherical array are selected for the AIR measurements. The height of the array was 1.5 m and the measurement locations were positioned at the heights of 0.9, 1.2, 1.5, 1.8, and 2.1 m from the floor level. The selected positions cover the whole azimuth and an approximate elevation range of $\pm 50^{\circ}$ above and below the horizontal plane. Fig. 4.3 shows the measurement positions.

Room impulse responses (RIRs) at the same positions are also measured using an Alctron M6 omnidirectional microphone. This second set of measurements are used to calculate the corresponding direct-to-reverberant (D/R) energy ratios for these positions. The following definition is used for D/R ratio:

$$D/R = 10 \log_{10} \left[\frac{\sum_{n=n_0-n_w}^{n_0+n_w} |h(n)|^2}{\sum_{n=n_0+n_w+1}^{N_h} |h(n)|^2} \right] \text{ (dB)}, \tag{4.15}$$

where $h(n)$ is the measured room impulse response with length N_h, n_0 is the sample at which the direct component attains its maximum value, $2n_w$ is the width of

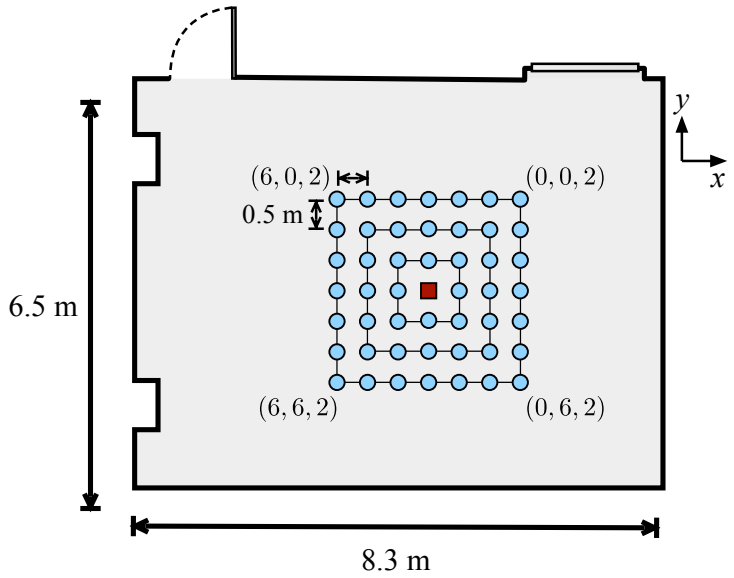


Figure 4.3: The measurement positions inside the classroom are displayed in 2D view. The centered square represents the position of the microphone array and circles represent the source locations. The walls are covered with plaster, the floor is carpeted, and the ceiling has acoustic tiles. One of the wall has a window and the other one has a wooden door that were kept closed during the measurements. [7]

a rectangular window used to select the direct component. n_w is found as 70 at the sampling rate of 48 kHz, which corresponds approximately to 1 m in propagation distance, was necessary and sufficient to eliminate reflections while calculating the D/R ratio for every tested measurement position.

White Gaussian sensor noise insertion : The evaluations also comprise sensor noise available conditions. Microphone signals are obtained via convolving each anechoic signal with the METU SPARG AIR dataset for a single source. A multi-source scenario consists of adding microphone signals obtained from a single source scenario. The noise insertion module is used to create the sensor noise [108] via covering the full frequency bandwidth. After calculating each full-length microphone signal's average power, this tool inserts the sensor noise concerning the defined noise level. Sensor noise in the defined levels of 0 dB, 10 dB, 20 dB, and 30 dB are inserted into the microphone signals according to the average power in each microphone channel before the computation starts. While the average noise power is fixed for the full

bandwidths, the sensor noise insertion would result in higher SNR levels for some of the frequency bandwidths that contain lower active source powers. The examples for full-bandwidth noise available time-frequency spectrograms are given in Fig. 4.4.

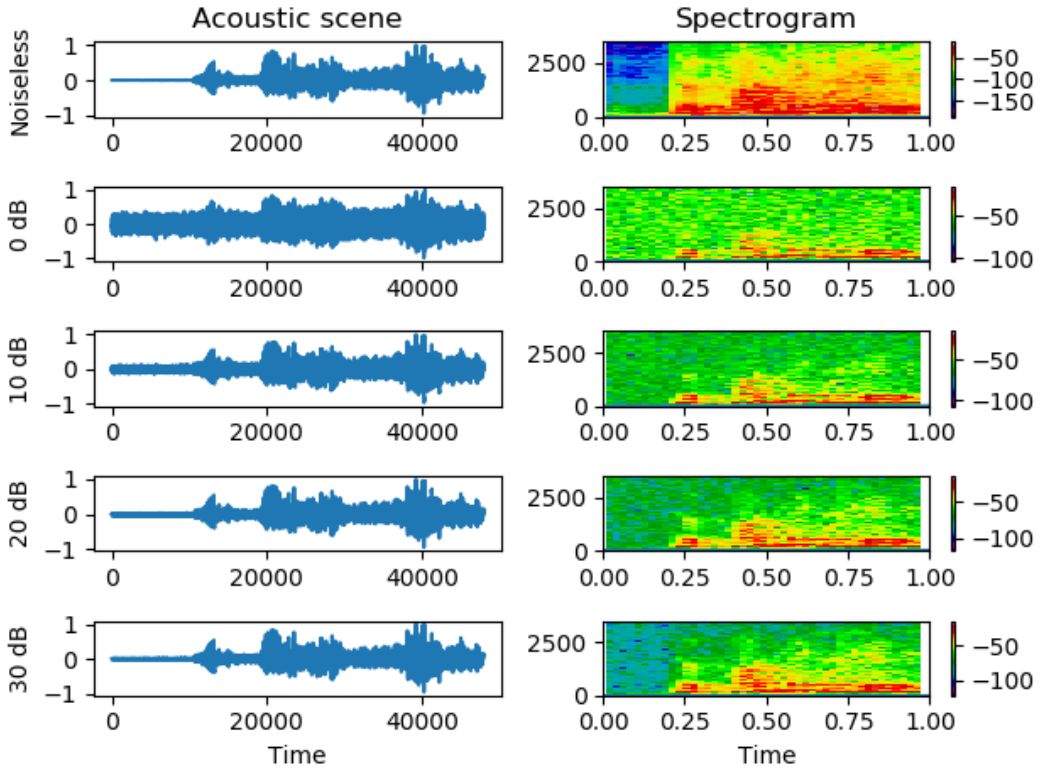


Figure 4.4: Examples of an omnidirectional acoustic scene for different levels of white Gaussian sensor noise inserted to the microphone signals.

4.2.2 AIR simulations with plane wave definitions

Owing to the closed form solution of a plane wave over a sphere, ideal conditions without reverberation can be numerically simulated. The condition obtained with zero reverberation and noise would also be used to measure the best output performances of the algorithms and compare the identified algorithms with the state of the art methods in the cases where the selected algorithms fail due to high reverberation. Therefore, we have simulated each AIR as a monochromatic plane wave with the corresponding azimuth ϕ_k and elevation θ_k given in equation (2.34). These spherical harmonic coefficients are multiplied with the anechoic recordings' magnitudes and

are mixed to create a non-reverberant acoustic scene over the array. These spherical harmonic coefficients are fed to the algorithms for the available computations in this chapter.

4.3 Evaluation Results

4.3.1 DOA estimation using HiGRID for multiple sources via Rigid Spherical Microphone Array

In this subsection, HiGRID is evaluated using acoustic impulse responses (AIRs) in the described AIR dataset [7], as well as the effect of additive white Gaussian noise to the robustness of the algorithm is introduced via emulated acoustic scenes.

The first scenario to evaluate the proposed algorithm with near-incoherent (i.e., nominally W-disjoint) sources contains the first 4 seconds of anechoic speech recordings from B&O Music for Archimedes CD [109]. These signals are speech recordings in English and Danish from two female and two male speakers and they are resampled to 48 kHz. Each recording is convolved with the randomly selected AIRs and mixed to obtain the acoustic scene. The speech signals were normalized according to ITU P.56 [110] using the `VOICEBOX` toolbox [111].

Four-source scenario having near-coherent sources, 4 seconds (01:00-01:04) of the anechoic recordings of Gustav Mahler's Symphony Nr. 1, fourth movement [112] was also processed with the proposed algorithm. Only the four violin tracks were used in which violins play the same musical phrase in unison. While there may be minor differences, the source signals have substantially similar magnitude spectrograms (see Fig. 4.5). Selected test signals constitute a very exacting and realistic scenario in object-based audio capture. The sampling rate was also selected as 48 kHz.

A maximum tree level of $K_{\max} = 4$ corresponding to a grid resolution of 7.33° is used to evaluate HiGRID. The maximum order of SHD is $N = 4$, while the frequency range is selected between 2608 Hz (i.e., $kr_a = 2$) and 5216 Hz (i.e., $kr_a = 3$) for order, $N = 4$. The FFT size and the hop size are 1024 and 64 samples, respectively.

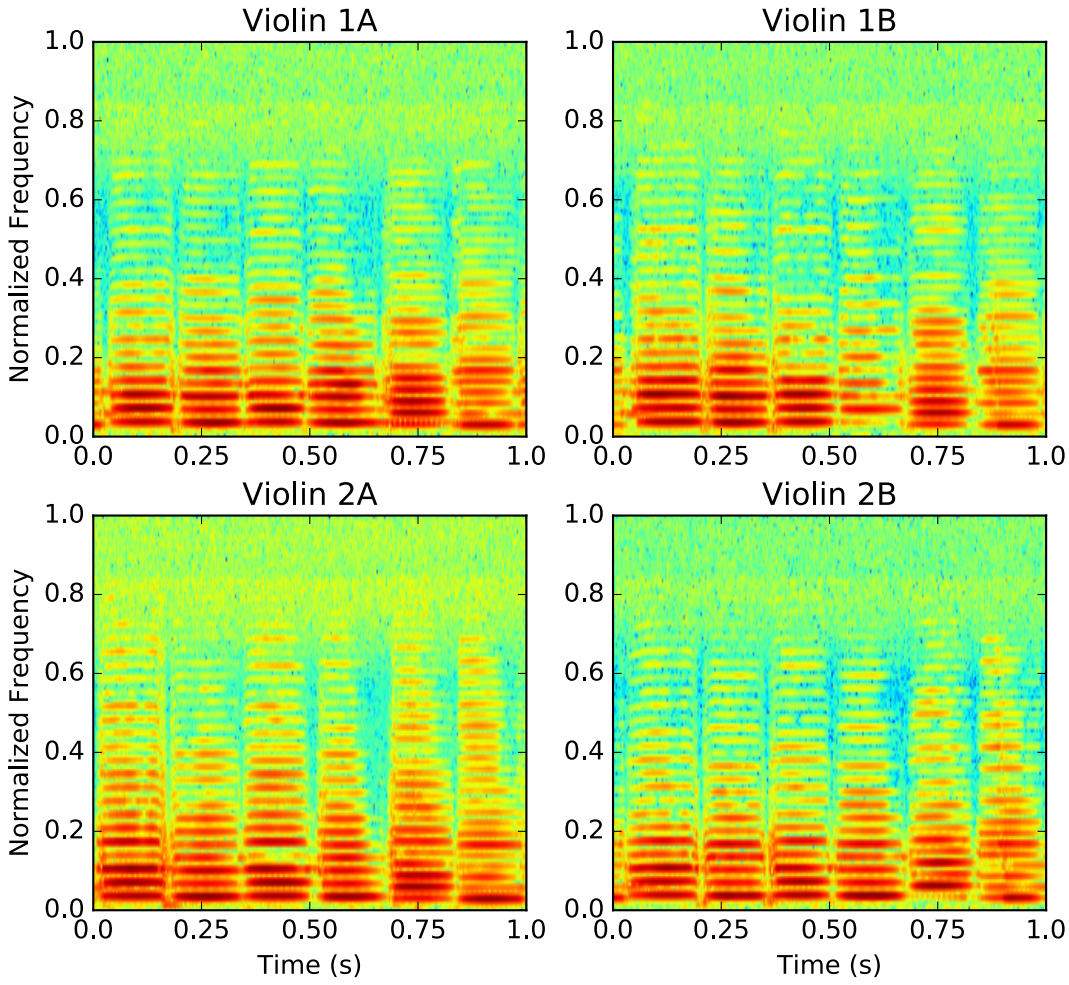


Figure 4.5: Magnitude spectrograms of the first second of the four violin tracks used in the evaluations.©2018 IEEE [1]

For each test case, the AIRs are clustered according to their D/R ratios with a width of 1.38 dB. The clusters' D/R ratios are ordered as -1.65, -0.27, 1.11, 2.50, 3.88, 5.26, 6.65 dB. Source combinations in each test scenario are selected only from one of the clusters aiming to create a proper setup that consists of sources having similar D/R ratios. The criterion for selecting sources over the same cluster is that the minimum angle difference to be greater than $\pi/4$. According to each cluster, ten test scenarios are randomly generated for each $S = 1$, 2, 3, and 4 source cases. DOA estimates are assigned to the ground truth via Kuhn-Munkres algorithm [113]. If the DOA estimation error was found to be larger than $\pi/4$, that case was identified as an extreme value and excluded from the analysis. These scenarios correspond to having zero sensor noise scenarios where the source signals (violins or speech) are convolved

with the measured AIRs and summed to obtain the test signals.

Evaluation results : According to the evaluation results for the DOA estimation of speech signals, 19 extreme values are obtained in 280 distinct test scenarios and 700 source positions. These are excluded from the calculation of average DOA estimation error. The calculated average DOA estimation errors for $S = 1$, 2, 3, and 4 sources are $2.71°$, $2.99°$, $3.20°$, and $3.39°$, respectively. The overall average was $3.18°$, and it is observed that the average DOA estimation error slightly increases with the source count. DOA estimation errors for speech signals can be observed in Fig. 4.6.

According to the evaluation results for the DOA estimation of violin signals, 5 extreme values are obtained in 280 distinct test scenarios and 700 source positions. These are excluded from the calculation of the average DOA estimation errors. The calculated average DOA estimation errors for $S = 1$, 2, 3, and 4 sources are $3.34°$, $3.44°$, $3.97°$, and $4.37°$, respectively. The overall average was $3.96°$. Similar to the case with speech signals, it is observed that the average DOA estimation error slightly increases with the source count. Fig. 4.7 shows the DOA estimation errors for violin signals.

The estimated average number of sources are also found for $S = 1$, 2, 3, and 4 source cases as, $S_{\text{avg}} = 1.67$, 2.97, 3.52, 4.65 for speech, and 1.02, 2.01, 2.78, 3.52 for violin signals, respectively. While the D/R ratio was high and when the source count was low, negative source count disparity is observed to be low. The worst performance occurred for the relatively low D/R ratio of -0.27 dB and $S = 4$ for speech signals. The average number of sources recognized for this condition was 5.7. The worst performance occurred for the moderate D/R ratio of 1.11 dB and $S = 4$ for violin signals. The average number of sources recognized for this condition was 3.2. Negative source count disparity was observed 8 times for speech and 47 times for violins over 280 test scenarios for each source type.

The second setup contains the simulations via the addition of sensor noise in different SNR levels according to the noise insertion method given in Section 4.2. The tested levels were 0, 10, 20, 30 dB SNR. Anechoic violin signals are used for the convolution with AIRs. The selection of the scenarios with the AIR recordings is made in the same way as the ideal case.

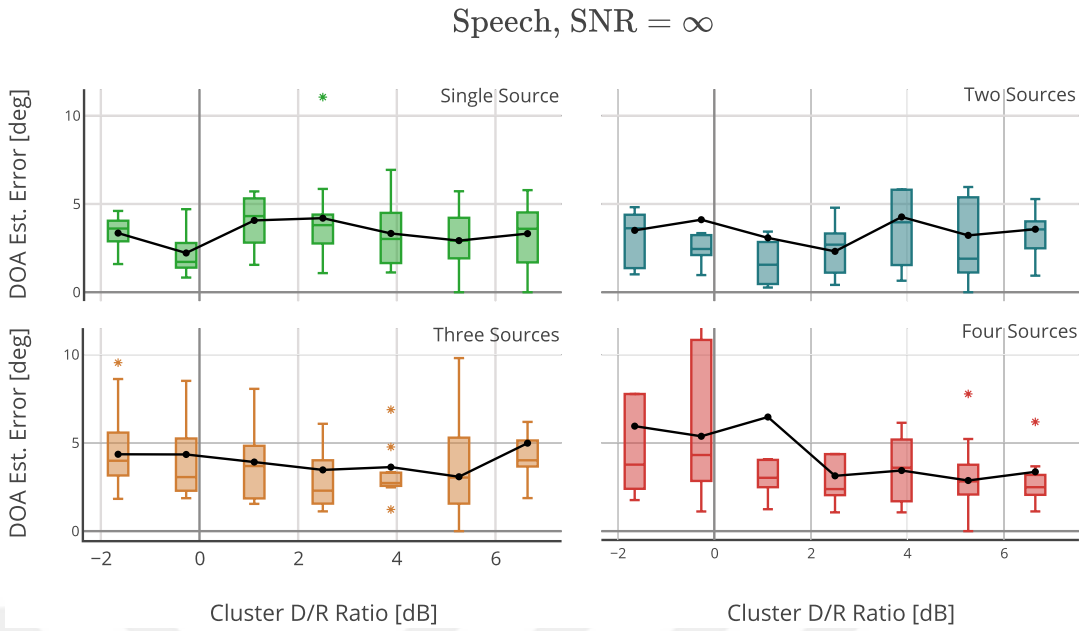


Figure 4.6: HiGRID DOA estimation results are given for speech signals for $S = 1$, 2, 3, and 4 sources and different D/R ratio levels. The box plots reveal the error distributions. The solid lines indicate the mean DOA error. Asterisks indicate outliers. Extreme values were excluded from the graph.©2018 IEEE [1]

Fig. 4.8 shows the distribution of the error for different number of sources and different noise levels tested. The rows in the figure indicate grouping (clustered regions for D/R ratio) according to the number of sources and the columns indicate grouping according to the noise level. The effect of noise level over the DOA estimation error is investigated with a one-way analysis of variance (ANOVA) by selecting the independent variable as SNR, which has four levels from 0 to 30 dB. On the other hand, the dependent variable is the DOA estimation error. Results of the ANOVA are significant at $\alpha = 0.05$ level ($F(3, 2564) = 3.625$, $p = 0.013$). Levene's test indicated that the variance is different across the tested groups ($F(3, 2564) = 3.796$, $p = 0.01$). It is observed that the difference between the error means between 0 dB SNR and 10 dB SNR as well as 0 dB SNR and 20 dB SNR are significant according to the post-hoc comparisons using the Games-Howell test. However, these differences (i.e., $0.51°$ and $0.49°$, respectively) can be neglected in a practical sound source localization scenario. Noise do not have a significant effect over the average number of identified sources and the negative source count disparity. The averaged number of

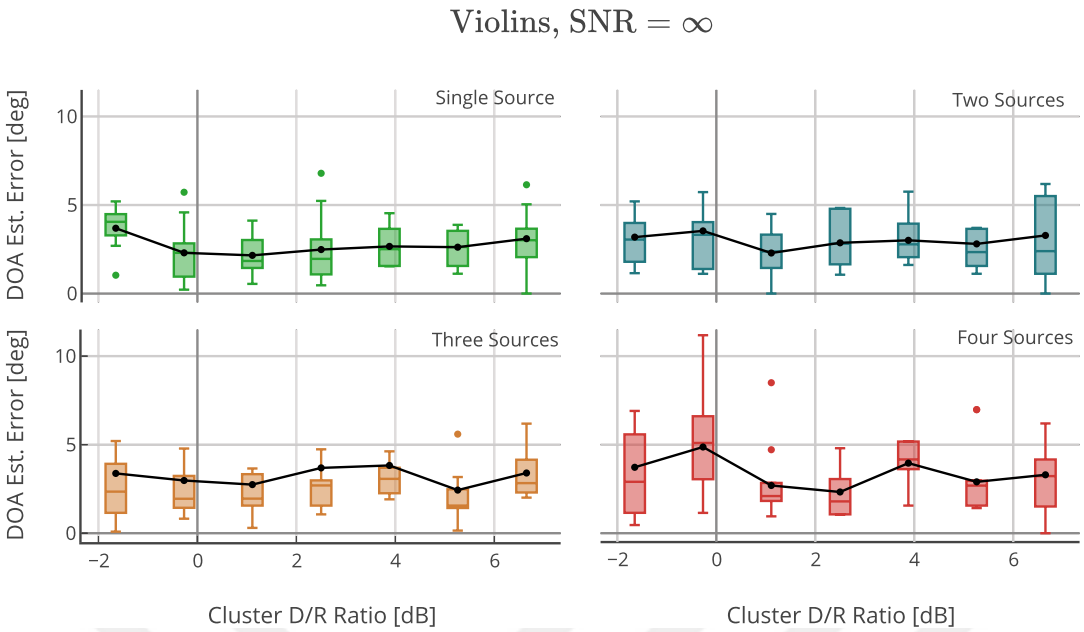


Figure 4.7: HiGRID DOA estimation results are given for violin signals for $S = 1$, 2, 3, and 4 sources and different D/R ratio levels. The box plots reveal the error distributions. The solid lines indicate the mean DOA error. Asterisks indicate outliers. Extreme values were excluded from the graph.©2018 IEEE [1]

SNR (dB)	$S = 1$	$S = 2$	$S = 3$	$S = 4$
0	1.22	2.08	2.92	3.57
10	1.01	2.30	2.74	3.75
20	1.00	2.12	2.78	3.38
30	1.05	2.16	2.84	3.65
∞	1.02	2.01	2.78	3.52

Table 4.1: Average number of localized sources for different SNRs ©2018 IEEE [1]

sources in different sensor noise level are revealed in Table 4.1. It may be noted that the source count is slightly overestimated on average, and vice versa for lower source counts (i.e., $S = 1$ and $S = 2$). Negative source disparity was 46, 44, 45, 43 for SNRs of 0, 10, 20, and 30 dB, over 280 trials each, respectively. As a conclusion, the proposed method is considered in general to be robust to additive, and uncorrelated white Gaussian noise.

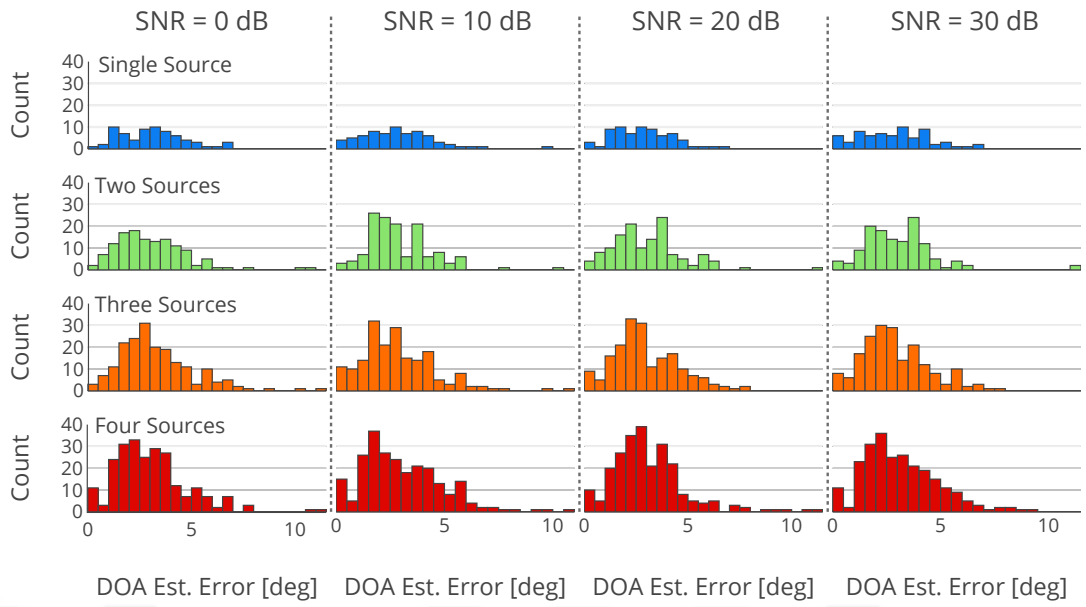


Figure 4.8: Histograms of DOA estimation errors for different number of sources and different noise levels. Bin size is 0.5° and each count corresponds to the DOA estimation error for a single source.©2018 IEEE [1]

The third evaluation corresponds to real-life original recordings without AIRs in order to test the performance of HiGRID. A pre-concert rehearsal of Nemeth Quartet (a classical quartet consisting of two violins, a viola, and a cello) is recorded using Eigenmike em32 in the small recital hall of Erimtan Museum in Ankara. The recording hall has a reverberation time $T_{60} = 1.19$ s. The spherical array was placed at a central position with the height of 1.5 m among the musicians. Accurate sound source positions can not be estimated precisely due to practical considerations such as musicians' movement during the performance. However, we made AIR measurements with a Genelec 6010A loudspeaker at positions that roughly coincide with the musical instruments' positions. In order to assess the accuracy of the proposed method, full grid search SRP-based DOA estimates were obtained from windowed direct-path components of AIRs and they were used as a reference. A five-second excerpt from the third movement of Ludwig van Beethoven's String Quartet Nr. 11, Op. 95 was used in the evaluation. Besides, all the instruments playing the individual parts were not in unison, unlike the near-coherent source scenario simulated in the previous section. Fig. 4.9 shows the recording setup.

Figure 4.9: Setup for the recording of the classical quartet. Position of the microphone array and directions of the instruments are indicated.©2018 IEEE [1]

(θ, ϕ)	**Violin 1**	**Violin 2**	**Viola**	**Cello**
Reference	$(116.1°, 92.4°)$	$(114.9°, 32.3°)$	$(124.4°, 327.8°)$	$(132.4°, 268.4°)$
$2608 - 5216$ Hz	$(109.7°, 89.4°)$	$(109.7°, 28.1°)$	$(112.3°, 340.4°)$	-
$1304 - 5216$ Hz	$(118.1°, 88.5°)$	$(107.1°, 30.9°)$	$(115.0°, 337.7°)$	$(138.1°, 276.4°)$

Table 4.2: Reference and estimated DOAs for the microphone array recording of the classical quartet ©2018 IEEE [1]

HiGRID is used with a maximum tree level of 4 comprising 3072 pixels. For the evaluation, two different frequency ranges were used: 1) 2608 Hz to 5216 Hz, and 2) 1308 Hz to 5216 Hz. The FFT size selected as 1024 with a hop size of 64 samples was used for the STFT. Table 4.2 represents the reference DOAs and estimation results. The observation obtained using the first frequency range is that three out of four instruments (two violins and the viola) can be localized. No spurious sources were detected in this case. Using the second frequency range, all instruments can be localized, along with six spurious DOAs were corresponding to strong room reflections (not shown in the table).

4.3.2 DOA estimation using RENT for multiple sources

The first evaluation scenario for RENT, including multiple coherent or near-coherent sources, are simulated using 6 concurrently active sources by convolving real AIRs with anechoic sound samples. The first four sources (S1-S4) were 4s-long violin tracks from the anechoic recordings of Mahler's Symphony Nr. 1, fourth movement [112]. The remaining two sources (S5-S6) were female and male anechoic speech recordings [109]. The dictionary elements used in RENT are Legendre kernels sampled on a HEALPix grid of resolution level $K_{\max} = 3$, resulting in 768 atoms of size 768×1. This corresponds to $7.33°$ resolution. The frequency range selected for the analysis is between 480 Hz and 3750 Hz. Window size and overlap are 2048 and 50%, respectively. The sampling rate is selected as 48 kHz. Multichannel acoustic impulse responses (AIRs) from the METU SPARG dataset were used [7]. For scenarios involving multiple sources, AIRs are randomly selected from the dataset subject only to the constraint that the angular separation between two sources is more significant than $\pi/10$. Random simulations for each source count is repeated 10 times. Fig. 4.10 shows the results for different number of sources. The average

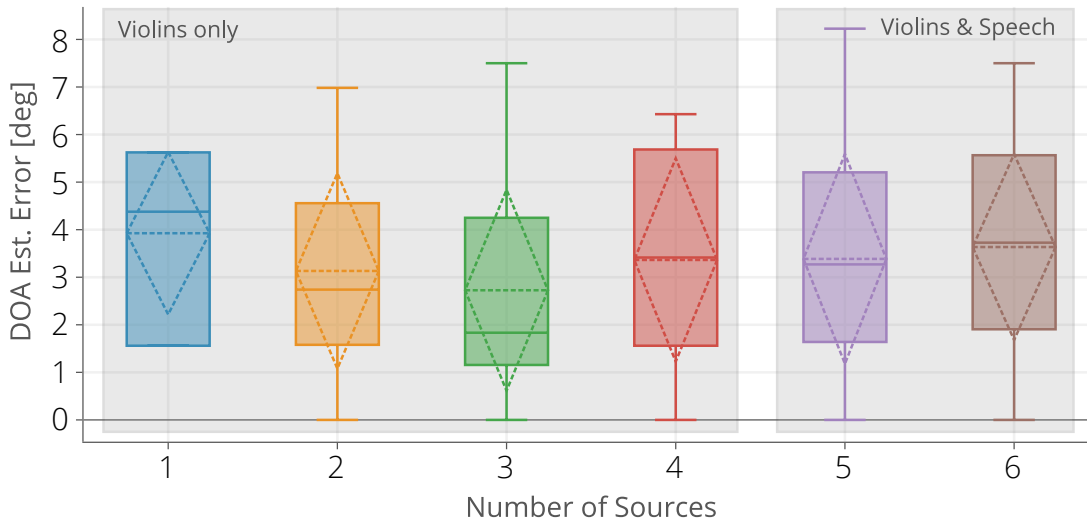


Figure 4.10: DOA estimation errors for different number of sources. The first four sources are violins playing in unison and the last two sources are anechoic speech signals. The dotted triangles indicate mean $\pm$ standard deviation points, respectively.©2019 IEEE [2]

	Number of Sources					
	1	**2**	**3**	**4**	**5**	**6**
RMSE	4.28°	3.74°	3.44°	3.98°	4.03°	3.44°
Mean	3.92°	3.13°	2.72°	3.36°	3.38°	3.63°
Min.	1.55°	0.02°	0.02°	0.02°	0.02°	0.02°
Max.	5.62°	6.98°	7.5°	6.42°	8.22°	7.5°
Stdev.	1.71°	2.05°	2.11°	2.13°	2.20°	1.95°
Avg. Src. Count	1	2	3	4	4.7	5.8

Table 4.3: Descriptive statistics of the DOA estimation experiments.

error is always less than 5° and the number of emulated sources does not significantly affect the error. Table 4.3 shows the statistics of the results. Average number of identified sources is also shown at the last row of the table. The performance of RENT in the identification of source count is observed as that up to 4 sources, RENT can identify all sources correctly. RENT missed one source each in 3 out of 10 scenarios consisting five sources and one source each in 2 out of 10 scenarios consisting six sources.

The second evaluation scenario consists of RENT's performance evaluation in the horizontal plane by RSMA recordings and an anechoic violin signal in order to characterize its DOA estimation accuracy when only a single source is present. Fig. 4.11 shows the DOA estimation errors for the 48 directions. The maximum error is 5.95°. RENT could not localize one of the sources as the error was greater than $\Psi = 10^\circ$, so it is excluded. The figure also shows that better accuracy is possible if the source directions and pixel centers overlap, it is possible that this is due to the perfect alignment of the source and dictionary atom directions.

The third evaluation is the measurement of RENT performance via observing the DOA estimation error among the cases of HEALPix resolutions as $K_{\max} = 2$, and 3. In these cases, six scenarios containing two sources are processed using different levels of pixelization for the sparse dictionary. The effect of the anechoic recording types over DOA estimation error is also investigated for $K_{\max} = 2$. The selected AIRs are convolved with the anechoic female/male recordings and violin/violin signals se-

lected from EBU SQAM (Sound Quality Assessment Material) Audio CD [114]. The positions for the source groupings in the room used for each scenario are given in Fig. 4.12, 4.13 and 4.14. Each scenario is 1 second in duration, and the processed interval is 0.5 s with an overlap of 0.25 s. Totally, 3 processing intervals are obtained for the scenarios. RENT threshold is selected as 0.5, and the analysis frequency range is from 0 Hz to 9 kHz. Tables 4.4, 4.5, 4.6, 4.7, 4.8 and 4.9 represent the DOA estimation errors obtained for the specified scenarios.

The measurements taken from the configured scenarios reveal that processing RENT

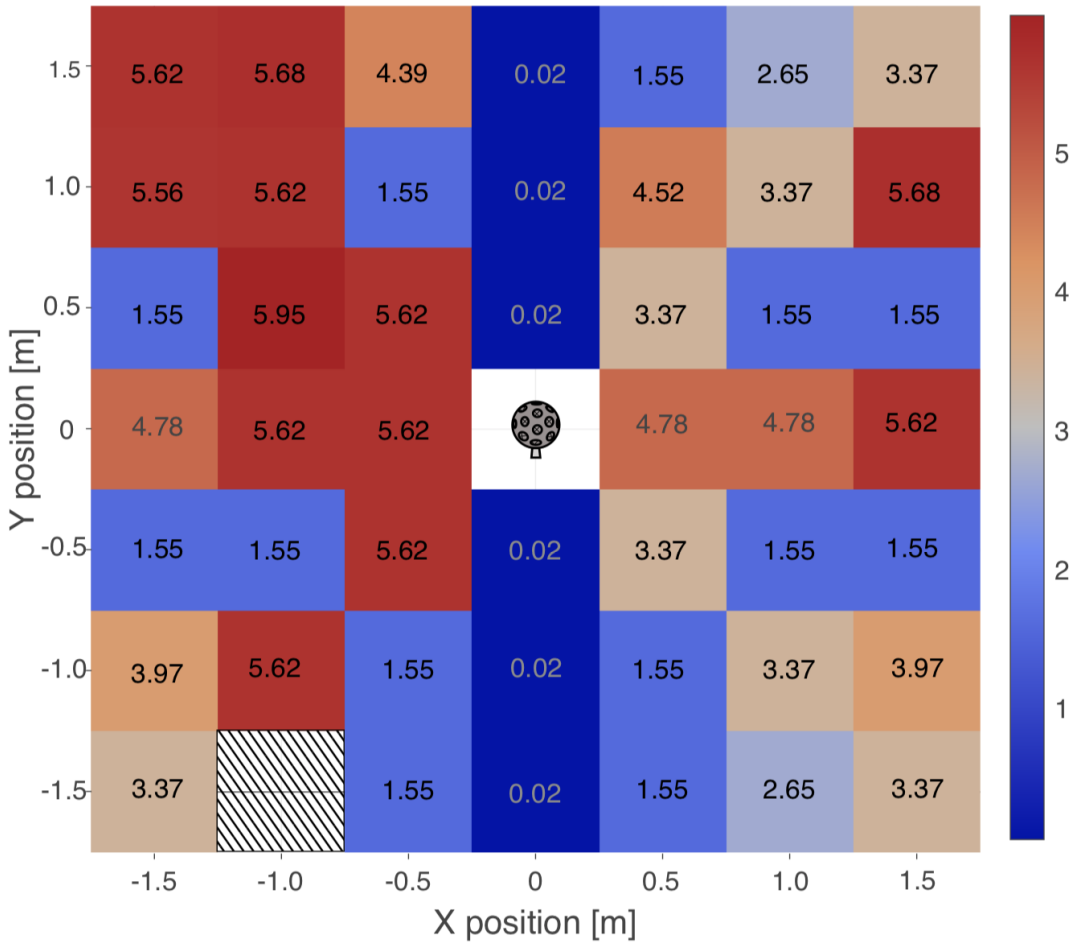


Figure 4.11: Heatmap showing the DOA estimation errors for a single source. Eigenmike is positioned at the center. The single source that cannot be localized is denoted by a hatched pattern. It is neglected in the map because the identified DOA estimation error is greater than 10°.©2019 IEEE [2]

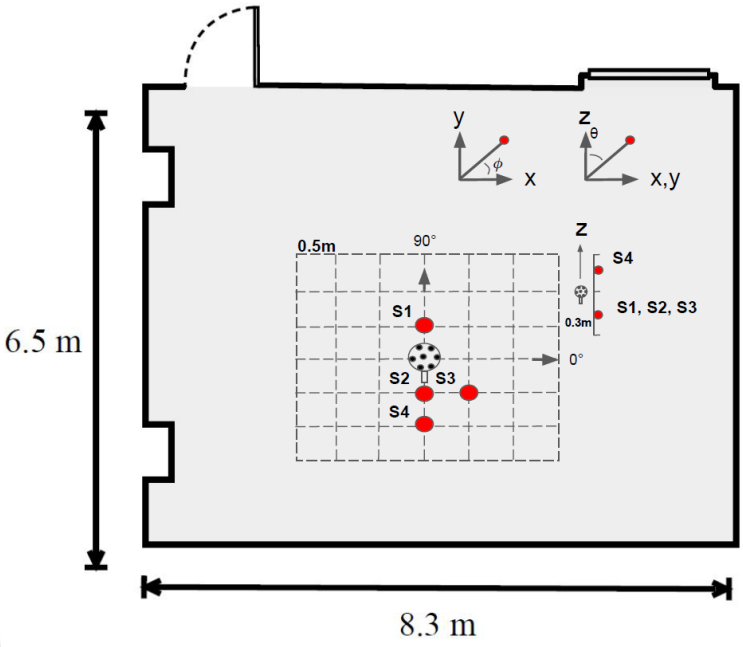


Figure 4.12: Scenarios for the evaluation of RENT performance in small intervals.Scenario 1 (S1,S2),Scenario 2 (S3,S4)

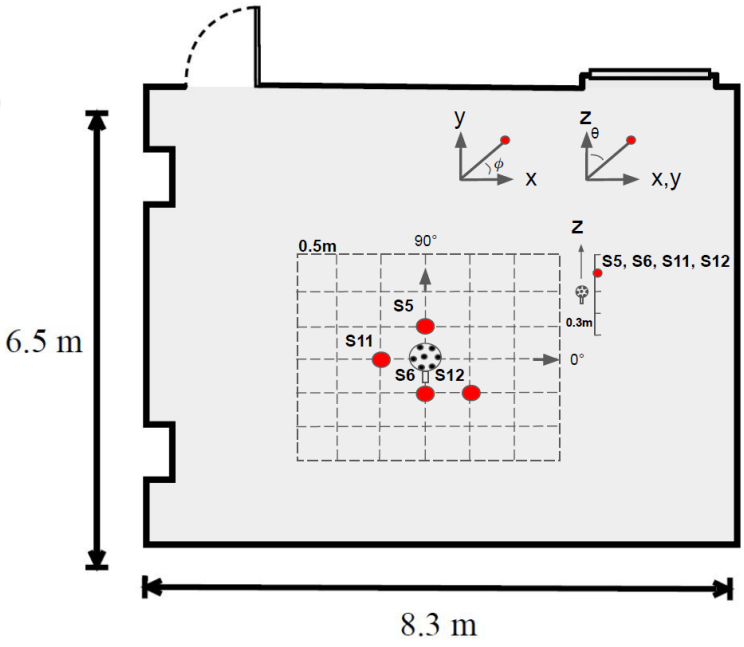


Figure 4.13: Scenarios for the evaluation of RENT performance in small intervals.Scenario 3 (S5,S6), Scenario 6 (S11,S12)

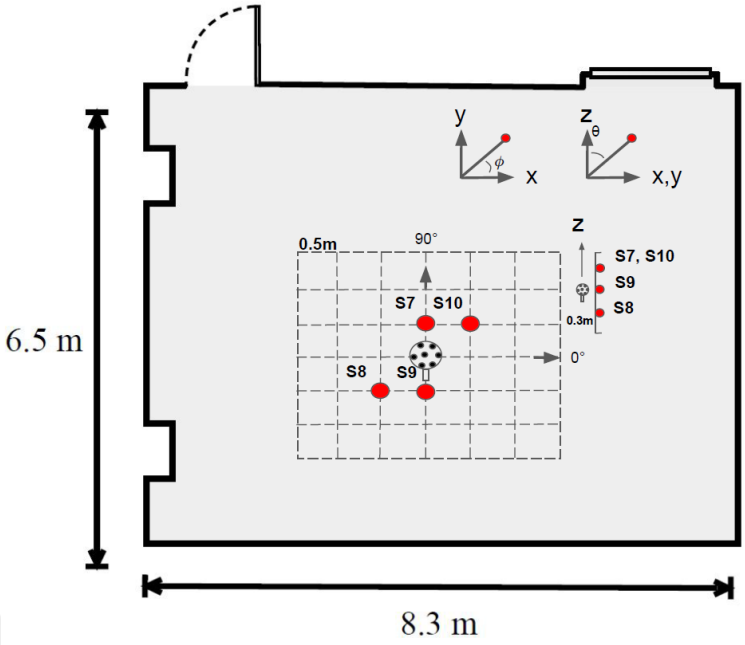


Figure 4.14: Scenarios for the evaluation of RENT performance in small intervals. Scenario 4 (S7,S8), Scenario 5 (S9,S10)

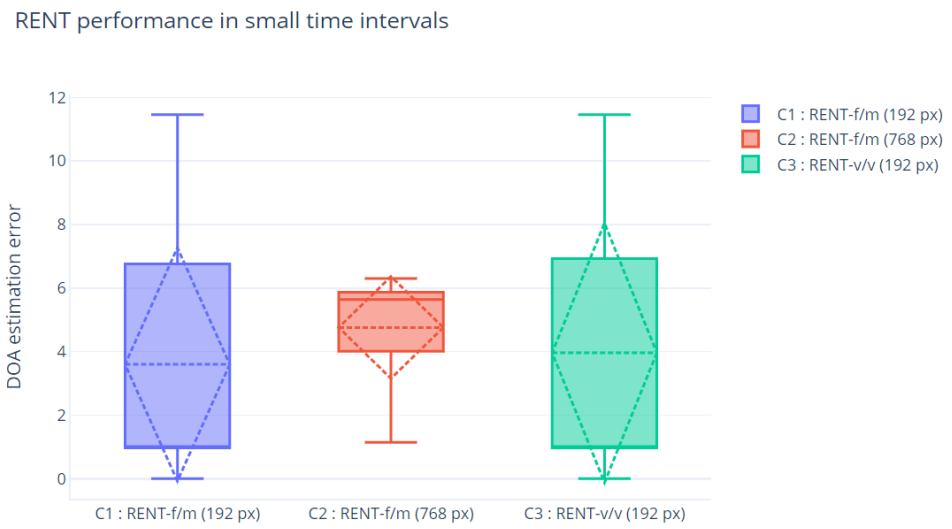


Figure 4.15: Mean and standard deviation for the RENT algorithm parameters over different scene configurations and resolution levels.

in 192 pixels instead of 768 pixels does not generate significant estimation error differences. Besides, we obtain more accurate results than high-level $K_{\max} = 3$ dic-

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S2	S1	S2	S1	S2	S1
Interval-1 (0.5s)	0.97°	0.97°	5.67°	5.67°	0.97°	0.97°
Interval-2 (0.5s)	0.97°	0.97°	5.67°	5.67°	0.97°	0.97°
Interval-3 (0.5s)	0.97°	0.97°	5.67°	5.67°	0.97°	0.97°

Table 4.4: Scenario 1 - Sources are located in elevation and azimuth $\{(120°, 270°), (120°, 90°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S3	S4	S3	S4	S3	S4
Interval-1 (0.5s)	7.10°	11.46°	6.07°	2.17°	11.46°	7.10°
Interval-2 (0.5s)	7.10°	4.92°	6.07°	2.17°	11.46°	7.10°
Interval-3 (0.5s)	7.10°	4.92°	6.07°	2.17°	11.46°	7.10°

Table 4.5: Scenario 2 - Sources are located in elevation and azimuth $\{(112°, 314°), (72°, 270°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S6	S5	S6	S5	S6	S5
Interval-1 (0.5s)	0.97°	0.97°	6.30°	4.70°	0.97°	0.97°
Interval-2 (0.5s)	0.97°	0.97°	6.30°	4.70°	0.97°	0.97°
Interval-3 (0.5s)	0.97°	0.97°	6.30°	4.70°	0.97°	0.97°

Table 4.6: Scenario 3 - Sources are located in elevation and azimuth $\{(60°, 270°), (60°, 90°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

tionary for some of the cases. However, estimation error variance of the 192-pixel RENT is higher than 768 pixel, as DOA estimation errors exceeding $10°$ would be observed in some cases. Using anechoic instrument recordings instead of speech to

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S8	S7	S8	S7	S8	S7
Interval-1 (0.5s)	0.97°	11.46°	6.07°	5.67°	0.97°	11.46°
Interval-2 (0.5s)	0.97°	11.46°	6.07°	5.67°	0.97°	11.46°
Interval-3 (0.5s)	0.97°	11.46°	6.07°	5.67°	0.97°	11.46°

Table 4.7: Scenario 4 - Sources are located in elevation and azimuth $\{(67°, 225°), (60°, 90°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S10	S9	S10	S9	S10	S9
Interval-1 (0.5s)	0.97°	6.76°	3.32°	4.70°	0.97°	6.76°
Interval-2 (0.5s)	0.97°	6.76°	3.32°	4.70°	0.97°	6.76°
Interval-3 (0.5s)	0.97°	6.76°	3.32°	4.70°	0.97°	6.76°

Table 4.8: Scenario 5 - Sources are located in elevation and azimuth $\{(90°, 45°), (60°, 270°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

	RENT-f/m (192px)		RENT-f/m (768px)		RENT-v/v (192px)	
	S12	S11	S12	S11	S12	S11
Interval-1 (0.5s)	0.0°	4.92°	5.61°	1.14°	0.0°	4.92°
Interval-2 (0.5s)	0.0°	4.92°	5.61°	1.14°	0.0°	4.92°
Interval-3 (0.5s)	0.0°	4.92°	5.61°	1.14°	0.0°	4.92°

Table 4.9: Scenario 6 - Sources are located in elevation and azimuth $\{(67°, 315°), (60°, 180°)\}$. DOA estimation errors are given for speech or instrument mixes in each interval using $K_{\max} = 2$, and 3

generate the scene also does not result in significant effects over the DOA estimation performance. Including five scenarios result in the same outputs for different source types, and only one scenario's output is different. Fig. 4.15 shows the mean

and standard deviation of the DOA estimation error according to the selected algorithm parameters. The mean and standard deviation calculated for C1, C2, and C3 are given as $3.59 \pm 3.67, 4.75 \pm 1.60, 3.96 \pm 4.14$, respectively. As a result, it is noted that the spread of the error would increase by decreasing the pixelization level of the sparse dictionary. However, it does not affect the DOA estimation error average significantly as the average value is smaller than higher pixelization angle (the angle between two neighbouring pixels). By using this fact, $K_{\max} = 2$ is selected to obtain lower computational complexity in the RENT-based algorithms.

The fourth evaluation : investigates the effect of sensor noise over DOA estimation with RENT. The same six scenarios (given in Fig. 4.12, 4.13, 4.14) are used to observe performance changes when white Gaussian noise is added to each Eigenmike em32 microphone signal [108]. Fig. 4.4 shows the cases which are original and noise (0 dB, 10 dB, 20 dB and 30 dB) inserted acoustic scenes. The displayed scenes are obtained via omnidirectional component.

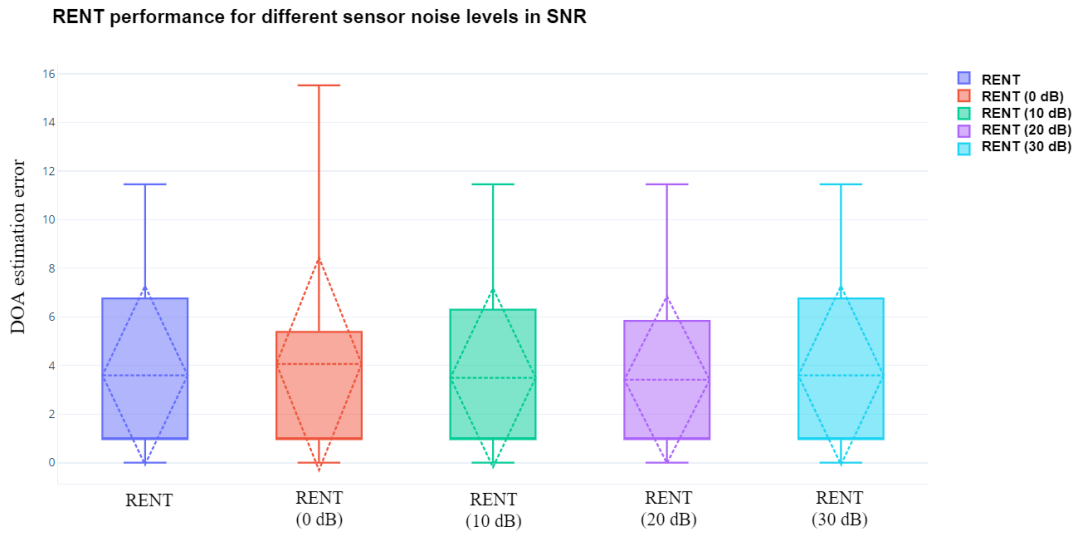


Figure 4.16: RENT DOA estimation errors given in different levels of sensor noise

Tables 4.10, 4.11, 4.12, 4.13, 4.14 and 4.15 represent the DOA estimation errors for the specified scenarios with 0 dB, 10 dB, 20 dB and 30 dB sensor noise. From the measurements, it is seen that the algorithm is robust to 10 dB sensor noise insertion except for 1 case where the algorithm can not estimate the DOA accurately. Besides, 0 dB sensor noise addition results in 3 cases where the DOA can not be estimated.

Fig. 4.16 represents the distribution of the DOA estimation error in the conditions where different levels of sensor noise is inserted to the microphone signals. The found DOA estimation errors for the conditions which contains 0 dB, 10 dB, 20 dB and 30 dB sensor noise are $4.06 \pm 4.37, 3.49 \pm 3.73, 3.41 \pm 3.43$ and 3.59 ± 3.57 respectively.

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S1	S2	S1	S2	S1	S2	S1	S2
Interval-1 (0.5s)	0.97°	-	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°
Interval-2 (0.5s)	0.97	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°
Interval-3 (0.5s)	15.53	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°

Table 4.10: Scenario 1 - Sources are located in elevation and azimuth $\{(120°, 270°), (120°, 90°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S4	S3	S4	S3	S4	S3	S4	S3
Interval-1 (0.5s)	4.92°	-	4.92°	-	4.92°	7.10°	7.10°	11.46°
Interval-2 (0.5s)	4.92°	13.29°	4.92°	7.10°	4.92°	7.10°	7.10°	4.92°
Interval-3 (0.5s)	4.92°	13.29°	4.92°	7.10°	4.92°	7.10°	7.10°	4.92°

Table 4.11: Scenario 2 - Sources are located in elevation and azimuth $\{(112°, 314°), (72°, 270°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S6	S5	S6	S5	S6	S5	S6	S5
Interval-1 (0.5s)	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°
Interval-2 (0.5s)	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°
Interval-3 (0.5s)	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97°	0.97

Table 4.12: Scenario 3 - Sources are located in elevation and azimuth $\{(60°, 270°), (60°, 90°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

Up to this part, two DOA estimation methods are proposed and validated to analyze the acoustic scene using rigid spherical microphone arrays. HiGRID outperforms in

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S8	S7	S8	S7	S8	S7	S8	S7
Interval-1 (0.5s)	4.92°	-	11.46°	0.97°	11.46°	0.97°	11.46°	0.97°
Interval-2 (0.5s)	11.46°	0.97°	11.46°	0.97°	11.46°	0.97°	11.46°	0.97°
Interval-3 (0.5s)	11.46	0.97°	11.46°	0.97°	11.46°	0.97°	11.46°	0.97°

Table 4.13: Scenario 4 - Sources are located in elevation and azimuth $\{(67°, 225°), (60°, 120°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S10	S9	S10	S9	S10	S9	S10	S9
Interval-1 (0.5s)	6.76°	0.97°	0.97°	6.76°	0.97°	6.76°	0.97°	6.76°
Interval-2 (0.5s)	6.76°	0.97°	0.97°	6.76°	0.97°	6.76°	0.97°	6.76°
Interval-3 (0.5s)	6.76°	0.97°	0.97°	6.76°	0.97°	6.76°	0.97°	6.76°

Table 4.14: Scenario 5 - Sources are located in elevation and azimuth $\{(90°, 45°), (60°, 270°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

	RENT (0 dB)		RENT (10dB)		RENT (20dB)		RENT(30dB)	
	S12	S11	S12	S11	S12	S11	S12	S11
Interval-1 (0.5s)	4.92°	0.0°	4.92°	0.0°	0.0°	4.92°	0.0°	4.92°
Interval-2 (0.5s)	4.92°	0.0°	4.92°	0.0°	0.0°	4.92°	0.0°	4.92°
Interval-3 (0.5s)	4.92°	0.0°	11.46°	0.0°	0.0°	4.92°	0.0°	4.92°

Table 4.15: Scenario 6 - Sources are located in elevation and azimuth $\{(67°, 315°), (60°, 180°)\}$ with 0 dB, 10 dB, 20 dB and 30 dB sensor noise

the DOA estimation accuracy as well as providing lower computational complexity according to the full grid search algorithms. On the other hand, RENT provides an average DOA estimation error of $5°$, and it can be used to identify time-frequency bins containing single sources. Providing low DOA estimation error with respect to the minimum audible angle [115] is significant for the development of source separation methods that will use the spatial information of the sources. In the next subsections,

the practical usage of spatial information in short time intervals will be validated by aiming to generate robust SPWD-based source separation algorithms.

4.3.3 Source Separation via Sparse Plane Wave Decomposition with Fixed Spatial Filtering (SPWD-FSF)

The evaluation setup consists of two different scenarios for different source counts. Multichannel acoustic impulse responses (AIRs) are used from the METU SPARG dataset [7] to validate the proposed algorithm. The anechoic recordings contain six tracks [112] which are 5 seconds of anechoic recordings used in the simulations. They are *Mi tradi quell' alma ingrata*, from Mozart's *Don Giovanni* [soprano (S1), cello (S2), violin 1 (S3), violin 2 (S4), clarinet (S5), and viola (S6)]. The loudness level is equalized, according to EBU R128 [116]. In the first scenario [given in Fig. 4.17], the sources are in the front half-space of the microphone array, and two (S1 and S2) to five (S1-S5) sources were evaluated in the corresponding scenario. The array is positioned at a 1.5 m height and the height of the sources are set as 1.8 m, 0.9 m, 1.2 m, 1.2 m, and 0.9 m for S1, S2, S3, S4 and S5, respectively. This simulation is generated as a more realistic recording scenario for a quartet by considering sound sources' specific positions. In the second scenario [given in Fig. 4.18], the sources are placed in the full azimuth covered space of the microphone array, and these cases include two (S1 and S2) to six (S1-S6) sources to be separated and evaluated. The array and the sources are placed at the same 1.5 m height.

Table 4.16: Source separation metrics for maximally directive beamformer and the proposed method (Scenario 1)

	Avg. SIR [dB]		**Avg. SDR [dB]**		**Avg. SAR [dB]**	
#	**maxDF**	**Prop.**	**maxDF**	**Prop.**	**maxDF**	**Prop.**
2	6.59	13.72	4.26	3.95	9.10	4.64
3	6.35	14.31	2.87	3.92	6.86	4.56
4	3.26	12.00	0.67	2.64	6.01	3.50
5	0.96	8.98	-1.27	0.80	5.37	2.13

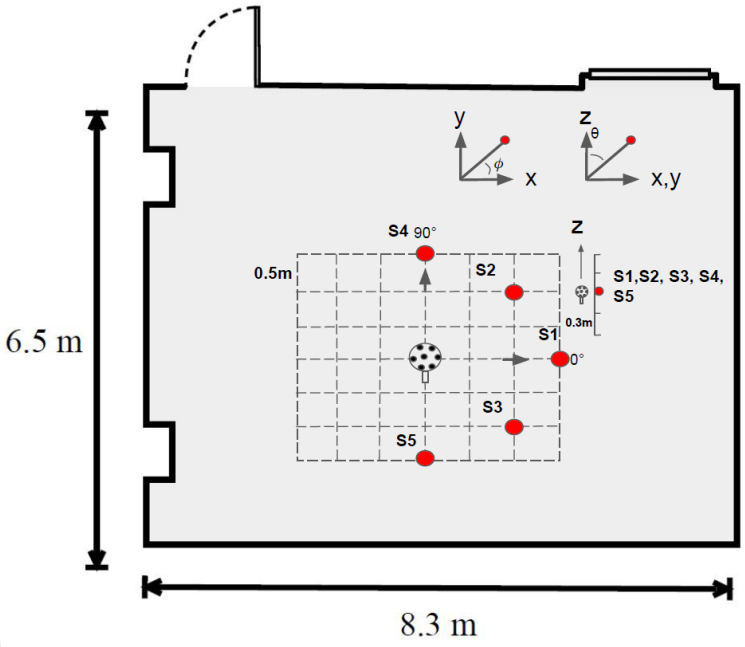


Figure 4.17: Top views of the two scenarios used in the evaluations. Red circles represent source positions. RSMA is positioned at the center.

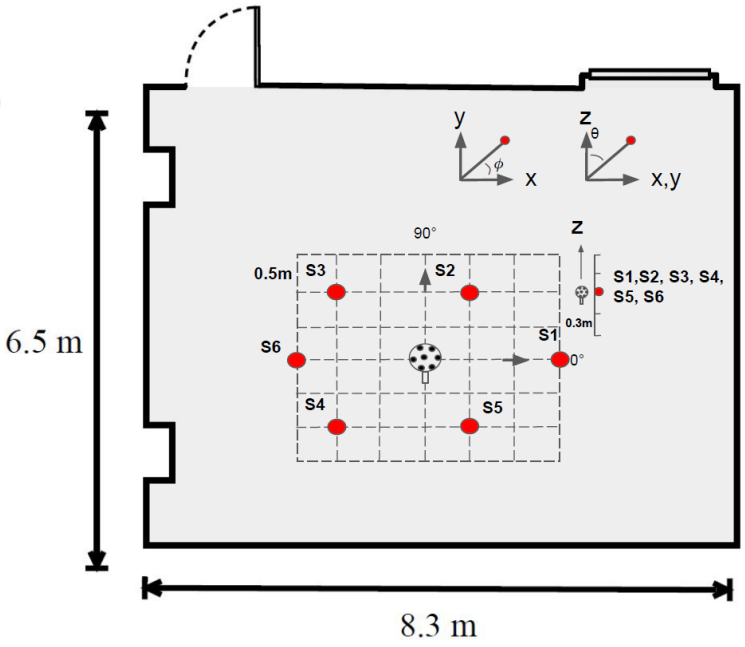


Figure 4.18: Top views of the two scenarios used in the evaluations. Red circles represent source positions. RSMA is positioned at the center.

Table 4.17: Source separation metrics for maximally directive beamformer and the proposed method (Scenario 2)

	Avg. SIR [dB]		Avg. SDR [dB]		Avg. SAR [dB]	
#	**maxDF**	**Prop.**	**maxDF**	**Prop.**	**maxDF**	**Prop.**
2	6.07	16.42	4.09	4.45	10.34	5.21
3	3.67	14.17	1.37	4.23	6.88	5.00
4	2.25	11.72	-0.42	2.68	5.06	3.59
5	1.07	9.72	-1.50	1.74	4.47	2.85
6	-0.79	5.78	-2.98	-0.56	4.61	1.66

The SPWD-FSF is used on a time-frequency bin basis after applying short-time Fourier transform with 1024-point FFTs and 50% overlap over the time domain microphone recordings. The dictionary used for the sparse recovery consists of Legendre kernels evaluated over 192 near-uniformly sampled points on HEALPix grid [86]. The center directions of the Legendre kernels are calculated at the same 192 points. OMP was executed for ten iterations at each time-frequency bin to obtain best-fitting pixels. It should be noted that the Algorithm 4 would also be executed up to the ratio between the residual energy and total energy satisfies a smaller value than the given threshold. Simulations were executed with $\kappa = 10$. For comparison, MaxDF steered in the source DOA directions are used. Reference signals used in the evaluation were obtained via simulating as if there is a single source in the room and the omnidirectional component is used. The mixture containing four sources, the reference signal for S3 (Violin 1), S3 obtained via beamforming, and S3 obtained via the proposed method, are given in Fig. 4.19.

Table 4.16 shows SIR, SDR and SAR for the tested cases in Scenario 1, calculated using BSSEval Toolbox [99]. The results indicate that the proposed method outperforms in a significantly better average SIR and a noticeably better average SDR performance than the maximum directivity beamformer. Besides, the MaxDF provided a better SAR performance. Both SIR and SDR decrease with the increasing number of sources as it would be expected in a complicated scenario.

The same source separation metrics are given in Table 4.17 for the cases given in

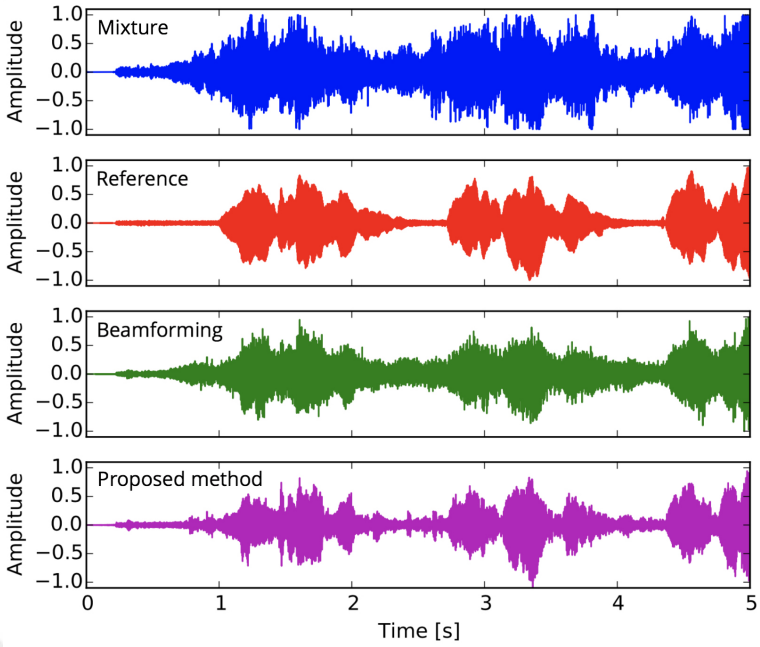


Figure 4.19: Waveforms of the reverberant mixture (blue), reference (red), maximally directive beamformer output (green) and the output of the proposed method (magenta) for the case with four sources (S1-S4) in Scenario 2.©IEEE 2018 [4]

Scenario 2. The results are similar to the results obtained for Scenario 1 except for the two source case where the MaxDF output has a better SDR performance. As with the first scenario, both SIR and SDR decrease with the increasing number of sources.

4.3.4 Sparse Plane Wave Decomposition with Adaptive Spatial Filtering (SPWD-ASF)

METU SPARG AIR dataset is also used in the emulations created for evaluating SPWD-ASF [7]. The test signals used are male and female speech signals, from tracks 49, 50 of the SQAM (Sound Quality Assessment Material) CD [114].

The evaluation setup of the proposed method in comparison with 1) MaxDF in the true DOAs of sources [117] and 2) SPWD-FSF is presented. Both the MaxDF and the fixed spatial filtering approaches require DOA as an input, while the proposed method estimates the source direction and separates the sources. The evaluation is based on SIR, and PESQ scores for eight different scenarios with two sources each.

Table 4.18: The source DOAs and D/R ratios for the emulated cases used in the evaluation.

	C1 (0-1 dB)		C2 (4-6 dB)	
	Source 1, Source 2		Source 1, Source 2	
	DOA $[\theta, \phi](^\circ)$	D/R(dB)	DOA $[\theta, \phi](^\circ)$	D/R(dB)
S1	[98,315],[90,225]	-0.2,-0.3	[67,225],[105,116]	7.8,2.9
S2	[59,90] , [67,315]	7.9,7.9	[130,135], [121,270]	2.1,6.3
S3	[100,45], [100,251]	2.3,2.2	[67,315] , [75,153]	7.9,3.7
S4	[113,45],[90,296]	6.3,5.9	[78,270],[71,123]	3.4,-0.8

Table 4.19: DOA estimation errors

	C1 (0-1 dB)		C2 (4-6 dB)	
	Source 1	**Source 2**	**Source 1**	**Source 2**
Scenario 1	1.09°	6.78°	6.39°	11.47°
Scenario 2	4.95°	0.96°	0.96°	11.31°
Scenario 3	4.20°	1.68°	7.52°	4.95°
Scenario 4	7.18°	11.47°	0.88°	1.71°

Scenarios are grouped into two categories: C1 has similar D/R ratios within 0 to 1 dB, and C2 has D/R ratios within 4 to 6 dB. Table 4.18 shows the directions of the emulated sources and their D/R ratios.

The recordings used in the evaluation were obtained by convolving anechoic speech signals with acoustic impulse responses (AIR) from METU SPARG AIR dataset. The algorithm parameters are given as the same for the previous evaluation. The spatial filter was estimated over intervals of 500 ms with 50% s overlap. The reference signals for the computation of SIR and PESQ scores were obtained by convolving the anechoic speech signals with single AIR related to the source. An omnidirectional response was calculated from the zeroth-order spherical harmonic components.

Table 4.19 shows the DOA estimation errors averaged across the whole duration of the tested scenarios. It may be observed that these errors are generally less than the pixel

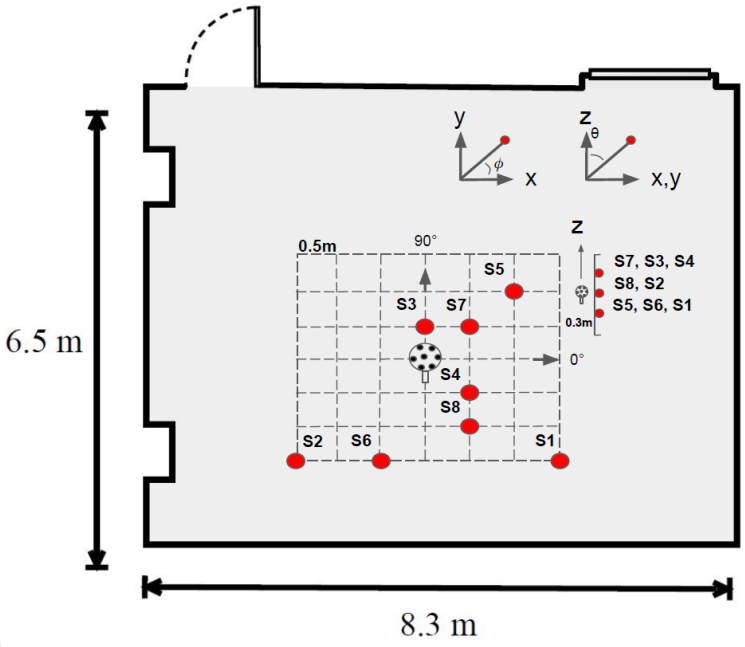


Figure 4.20: Scenarios (C1) for the evaluation of source separation performances.{Scenario 1(S1,S2),Scenario 2(S3,S4),Scenario 3(S5,S6),Scenario 4(S7,S8)}

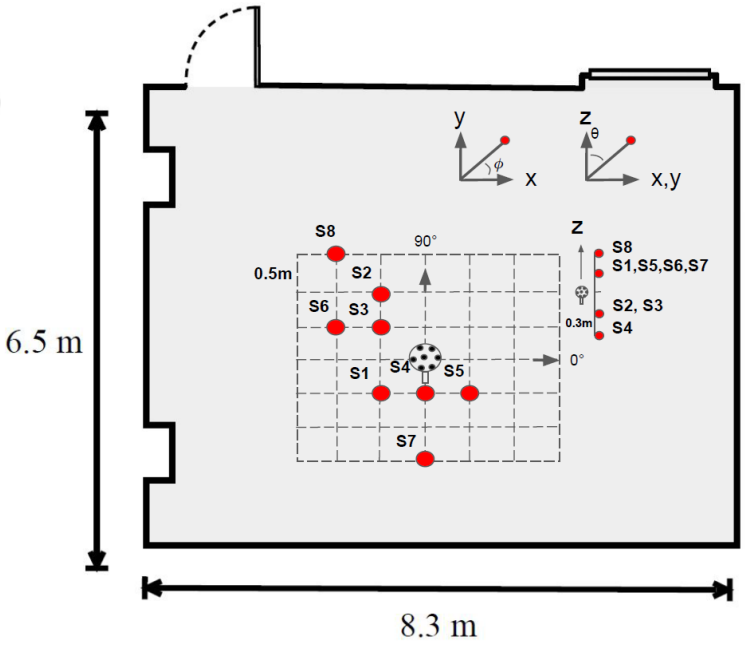


Figure 4.21: Scenarios (C2) for the evaluation of source separation performances.{Scenario 1(S1,S2),Scenario 2(S3,S4),Scenario 3(S5,S6),Scenario 4(S7,S8)}

Table 4.20: SIR and PESQ scores (Mean $\pm$ Standard deviation)

	C1 (0-1 dB)		**C2 (4-6 dB)**	
	SIR (dB)	**PESQ**	**SIR (dB)**	**PESQ**
Reference	∞	4.5	∞	4.5
MaxDF	10.9 ± 3.3	1.8 ± 0.2	11.8 ± 3.4	1.9 ± 0.1
SPWD-FSF [4]	16.6 ± 3.4	1.8 ± 0.2	18.9 ± 5.6	1.9 ± 0.2
SPWD-ASF	19.6 ± 3.2	1.6 ± 0.2	21.9 ± 3.8	1.7 ± 0.2

resolution, $\theta_{\text{pix}} = 14.7°$, of the $M = 192$ element HEALPix grid order, indicating a good DOA estimation accuracy. Table 4.20 shows the source separation performance indicators for the two scenarios. The results show that the proposed approach achieves noticeable SIR improvements over the compared methods, especially when the D/R ratio differences are high.

On average, the method proposed provides 9.7 dB and 3.0 dB SIR improvements in C1, 10.1 dB, and 3.0 dB SIR improvements in C2, over the MaxDF beamforming and fixed-beamwidth spatial filtering methods, respectively. While the PESQ score is slightly lower for the proposed method than fixed-beamwidth spatial filtering in C2, the difference is not massive, and the separated speech still has a PESQ score representative of acceptable quality.

4.3.5 Evaluation of the Wiener post-filtering (WPF)

The effect of Wiener post-filtering presented in Section 3.3 is evaluated with three different setups comprising $S = 2$, $S = 3$, and $S = 4$ sources. The metrics used for the physical and perceptual evaluation criteria are selected as SDR, SIR, SAR, Speech Intelligibility (SI), and PESQ. The comparisons are performed in 2 algorithms that use RENT as a common source DOA estimation block. These methods are MaxDF and SPWD-FSF where DOA is estimated via RENT for all. The effect of WPF by applying to the proposed methods as a post-processing stage will be investigated using the evaluation metrics.

The evaluation setup consists of multi-channel acoustic impulse response (AIR)

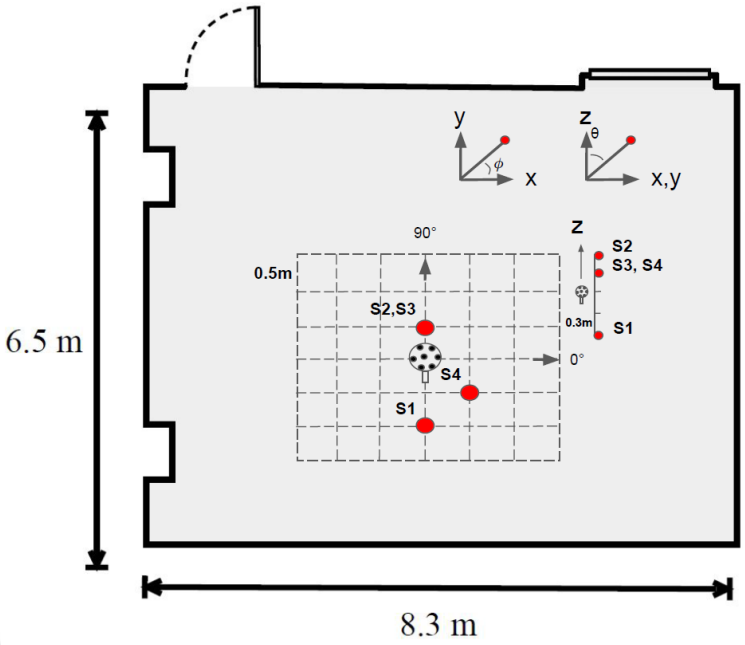


Figure 4.22: Performance setup consists of two scenarios and each scenario contains S=2 sources.{Scenario 1(S1,S2), Scenario 2(S3,S4)}

recordings from the METU SPARG AIR dataset [7]. AIR recordings are selected randomly. The test signals used are speech signals, which are from track 49, 50, 51, 52 of the SQAM (Sound Quality Assessment Material) CD [114]. The scenarios' duration is 5 seconds, and the microphone array recordings in the number of 2, 3 and 4 sources are separated with the MaxDF, and SPWD-FSF methods. WPF is applied to the outputs of these separation blocks. The scenarios are given in Figures 4.22, 4.23, 4.24.

The frequency range selected for the analysis of RENT is between 0 Hz and 9000 Hz. FFT window size and overlap are 2048 and 50%, respectively. The sampling rate is selected as 48 kHz. Source separation covers the full frequency range. The minimum angle required for matching each frame's DOA results to the previous ones is $\pi/4$. Reference signals for the computation of SDR, SIR, SAR [99], and SI [107] are calculated by averaging the microphone array recordings via dividing with the mode strength component as in the condition where there is only one single source active in the same position.

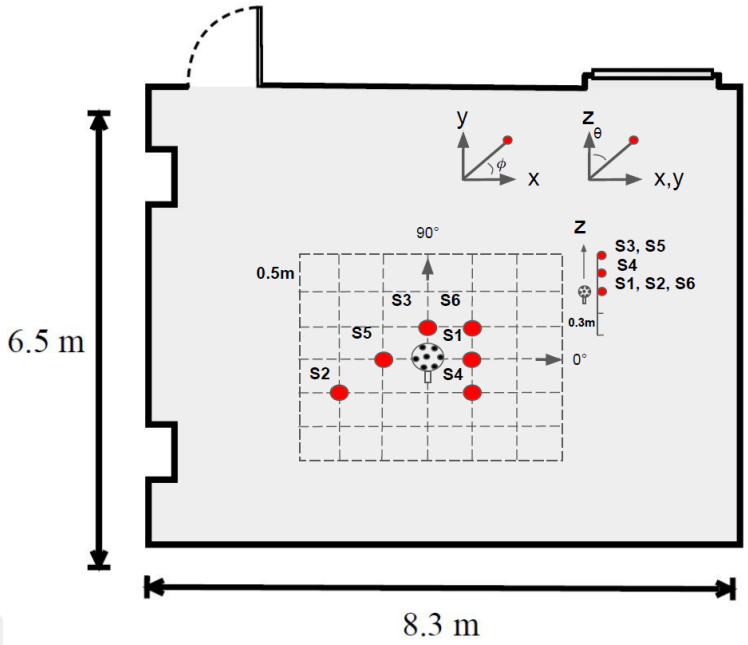


Figure 4.23: Performance setup consists of two scenarios and each scenario contains S=3 sources.{Scenario 1(S1, S2, S3), Scenario 2(S4, S5, S6)}

The evaluation results are given in the Tables 4.21, 4.22 and 4.23. The results observed for the WPF effect over SDR, SIR, SAR, and SI are positive. Weighting time-frequency bins using WPF improves the SIR separation performance by 8-9 dB for the MaxDF, and 4-5 dB average for the SPWD-FSF. SDR and SAR are also slightly improved. The results indicate that WPF does not significantly improve the SI measurements and SI remains at a perceptually acceptable level.

S=2 Case	Scenario 1 (Avg. dB)				Scenario 2 (Avg. dB)			
	SDR	SIR	SAR	SI	SDR	SIR	SAR	SI
MaxDF	2.43	12.93	3.07	0.75	3.57	12.25	4.61	0.82
MaxDF + WPF	3.98	20.96	4.18	0.78	5.28	23.48	5.43	0.83
SPWD-FSF	3.54	18.75	3.73	0.76	5.04	19.55	5.24	0.83
SPWD-FSF + WPF	4.34	22.43	4.44	0.77	5.50	26.74	5.58	0.83

Table 4.21: Evaluation metrics of 4 algorithms for the scenarios having 2 sources.

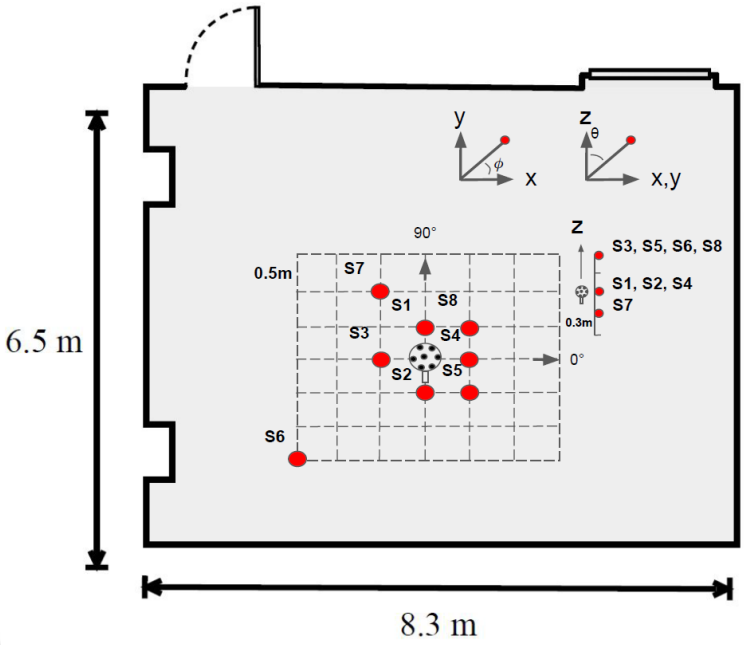


Figure 4.24: Performance setup consists of two scenarios and each scenario contains S=4 sources.{Scenario 1(S1,S2,S3,S4), Scenario 2(S5,S6,S7,S8)}

S=3 Case	Scenario 1 (Avg. dB)				Scenario 2 (Avg. dB)			
	SDR	SIR	SAR	SI	SDR	SIR	SAR	SI
MaxDF	1.21	10.11	2.49	0.78	0.99	5.88	4.29	0.78
MaxDF + WPF	3.77	18.97	3.96	0.79	4.39	16.78	4.78	0.79
SPWD-FSF	2.90	15.23	3.36	0.80	2.99	13.19	3.83	0.79
SPWD-FSF + WPF	4.00	21.33	4.12	0.79	4.21	19.21	4.42	0.78

Table 4.22: Evaluation metrics of 4 algorithms for the scenarios having 3 sources.

4.3.6 The Joint Source Separation Results from Real Recordings

This section consists of separation for a recorded pre-concert rehearsal of Nemeth Quartet (a classical quartet consisting of two violins, a viola, and a cello) that was given in the HiGRID DOA estimation results section. The program material used

S=4 Case	Scenario 1 (Avg. dB)				Scenario 2 (Avg. dB)			
	SDR	SIR	SAR	SI	SDR	SIR	SAR	SI
MaxDF	0.00	3.58	4.61	0.75	-2.43	3.00	1.09	0.71
MaxDF + WPF	3.63	11.94	4.61	0.78	0.01	10.86	1.18	0.74
SPWD-FSF	2.59	10.34	4.00	0.80	-0.55	8.64	0.72	0.74
SPWD-FSF + WPF	3.85	14.23	4.57	0.78	0.65	13.65	1.16	0.74

Table 4.23: Evaluation metrics of the 4 algorithms for the scenarios having 4 sources.

in the evaluation is a ten-second excerpt from the third movement of Ludwig van Beethoven's String Quartet Nr. 11, Op. 95. All instruments are playing in this section. The recording setup is given in Fig. 4.9. The duration of the recorded scene to be analyzed is 10 seconds.

SPWD-FSF + WPF algorithm is used for obtaining the separation results. The estimated instruments' DOAs via RENT are given for Viola, Violin 2, Violin 1 and Cello as $(110°, 348°)$,$(110°, 33°)$,$(110°, 78°)$, $(132°, 280°)$, respectively. The acoustic scenes and spectrograms obtained using the SPWD-FSF + WPF algorithm blocks are also given in Fig. 4.25. The spectrogram of Cello is given blank up to 0.5 s because the source is not active at this duration, and RENT can not identify it. Since, the reference signals are not available, the objective metrics as SIR, SDR, SAR, and PESQ can not be measured. However, informal listening tests indicate that the source separation performance is very good.

Another scenario contains separation results for Eigenmike recording of a group of musicians playing in a large, open barn in Vermont [118]. The recording is reported to be made by positioning the Eigenmike em32 RSMA at a central position between the musicians. The duration of the recording to be used in the separation is 10 seconds. The musicians' positions can be revealed from the Fig. 4.26 with the viewing angle icon indicated via red rectangle on the right hand side of each image.

SPWD-ASF + WPF is used for the separation. The estimated DOAs by RENT are $(120°, 0°)$,$(100°, 131°)$, $(100°, 180°)$, and $(120°, 270°)$ in elevation and azimuth. The acoustic scenes and spectrograms obtained using SPWD-ASF + WPF algorithm

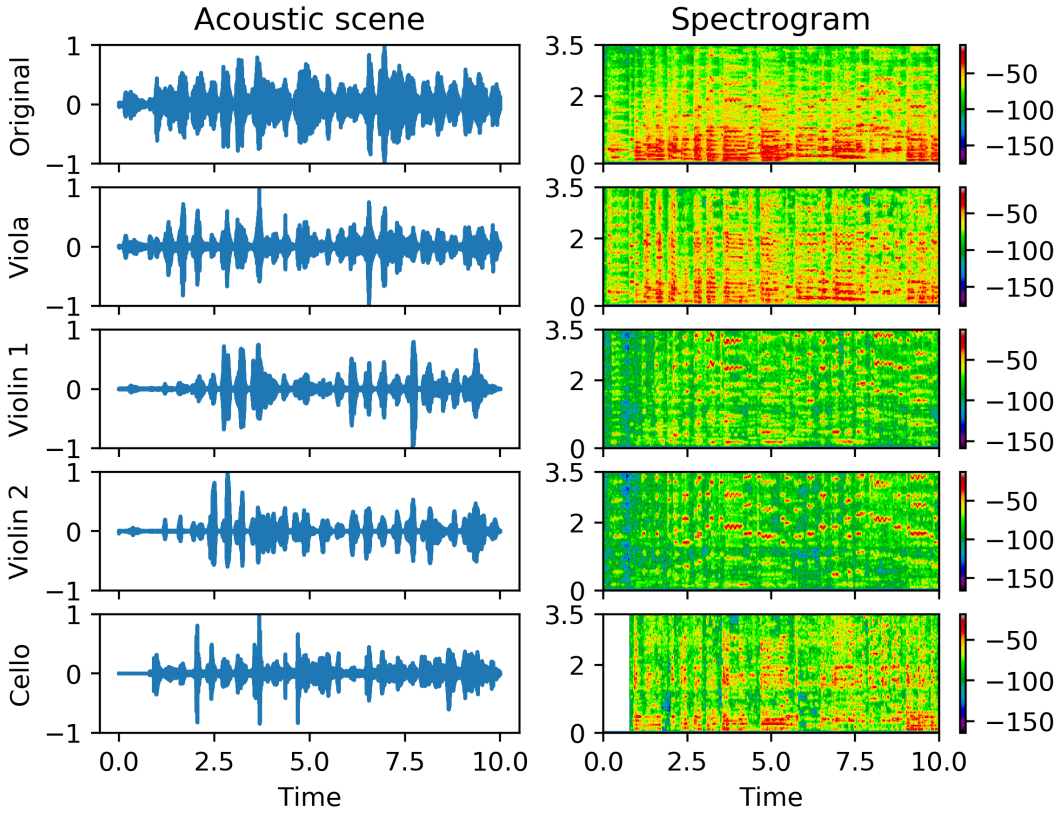


Figure 4.25: Waveforms and magnitude of spectrograms of the original Nemeth Quartet recording and separated signals. Frequency region selected for the visualization at the spectrograms is in the range of 0 to 3500 Hz. The duration of each sample is 10 seconds.

blocks are also given in Fig. 4.27. The objective metrics such as SIR, SDR, SAR, and PESQ can not be calculated due to non-existence of the reference.

4.4 Statistical Evaluation of Source Separation Performance

This section involves emulated speech mixtures using AIR recordings. We used common source separation objective metrics to make a comparison between different methods. The configuration consists of two sets of evaluations: The first set of evaluations presents several different emulations to observe different algorithms evaluated

Figure 4.26: Eigenmike recording of a group of musicians playing in a large, open barn in Vermont. Red rectangle indicates the viewing angle range from $0°$ to $270°$.

for different source counts (S=2, S=3, and S=4) under noise-free conditions. The robustness of the same algorithms to additive noise is also evaluated for different levels of sensor noise 0, 10, 20, and 30 dB in scenes comprising two sources.

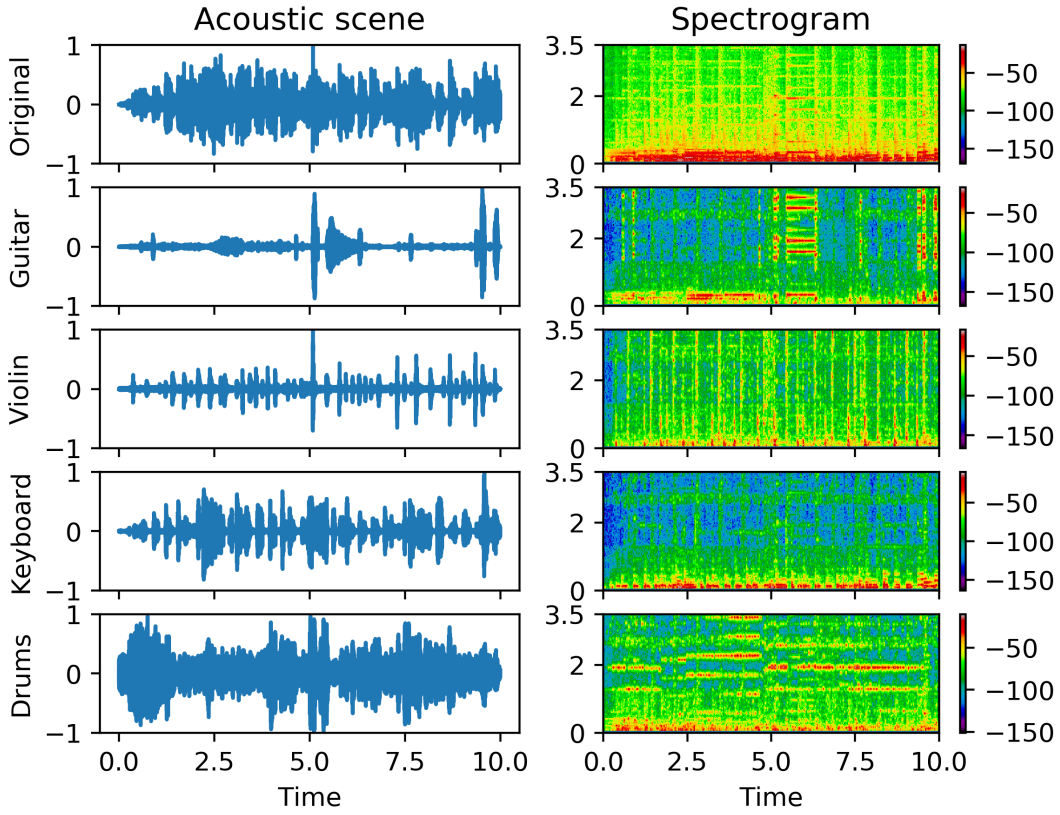


Figure 4.27: Waveforms and magnitude of spectrograms of the original Showdown Showcase recording and separated signals. Frequency region selected for the visualization at the spectrograms is in the range of 0 to 3500 Hz. The duration of each sample is 10 seconds.

4.4.1 Acoustic scene emulations

The microphone array signals used in the objective evaluation were obtained by convolving anechoic speech signals with randomly selected acoustic impulse responses from METU SPARG AIR dataset [7]. The employed test signals are anechoic male and female speech signals selected from EBU SQAM (Sound Quality Assessment Material) Audio CD [114].

Each sources from the scenario were randomly selected and 10 scenarios are created that consist of 540 separated sources for the noise-free case and noisy case comprises

480 separated sources. This way, different types of scenarios having a variety of source DOAs and D/R ratios were obtained. The emulated AIRs are selected by satisfying the minimum angle difference of $\pi/4$. Each of the tested cases had a 4 s duration.

4.4.2 Methods in the statistical scenario

Several methods are used for comparison to assess the relative importance of different stages for the proposed approach. The basic methods are MaxDF [90], SPWD-FSF [4], and SPWD-ASF as proposed in this thesis. Two different versions, with and without Wiener post-filtering (WPF) are evaluated. In all cases the common DOA estimation method is selected as RENT.

The employed steering direction count and the number of kernels are selected as $M = D = 192$. DOAs are estimated for every intervals of length $\tau_w = 500$ ms with 50% s overlap. The dilation parameter in the adaptive approach was selected as $\varrho = 1$, and the spread parameter for SPWD-FSF was selected as $\kappa = 4$. RENT thresholds were selected as 0.5 and 0.1 for the single source and noise/diffuse field time-frequency groupings to generate a Wiener post-filter.

4.4.3 Objective metrics

After SIR, SDR, SAR, and PESQ were calculated, the statistics indicate that PESQ was positively correlated with the SDR and SAR. The correlations of PESQ with SAR [$r(538) = 0.674, p < 0.001$] and with SDR [$r(538) = 0.617, p < 0.001$] were statistically significant for the noise-free scenarios. Similarly the correlations of PESQ with SAR [$r(538) = 0.649, p < 0.001$] and with SDR [$r(538) = 0.664, p < 0.001$] were statistically significant for the noisy scenarios. This correlation reveals that the conditions having higher PESQ scores would also output higher SAR/SDR and vice versa. According to the correlation between PESQ and SAR/SDR, SIR and PESQ scores are provided for more detailed statistical analysis.

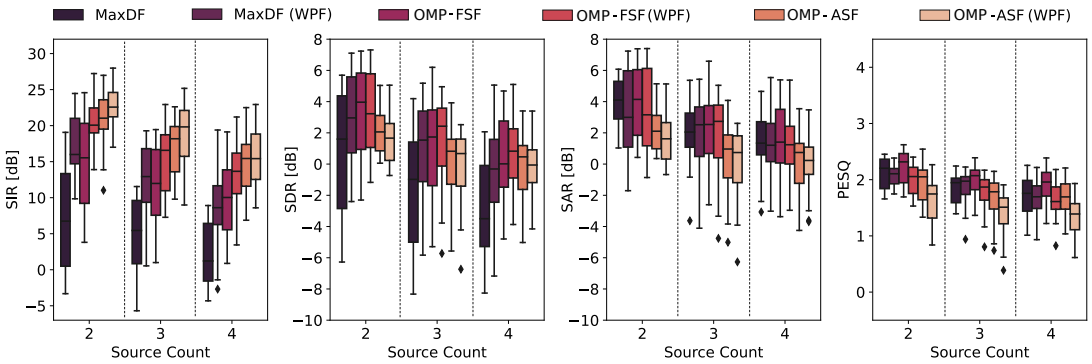


Figure 4.28: SIR, SDR, SAR and PESQ scores for the tested algorithms by using scenarios with different number of randomly selected source positions under ideal conditions with no sensor noise.

4.4.4 Noise-free conditions

The average DOA estimation error was $5.76° \pm 3.60°$, and it is noted that half the grid resolution for the employed HEALPix grid with $M = 192$ pixels is greater than the error. Besides, the calculated RMSE values are $6.49°$, $6.69°$, $7.03°$ for 2, 3 and 4 sources respectively. The observed value is close to the average obtained with the comparisons made before. Multiple comparisons revealed that the number of sources in the scenario did not significantly affect the DOA estimation error. While the source separation task that relies on it, this level of accuracy is sufficient. It is noted that increasing the number of kernels which are encoded in the corresponding dictionary for the candidate DOA directions would have resulted in more robust DOA estimations but the computational cost can be sufficiently increased. For example, selecting $D = 768$ instead of $D = 192$ would have result in a fourfold increase in the computational cost of RENT, despite reducing the DOA estimation error [see in 4.7.2].

SIR, SAR, SDR, and PESQ results are revealed in Fig. 4.28 for the noise-free case by evaluating the scenarios containing different number of sources. Regardless of the source count in the scenario, SPWD-ASF with Wiener post-filtering achieves the highest SIR. In addition, SPWD-ASF with Wiener post-filtering results in the mean SIR scores of 22.75, 18.63, and 15.69 dB for 2, 3, and 4 sources, respectively. The highest mean SDR scores of 3.42, 1.26 and 0.54 dB are provided by SPWD-FSF

with Wiener post-filtering for 2, 3, and 4 sources, respectively. The highest mean SAR scores of 3.97, 2.11, and 1.37 dB are also achieved by SPWD-FSF without Wiener post-filtering for 2, 3, and 4 sources, respectively. Finally, the highest mean PESQ scores of 2.21, 1.99, and 1.91 are observed by SPWD-FSF without Wiener post-filtering for 2, 3, and 4 sources, respectively.

In addition to the scenario-based objective statistical evaluations, acoustical properties associated with each source position and/or scenario had statistically significant correlations with the objective metrics at $p < 0.05$ or lower for all combinations.

In detail, increasing angular difference between the desired source and its nearest interference would improve the objective metric scores. The metrics can also be improved by increasing D/R ratio at the source position, but worsen with increasing DOA estimation error. A statistical model is employed to include all of these covariates.

SIR and PESQ are statistically assessed with the analysis of covariance (ANCOVA) method. The employed statistical models involve the fixed factors that were *Separation method* (METHOD), *Post-filtering* (WPF), and *Source count* (COUNT). We have used the model which includes all main terms as well as two-way interactions of the independent variables and the identified covariates outlined above. The Wiener post-filtering is used as a distinct factor to observe its effect on the results. JASP was used to run the analysis [119].

4.4.4.1 SIR results

The SIR as the dependent variable is analyzed with the ANCOVA model. The observations revealed that WPF [$F(1, 523) = 190.329,\ p < 0.001,\ \eta^2 = 0.107$], METHOD [$F(2, 523) = 285.019,\ p < 0.001,\ \eta^2 = 0.320$], and COUNT [$F(2, 523) = 47.232,\ p < 0.001,\ \eta^2 = 0.053$] were all statistically significant factors. Two-way interactions METHOD×WPF [$F(2, 523) = 32.199,\ p < 0.001,\ \eta^2 = 0.036$] and WPF×COUNT [$F(2, 523) = 4.651,\ p = 0.01,\ \eta^2 = 0.005$] were also statistically significant. The analysis reveals that different methods containing Wiener post-filtering behave differently and the effects of Wiener post-filtering also depend

Table 4.24: Correlations between objective metrics and experimental covariates

Variable		Nearest interf.	D/R ratio	DOA est. err.
SIR	Pearson's r	0.267	0.413	-0.091
	p-value	$< .001$	$< .001$	0.035
SAR	Pearson's r	0.101	0.652	-0.111
	p-value	0.021	$< .001$	0.010
SDR	Pearson's r	0.183	0.658	-0.118
	p-value	$< .001$	$< .001$	0.006
PESQ	Pearson's r	0.122	0.637	-0.006
	p-value	0.004	$< .001$	0.892

on the number of sources.

Among the defined covariates, the D/R ratio at the source position and angular difference with the nearest interference were both statistically significant at $p < 0.001$ level, where the effect size for the latter was $\eta^2 = 0.146$. It demonstrates that increasing the relative level of reverberation would have a degrading effect on the SIR performance. Post-hoc comparisons with Tukey correction disclosed that pairwise differences between all tested methods were statistically significant at $p < 0.001$ level. The ordered SIR performance from the highest to lowest was SPWD-ASF, SPWD-FSF, and MaxDF beamforming, respectively. SIR improvement of SPWD-ASF with respect to SPWD-FSF and MaxDF was observed as 9.58 dB and 3.88 dB. Similarly, SIR improvement of SPWD-FSF with respect to MaxDF was observed as 5.70 dB on average.

Post-hoc comparisons with Tukey correction were also applied to observe the effect of post-filtering varied across different methods. Applying Wiener post-filtering to the MaxDF beamformer's output would achieve an additional SIR improvement of 7.84 dB, whereas observed SIR improvements for SPWD-FSF and SPWD-ASF were 4.16 dB and 1.65 dB, respectively. While the achieved SIR improvements were statistically significant, the addition of post-filtering to the output of SPWD-ASF would result in a redundant extra computational cost because SIR improvement of SPWD-ASF without

post-filtering averaged across all cases is already sufficiently high at 17.18 dB.

Wiener post-filtering was also statistically significant at $p < 0.001$ level with difference between SIR performance. SIR improvement of an additional 4.55 dB is obtained via Wiener post-filtering on average. The pairwise comparisons among the source counts in the scenarios also disclosed that the decrease in SIR was on average 1.78 dB between two- and three-source scenarios, and 2.48 dB between three- and four-source cases.

4.4.4.2 PESQ results

The PESQ as the dependent variable is analyzed with the ANCOVA model. The observations revealed that WPF [$F(1, 523) = 96.663,\ p < 0.001,\ \eta^2 = 0.063$], METHOD [$F(2, 523) = 93.504,\ p < 0.001,\ \eta^2 = 0.121$], and COUNT [$F(2, 523) = 38.867,\ p < 0.001,\ \eta^2 = 0.050$] were statistically significant main effects. Among the two-way interactions, only METHOD$\times$WPF [$F(2, 523) = 25.160,\ p < 0.001,\ \eta^2 = 0.033$] was statistically significant.

The difference between the mean PESQ scores obtained for MaxDF and SPWD-FSF methods were not statistically significant as indicated with post-hoc comparisons with Tukey correction. However, the mean differences for the pairs MaxDF/SPWD-ASF and SPWD-FSF/SPWD-ASF were statistically significant with 0.252 and 0.3. That means SPWD-ASF performance degrades in PESQ less well than the other two methods.

Mean PESQ scores were reduced by applying Wiener post-filtering in 0.189 at $p < 0.001$ level. PESQ scores were also reduced by 0.11 on average per each additional source in terms of source count. While PESQ score was not statistically significantly affected by applying Wiener post-filtering for MaxDF, post-filtering had decreased PESQ (at $p < 0.001$ level) for SPWD-FSF and SPWD-ASF, by 0.238 and 0.321, respectively.

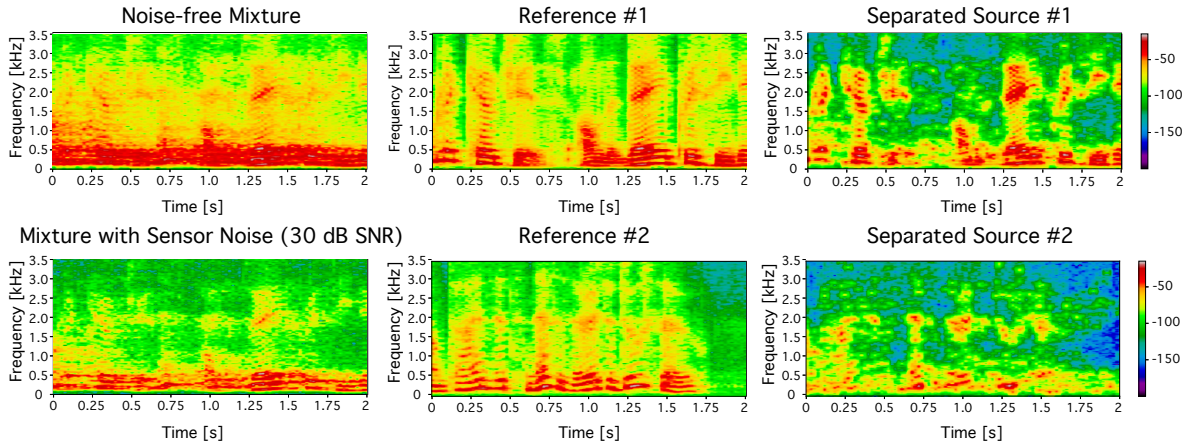


Figure 4.29: Spectrograms of the noise-free mixture, the mixture with additive sensor noise at 30 dB SNR, the two reference signals and the two separated sources obtained using SPWD-ASF with Wiener post-filtering.

4.4.5 Robustness to Sensor Noise

The robustness of different algorithms to sensor noise was investigated by inserting uncorrelated white Gaussian noise to each one of the emulated microphone signals at 0, 10, 20, and 30 dB SNR levels for each of 10 randomly generated two-source scenarios [108]. Fig. 4.29 shows the time-frequency spectrograms of the pressure signal due to ideal mixture, the noisy mixture with 30 dB SNR, the reference signals and the signals separated with SPWD-ASF with Wiener post-filtering. By using RENT as the time-frequency grouping algorithm as a preliminary step for the Wiener post-filtering, Fig. 4.31 represents the time-frequency groupings for the noise-free and 20 dB noise inserted case. The average DOA estimation errors were $7.13^\circ \pm 4.81^\circ$, $5.49^\circ \pm 3.48^\circ$, $5.12^\circ + 3.49^\circ$, and $3.74^\circ + 3.17^\circ$ for 0 dB SNR, 10 dB SNR, 20 dB SNR, and 30 dB SNR, respectively. A better DOA estimation performance was achieved at higher SNR levels in general. The results are given in Table 4.25.

Fig. 4.30 shows the SIR, SAR, SDR and PESQ results for the different SNR levels discussed. The highest SIR score of 21.49 dB is obtained via SPWD-ASF with Wiener post-filtering outputs by averaging across all cases. An average of SIRs 16.65 dB, 21.38 dB, 23.95 dB, and 24.00 dB values for 0, 10, 20, and 30 dB SNR cases were also provided with SPWD-ASF with Wiener post-filtering. Statistical analysis

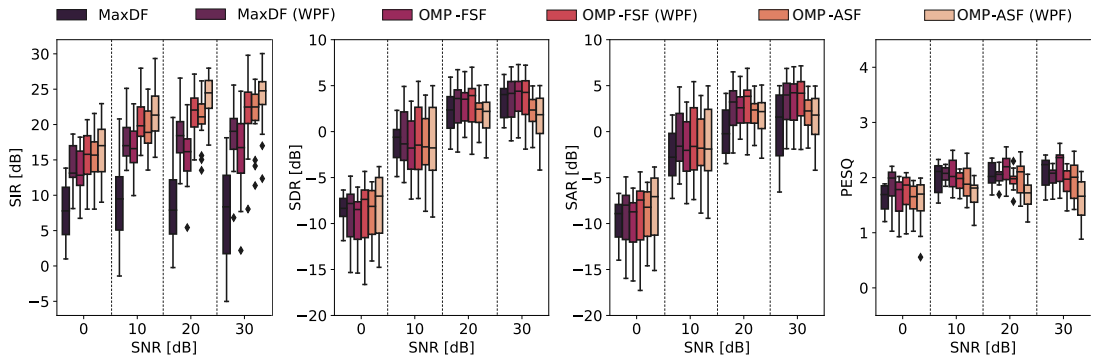


Figure 4.30: SIR, SDR, SAR and PESQ scores for the tested algorithms for scenarios with comprising two randomly selected source positions with 0 dB, 10 dB, 20 dB, and 30 dB SNR, respectively.

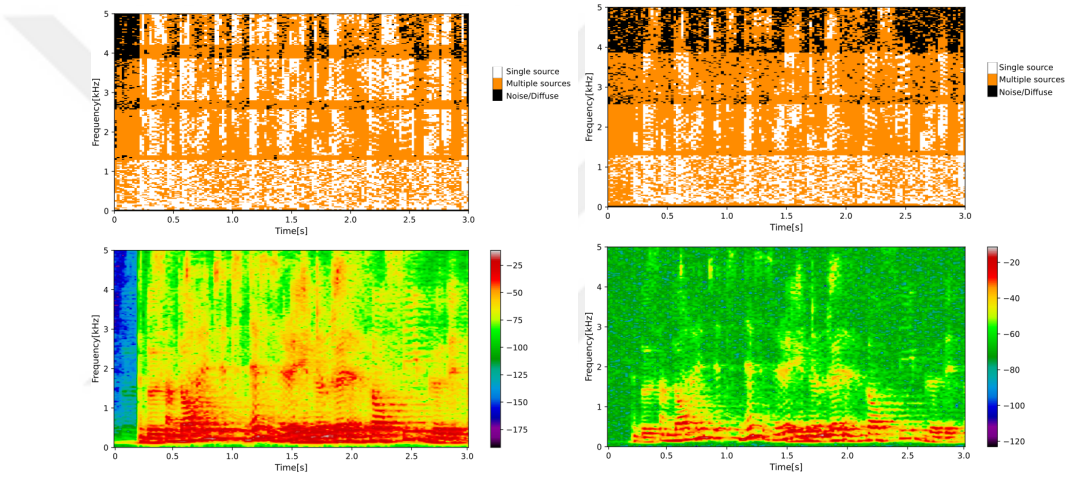


Figure 4.31: Time-frequency groupings obtained with RENT in noise-free (left) and 20 dB sensor noise cases (right). RENT thresholds are selected as 0.5 and 0.1 for single source and noise/diffuse, respectively.

revealed that there was no statistically significant difference between the averaged SDR performances of the tested methods. Suffer in the SDR performance is observed equally across different methods with additive sensor noise. The best mean SAR performance of -0.73 dB, 2.80 dB, and 3.40 dB for 10, 20, and 30 dB SNR cases were obtained via SPWD-FSF with Wiener post-filtering. However, a better performance with a mean SAR of -8.46 dB, only for the 0 dB SNR case can be provided by SPWD-ASF with Wiener post-filtering . Finally, the best averaged PESQ score of 2.02 was observed with SPWD-FSF without Wiener post-filtering.

The dependence of SIR and PESQ on different experimental parameters is estimated by applying Analysis of covariance (ANCOVA) test. The fixed factors were *Post-filtering* (WPF), *Separation method* (METHOD), and *Signal-to-noise ratio* (SNR). The covariances from the noise-free evaluation model were used to include all main terms and two-way interactions.

4.4.5.1 SIR results

The SIR as the dependent variable is analyzed with the ANCOVA model. The observations revealed that WPF [$F(1,459) = 318.381,\ p < 0.001,\ \eta^2 = 0.184$], METHOD [$F(2,459) = 230.874,\ p < 0.001,\ \eta^2 = 0.266$], and SNR [$F(3,459) = 22.190,\ p < 0.001,\ \eta^2 = 0.038$] were all statistically significant factors. Among the two way interactions, METHOD×WPF [$F(2,459) = 46.346,\ p < 0.001,\ \eta^2 = 0.053$], WPF×SNR [$F(3,459) = 5.615,\ p < 0.001,\ \eta^2 = 0.010$], and METHOD×SNR [$F(6,459) = 4.244,\ p < 0.001,\ \eta^2 = 0.015$] were also statistically significant.

It is observed that all pairwise comparisons between different methods were statistically significant at $p < 0.001$ level by post-hoc comparisons with Tukey correction. The best average SIR performance was obtained by SPWD-ASF with Wiener post-filtering similar to the noise-free case, and an SIR improvement of 7.70 dB over MaxDF and 2.77 dB over SPWD-FSF was provided by Wiener post-filtering addition. Similarly, Wiener post-filtering improved the average SIR by 5.23 dB.

The worst SIR performance case is obtained with 0 dB SNR, which is also significantly ($p < 0.001$) worse than all other SNR levels. The SIR differences among the case with 0 dB SNR and 10, 20, and 30 dB SNR cases are -2.78, -3.13, and -3.26 dB, respectively. The data suggests that the average SIR performance is not significantly affected, unless the SNR is as low as 0 dB. We should note that SNR level would affect different methods in the evaluation differently. According to different noise levels, the differences between the mean SIR performance of MaxDF were not statistically significant. Besides, the SIR performances of SPWD-FSF and SPWD-ASF were statistically significantly ($p < 0.001$) different only between the 0 dB SNR case and other cases. This information discloses that the proposed methods are robust to noise in their SIR performances unless the SNR is as low as 0 dB.

As a result, Wiener post-filtering application significantly improves the SIR performances of all methods in the evaluation at $p < 0.001$ level. Application of post-filtering achieves the highest SIR for MaxDF at 9.12 dB, followed by SPWD-FSF at 4.47 dB and SPWD-ASF at 2.27 dB, respectively. SIR improvement with the application of post-filtering was also statistically significant at $p < 0.001$ level for all noise levels. The mean SIR improvements achieved by Wiener post-filtering are 3.40 dB, 5.06 dB, 6.48 dB and 6.22 dB for 0 dB, 10 dB, 20 dB, and 30 dB SNR levels, respectively.

4.4.5.2 PESQ results

The PESQ as the dependent variable is analyzed with the ANCOVA model. The observations revealed that METHOD [$F(2, 459) = 85.503,\ p < 0.001,\ \eta^2 = 0.098$], WPF [$F(1, 459) = 35.429,\ p < 0.001,\ \eta^2 = 0.021$], SNR [$F(3, 459) = 49.881,\ p < 0.001,\ \eta^2 = 0.088$], METHOD $\times$ WPF [$F(2, 459) = 27.583,\ p < 0.001,\ \eta^2 = 0.033$], WPF $\times$ SNR [$F(3, 459) = 27.583,\ p < 0.001,\ \eta^2 = 0.04$], METHOD $\times$ SNR [$F(6, 459) = 2.156,\ p < 0.046,\ \eta^2 = 0.008$] were all significant.

It is observed that PESQ scores decrease for SPWD-ASF and they were significantly lower than both MaxDF and SPWD-FSF with mean differences of 0.212 and 0.2, respectively. The differences between the mean PESQ scores of MaxDF and SPWD-FSF were not significant according to the post-hoc comparisons with Tukey correction. The difference between mean PESQ scores for cases with and without post-filtering was -0.09 which was significant at $p < 0.001$ level. As PESQ scores for the 0 dB SNR case were significantly lower than other cases at $p < 0.001$ level similarly to the SIR, there is not a statistical evidence for the comparison of PESQ scores between the other SNR levels.

The PESQ scores for MaxDF were not significantly reduced via post-filtering. While PESQ scores for both SPWD-FSF and SPWD-ASF were negatively affected via employing post-filtering as statistically significant ($p < 0.001$) , the difference is only practically meaningful for SPWD-ASF where the application of post-filtering reduces the mean PESQ score by 0.21.

	SNR (dB)			
	0	10	20	30
RMSE	8.596	6.490	6.191	4.897
Mean	7.131	5.483	5.121	3.740
Stdev Deviation	4.814	3.482	3.489	3.172

Table 4.25: DOA Estimation Errors for Different SNR Levels

As a result, the mean PESQ scores is slightly improved via post-filtering at lower SNR values while it distorts them for higher SNR values. More specifically, while application of post-filtering reduces the mean PESQ score by 0.109 in statistically significant at $p = 0.008$ for 0 dB SNR, the difference is not statistically significant for 10 dB SNR, and it improves the PESQ scores by 0.172 and 0.214 for 20 and 30 dB SNR levels, respectively.

4.4.6 Discussion of results

The evaluation demonstrates a correlation between better source separation performance in terms of the SIR measure and the SPWD-ASF with post-filtering. This happens with the cost of lowering the PESQ scores which are correlated with the perceived quality of the separated sources. The evaluations verified the initial expectations that higher D/R ratio at the source position, higher the associated objective metrics. Also, there is a negative impact between the increasing source count or decreasing SNR over all objective metrics.

SPWD-ASF method with Wiener post-filtering achieves the highest SIR performance in the noise-free case. We have obtained the highest PESQ score with SPWD-FSF without Wiener post-filtering. The Wiener post-filtering addition for the SPWD-ASF in the noise-free case slightly improves the SIR performance. WPF selection can depend on the trade-off between this small improvement and additional computational cost.

The proposed source separation approach (SPWD-ASF with WPF) is robust and pro-

vides best SIR performance to noise where it can provide mean SIR improvements in excess of 15 dB even. The proposed Wiener post-filtering approach also improves the SIR performance according to the MaxDF for all SNR levels. In addition, it is observed that the combination of a high level of reverberation and sensor noise especially at 0 and 10 dB SNR, causes SAR and SDR to be very low. While these conditions represent very unrealistic situations, especially since the nominal SNR for low-cost piezoelectric or MEMS microphones are typically less than 20 dB [120]. In this result, the more realistic scenario among those evaluated is the one that emulates 30 dB SNR.

4.5 Comparison of the methods with the state of the art

4.5.1 DOA estimation error comparison

Three state-of-the-art methods, as well as SRP-based high-resolution DOA estimation, are used for the comparison of the DOA estimation performances. The state-of-the-art methods chosen for comparison are PIV [25], SSPIV [26], and DPD-MUSIC [31]. A general overview of the compared methods, the specific test setup, and the evaluation results are presented in this section. Specific details of the algorithms are beyond what is described in the background section. SRP method based on plane-wave decomposition is chosen as the baseline as full grid search SRP with local maxima detection. The parameters used for PIV, SSPIV and DPD-MUSIC methods (e.g., time and frequency resolutions, amount of overlap in the time-frequency representation, and the employed thresholds) comparison are selected to match those in the original publications.

DOA estimation algorithms' parameters : In the comparisons, the frequency range of PIV was between 0 and 4 kHz, the window size was 8 ms, and the overlap was 4 ms [25]. The estimated local DOAs from the time-frequency ranges are established into a global DOA histogram map to cluster. SSPIV vector is calculated from the first three elements of the signal subspace vector [26]. In the comparisons, the frequency range was between 500 and 3850 Hz, the window size was 8 ms, the

overlap was 6 ms, and the decomposition order was $N = 3$ [121]. The output is a global DOA map used to cluster. In the comparisons, DPD-MUSIC's working frequency range was between 500 and 3875 Hz, the window size was 16 ms, the overlap was 75%, and the decomposition order was $N = 3$ [31]. HEALPix tessellation level is selected as three for the EB-MUSIC spectrum. The maxima values at the pixel centroids are selected as the DOAs, which are estimated for each time-frequency bin.

The selected frequency range for both SRP and HiGRID was between 2608 and 5214 Hz, and the window size was 21.3 ms, the overlap was 2 ms, and the decomposition order was $N = 4$. HEALPix tessellation level is selected as three for the EB-MUSIC spectrum. SRP was calculated at all the pixel centroids. For HiGRID, the maximum depth of the analysis tree was 3. For both methods, the NNL method is used for DOA estimation, and that was applied over the global DOA histograms. The source count parameter was needed for PIV, SSPIV, and DPD-MUSIC because these algorithms are required to get the source count for the clustering. Source count is given as input to these algorithms, and no such assumption was made for SRP and HiGRID.

4.5.1.1 HiGRID Test Scenarios

The created scenario involves four concurrent sources positioned at S1: $(90°, 45°)$, S2: $(90°, 135°)$, S3: $(90°, 225°)$ and, S4: $(90°, 315°)$ at a distance of 1.41 m. The AIRs were emulated using the AIR measurements from METU SPARG AIR dataset. The D/R ratios are selected as similar values for these specific source positions were, 3.04, 3.29, 3.56, and 2.39 dB, respectively. This setup is generated to construct a simple scenario with spatially well-separated sources and a high D/R ratio.

The comparison is executed for nearly coherent and non-coherent types of sources (speech and violin) to evaluate the HiGRID approach. The sampling rate and the duration were $F_s = 48$ kHz and 4 s is selected as in the previous section, respectively. The employed speech signals for sources 1 to 4 are female speech (English), male speech (English), female speech (Danish), and male speech (Danish), respectively.

Source	SRP	HiGRID	PIV	SSPIV	DPD-MUSIC
1	1.54°	1.15°	13.12°	0.87°	0.69°
2	1.87°	1.34°	5.76°	1.21°	2.04°
3	1.39°	0.32°	6.86°	1.56°	1.17°
4	0.80°	0.36°	3.51°	1.28°	0.56°
Average	1.40°	0.79°	7.31°	1.23°	1.12°

Table 4.26: DOA estimation errors for four concurrent speech sources using different methods.©IEEE 2018 [1]

Source	SRP	HiGRID	PIV	SSPIV	DPD-MUSIC
1	1.21°	1.15°	4.89°	6.60°	3.45°
2	1.05°	1.34°	10.07°	4.11°	4.21°
3	1.15°	0.70°	20.41°	2.31°	2.28°
4	1.05°	1.18°	11.92°	7.20°	1.12°
Average	1.12°	1.09°	11.82°	5.06°	2.77°

Table 4.27: DOA estimation errors for four concurrent violin sources using different methods.©IEEE 2018 [1]

Discussion of the results : Tables 4.26 and 4.27 show the DOA estimation results for the two different source types. While calculating the DOA estimation errors, the Kuhn-Munkres algorithm [113] is used for the association between the identified sources and the ground truth.

It is seen that HiGRID outputs a good performance that is very similar to SRP with consistent DOA estimates. It also outputs small estimation errors both for individual sources and on average. This result indicates a distinct advantage over SRP since the computational cost associated with HiGRID is much lower, as explained in Sec. 4.7.2.

Both the algorithms, with the exception of PIV, provide an acceptable level of accuracy with DOA estimation errors in the order of one degree for speech signals. It can be observed seen from Table 4.26. Overall, the worst performance is obtained via the PIV method, while HiGRID provides the best performance. However, the differ-

ences between SRP, HiGRID, SSPIV, and DPD-MUSIC are not significantly high in a practical context.

The results in Table 4.27 show the detrimental effects of near-coherent sources on PIV and, to a lesser extent, on SSPIV and DPD-MUSIC. The average DOA estimation errors except for HiGRID and SRP are much higher than the test using speech signals. HiGRID and SRP provide better DOA estimation performance that is similar to the case with speech sources. However, the performance of PIV and SSPIV are reduced sharply, where the DOA estimation from PIV will be entirely unusable for Source 3. DOA estimates from DPD-MUSIC for individual sources are all higher than those for the first test case. The average error is larger than the first case by a factor of 2.5.

4.5.1.2 RENT Test Scenarios

RENT DOA estimation method is also compared with other state of the art methods: HiGRID [1], SSPIV [26] and DPD-MUSIC [31]. The evaluation involves using measured AIRs for four sources at diagonal positions at a distance of 1.41 m [7]. These AIRs are convolved with the anechoic speech recordings to compose a mixture. The simulated sources were positioned in the horizontal plane at the directions of $(\theta, \phi) = (90°, 45°)$, $(90°, 135°)$, $(90°, 225°)$ and, $(90°, 315°)$.

Source	HiGRID	SSPIV	DPD-MUSIC	RENT
1	$1.15°$	$6.6°$	$3.45°$	$3.38°$
2	$1.34°$	$4.11°$	$4.21°$	$3.38°$
3	$0.7°$	$2.31°$	$2.28°$	$5.62°$
4	$1.18°$	$7.2°$	$1.12°$	$3.38°$
Avg.	$1.09°$	$5.06°$	$2.77°$	$3.94°$

Table 4.28: DOA estimation errors of RENT in comparison with the state of the art DOA estimation methods.©IEEE 2019 [2]

Discussion of the results : The results shown in Table 4.28 indicate that RENT provides worse DOA estimations than HiGRID and DPD-MUSIC, but better DOA

estimations than SSPIV for the tested condition. RENT can be used as a DOA estimation algorithm for the cases where computational cost is a design factor, while it is also robust to noise and can be applied in short time intervals. On the other hand, it has significant outputs that have been used for the source separation algorithms. These are the findings such as identification of frequency regions having single source, noise/diffuse field and multiple sources to calculate noise and active sources' PSD.

4.5.2 Source separation performance

The source separation literature is mostly available with source separation algorithms processed in the cases where the DOA of the sources are initially given as inputs. Two state of the art methods available for RSMAs are selected. Kalkur et al. propose an algorithm related to the sparse recovery of the original recordings via spherical harmonics dictionary using sparsity-based Bregman Iteration [70]. The other work relies on Wiener post-filtering designed to obtain a weighting function for the maximum directivity factor (MaxDF) beamforming [66].

In the first performance evaluation, we have simulated the non-reverberant scenarios to compare the proposed algorithms with the state of the art. Since the method proposed via Fahim et al. does not have a fixed offered technique for DOA estimation, RENT was employed for that purpose. The originally cited DOA estimation technique was the non-convex optimization as Bregman iteration for the article [70] in the non-reverberant case. The simulated scenarios contain AIR calculated for a non-reverberant environment and convolved with the anechoic male and female speech signals selected from EBU SQAM (Sound Quality Assessment Material) Audio CD [114]. Five scenarios each consisting two sources are prepared. The sources are positioned in elevation and azimuth of $\{(90°, 270°), (45°, 0°)\}$, $\{(45°, 90°), (90°, 0°)\}$, $\{(90°, 270°), (45°, 180°)\}$, $\{(90°, 270°), (45°, 90°)\}$, and $\{(45°, 180°), (90°, 270°)\}$. The reference signals are selected as the convolved output of single AIR with the single anechoic recording. First source is selected as male and the second source is female speech. Total duration of the recordings are two seconds. The dilation and spread parameters are selected as $\varrho = 1.0$, $\kappa = 4.0$ and RENT single source and

diffuse field thresholds are selected as 0.5 and 0.1, respectively. The parameters for the state of the art are selected as in the evaluation sections of the original articles.

	Kalkur et al. [70]		Fahim et al. [66]		MaxDF [90]		SPWD-A.+WPF	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Scn. 1	21.77	26.77	26.21	27.84	12.56	13.01	25.14	29.73
Scn. 2	29.84	25.30	27.99	26.88	12.48	12.00	29.26	26.05
Scn. 3	20.12	25.35	24.19	29.26	12.80	18.94	25.31	29.57
Scn. 4	15.74	26.40	24.71	26.37	16.87	13.69	26.03	26.88
Scn. 5	11.90	29.04	26.95	24.24	16.22	12.49	27.04	29.23
Av.±Stdev	$\mathbf{23.22 \pm 5.82}$		$\mathbf{26.46 \pm 1.69}$		$\mathbf{14.11 \pm 2.37}$		$\mathbf{27.42 \pm 1.84}$	

Table 4.29: SIR (dB) Measurements for SPWD-ASF+WPF in different scenarios with the comparison of the state of the art.

	Kalkur et al. [70]		Fahim et al. [66]		MaxDF [90]		SPWD-A.+WPF	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Scn. 1	3.46	7.13	8.30	8.70	7.05	7.37	6.52	8.60
Scn. 2	10.40	3.50	8.65	8.30	7.19	6.85	7.15	5.55
Scn. 3	3.43	8.78	8.30	8.80	7.14	8.53	6.56	8.13
Scn. 4	3.24	8.89	8.34	8.80	8.22	7.63	4.96	7.67
Scn. 5	2.51	9.02	8.41	8.63	8.06	7.19	5.63	7.57
Av.±Stdev	$\mathbf{6.04 \pm 3.07}$		$\mathbf{8.52 \pm 0.21}$		$\mathbf{7.52 \pm 0.56}$		$\mathbf{6.83 \pm 1.16}$	

Table 4.30: SDR (dB) measurements for SPWD-ASF+WPF in different scenarios with the comparison of the state of the art.

Tables 4.29, 4.30, 4.31 and 4.32 show the objective source separation metrics of SIR, SDR, SAR and PESQ. As these scenarios represent the perfect recording with zero reverberation, the time-frequency bin groupings for the signal to be separated can be identified easier than the reverberant case. The noise power spectral density estimation is irrelevant since each of the time-frequency regions would contain the active sources or zero power. Therefore, the proposed Wiener post-filtering (given in Chapter 3.2.3) can not obtain a significant improvement over the state of the art

	Kalkur et al. [70]		Fahim et al. [66]		MaxDF [90]		SPWD-A.+WPF	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Scn. 1	3.56	7.19	8.38	8.76	8.72	8.96	6.59	8.64
Scn. 2	10.45	3.55	8.71	8.37	8.96	8.70	7.18	5.60
Scn. 3	3.53	8.89	8.43	8.84	8.74	9.00	6.63	8.16
Scn. 4	3.61	8.98	8.46	8.89	8.94	9.05	5.00	7.71
Scn. 5	3.31	9.06	8.48	8.77	8.88	8.95	5.67	7.60
Av.±Stdev	$\mathbf{6.21 \pm 2.95}$		$\mathbf{8.61 \pm 0.20}$		$\mathbf{8.89 \pm 0.12}$		$\mathbf{6.88 \pm 1.19}$	

Table 4.31: SAR (dB) measurements for SPWD-ASF+WPF in different scenarios with the comparison of the state of the art.

	Kalkur et al. [70]		Fahim et al. [66]		MaxDF [90]		SPWD-A.+WPF	
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**S1**	**S2**
Scn. 1	2.62	2.09	3.34	3.20	2.49	1.99	3.35	2.83
Scn. 2	2.11	2.56	3.26	3.51	1.98	2.46	2.81	3.16
Scn. 3	2.63	2.34	3.20	3.18	2.51	2.25	3.39	2.70
Scn. 4	2.91	2.26	3.48	2.78	2.80	2.06	3.21	2.67
Scn. 5	2.74	2.16	3.45	2.80	2.74	1.98	3.30	2.33
Av.±Stdev	$\mathbf{2.44 \pm 0.28}$		$\mathbf{3.22 \pm 0.25}$		$\mathbf{2.33 \pm 0.31}$		$\mathbf{2.98 \pm 0.35}$	

Table 4.32: PESQ measurements for SPWD-ASF+WPF in different scenarios with the comparison of the state of the art.

algorithms [70] [66] in SIR. The improvement (14-15 dB) can be revealed over the MaxDF with the post-filtering. RENT results in average DOA estimation errors of $\{8.92°\}$, $\{8.90°\}$, $\{8.51°\}$, $\{8.71°\}$, and $\{8.98°\}$ for the given scenarios respectively. These errors occur from the pixelization differences between the actual source directions and selected HEALPix pixel centers in the tessellation level of $K_{max} = 2$. From the SDR, SAR and PESQ comparison, there is not a major change for the proposed algorithm but the work [66] can provide a 1-2 dB SDR and SAR improvements according to our proposed method.

	Scenario 1		Scenario 2		Scenario 3		
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**Avg± Stdev**
MaxDF [90]	11.21	11.15	5.83	7.28	15.56	12.32	10.56±3.52
Kalkur et al. [70]	18.79	7.50	13.48	4.85	14.05	11.81	11.75±4.97
Fahim et al. [66]	13.56	12.17	11.53	11.05	16.71	15.11	13.36±2.21
SPWD-FSF$_{\kappa=4}$	22.85	20.19	15.62	16.85	18.65	19.64	18.96±2.56
SPWD-ASF$_{\varrho=0.5}$	24.53	22.92	16.22	17.10	18.91	21.25	20.16±3.28
SPWD-ASF$_{\varrho=1.0}$+WPF	22.69	22.63	21.39	23.43	23.82	22.68	22.77±0.83

Table 4.33: SIR (dB) Measurements for SPWD-ASF + WPF in different scenarios with the comparison of the state of the art.

Object-based audio source quality is also validated by changing the gain (A) parameter at the renderer. The normalized separated sources' gains are set to A and $1 - A$ and then a rendered scenario is created. The effect of the gain parameter over PESQ and SDR can be observed in Fig. 4.32. It is seen that rendering the separated sources with gain parameters would improve both sources' quality metrics defined via PESQ.

The second performance evaluation setup consists of multi-channel acoustic impulse response (AIR) recordings from the METU SPARG AIR dataset [7]. RSMA (Eigenmike em32) is positioned at the center of this classroom ($T_{60} \approx 1.12$s) [7]. The same anechoic signals in the previous evaluation are also used in this evaluation. They are convolved with the randomly selected AIR recordings and mixed to generate the audio scene. White Gaussian sensor noise is also inserted to each microphone signal at 30 dB SNR [108]. Selected locations in the defined scenarios for the acoustic sources are marked as red circles in Fig. 4.33 and the sources are stationary. While three scenarios are generated in a reverberant environment having a strong reverberation, the DOA estimation method of Kalkur et. al [70] fails to localize the sources accurately and therefore, RENT is used as common DOA estimation technique to make a fair comparison between the source separation techniques.

Tables 4.33, 4.34, 4.35 and 4.36 show the objective source separation metrics, SIR, SDR, SAR and PESQ. The compared algorithms degrade in the performance of SIR for the strong reverberation cases. The improvements in SDR and SAR resemble

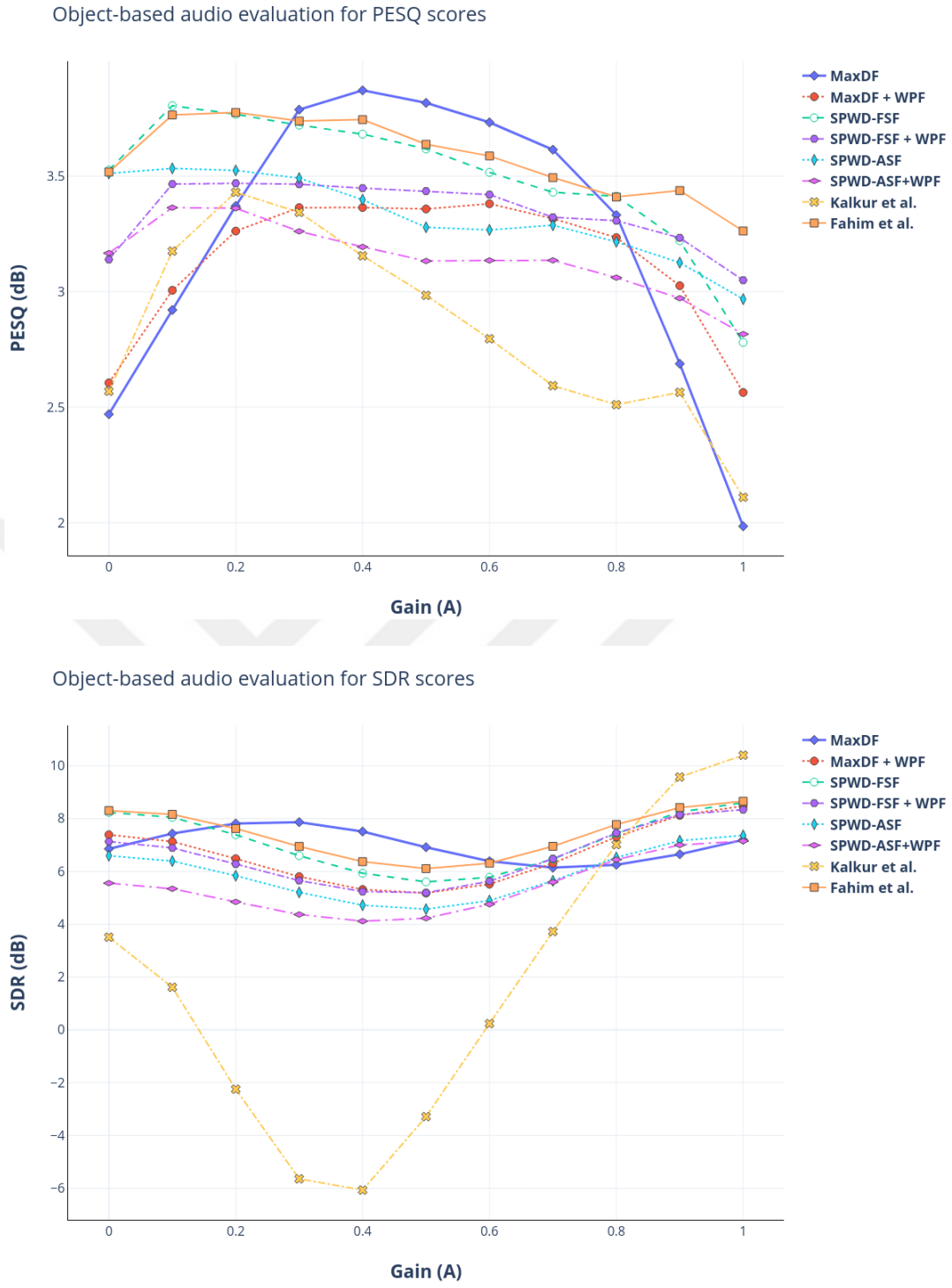


Figure 4.32: Object-based audio source quality (PESQ, SDR) measure at the rendering output by changing the gain in Scenario 2. Before rendering, the first source's gain is set to A and the second source's gain is set to $1 - A$.

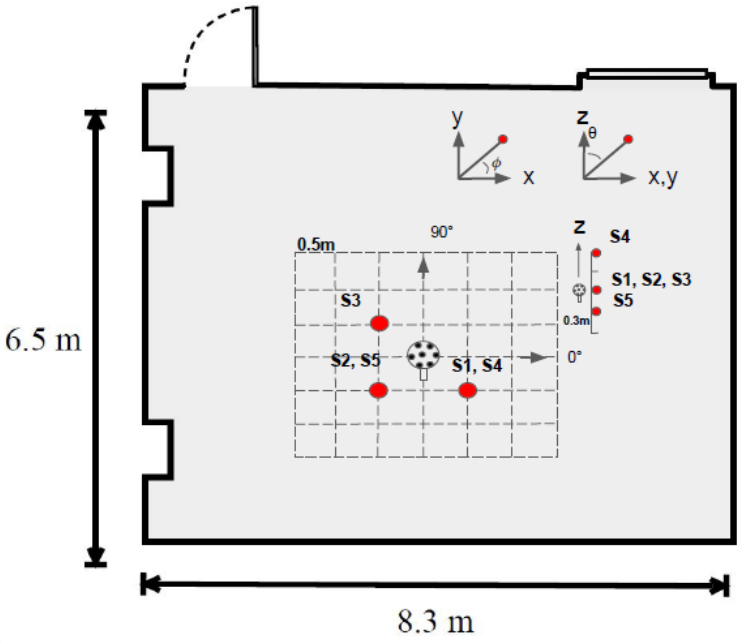


Figure 4.33: Setup generated for comparison between the proposed algorithm and the state of the art. Scenario 1 (S1,S2), Scenario 2 (S1,S3), Scenario 3 (S4,S5)

to each other except for Kalkur et. al. with the worst performance for the source separation in strong reverberant environments. As a conclusion for non-reverberant and reverberant cases, the proposed approach satisfies a better separation performance in the interference suppression. It also provides an acceptable level for an average PESQ score of 1.87 for the reverberant and 2.98 for the non-reverberant case. SPWD-FSF outputs the best source quality via evaluating SDR, SAR and PESQ.

Object-based audio source quality is also validated by changing the gain (A) parameter at the renderer. The normalized separated sources' gains are set to A and $1 - A$ and then a rendered scenario is created. The effect of the gain parameter over PESQ and SDR can be observed in Fig. 4.34. It is seen that rendering the separated sources with gain parameters improve the source quality metric of the source having higher distortion and artefacts.

	Scenario 1		Scenario 2		Scenario 3		
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**Avg± Stdev**
MaxDF [90]	2.45	1.98	0.46	0.97	3.74	3.44	2.17±1.30
Kalkur et al. [70]	-1.00	-0.17	-0.97	-1.16	-0.72	-4.37	-1.40±1.49
Fahim et al. [66]	3.03	2.41	2.35	2.24	4.09	4.14	3.04±0.87
SPWD-FSF$_{\kappa=4}$	3.63	3.73	3.33	3.76	3.67	3.19	3.55±0.23
SPWD-ASF$_{\varrho=0.5}$	3.45	3.35	3.23	3.35	2.80	2.65	3.13±0.33
SPWD-ASF$_{\varrho=1.0}$+WPF	1.19	2.39	1.41	3.08	2.89	3.89	2.48±1.03

Table 4.34: SDR (dB) Measurements in different scenarios with the comparison of the state of the art.

	Scenario 1		Scenario 2		Scenario 3		
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**Avg± Stdev**
MaxDF [90]	3.39	2.86	2.95	2.87	4.16	4.29	3.42±0.65
Kalkur et al. [70]	-0.90	1.34	-0.62	1.30	-0.40	-3.98	-0.54±1.94
Fahim et al. [66]	3.62	3.15	3.21	3.18	4.43	4.64	3.71±0.66
SPWD-FSF$_{\kappa=4}$	3.71	3.87	3.71	4.06	3.87	3.34	3.76±0.24
SPWD-ASF$_{\varrho=0.5}$	3.50	3.42	3.55	3.62	2.96	2.74	3.29±0.36
SPWD-ASF$_{\varrho=1.0}$+WPF	1.25	2.45	1.48	3.14	2.94	3.97	2.54±1.03

Table 4.35: SAR (dB) Measurements in different scenarios with the comparison of the state of the art.

	Scenario 1		Scenario 2		Scenario 3		
	S1	**S2**	**S1**	**S2**	**S1**	**S2**	**Avg± Stdev**
MaxDF [90]	1.87	2.63	1.71	2.46	2.62	1.91	2.20±0.41
Kalkur et al. [70]	1.78	2.62	1.83	2.66	2.66	1.75	2.22±0.47
Fahim et al. [66]	1.96	2.77	1.81	2.60	2.67	1.92	2.29±0.43
SPWD-FSF$_{\kappa=4}$	2.15	2.73	2.00	2.67	2.60	1.84	2.33±0.38
SPWD-ASF$_{\varrho=0.5}$	2.02	2.50	1.82	2.59	2.51	1.87	2.21±0.35
SPWD-ASF$_{\varrho=1.0}$+WPF	1.45	2.07	1.41	2.41	2.21	1.68	1.87±0.41

Table 4.36: PESQ Measurements in different scenarios with the comparison of the state of the art.

4.6 Subjective Evaluation with MUSHRA

MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) is a methodology for obtaining a codec listening test to evaluate lossy audio compression algorithms in the output's perceived quality. In this section, MUSHRA is used to check the algorithms' separation quality to sort them perceptionally from high to low in descending order. The participants are selected from students who are members of the Spatial Audio Research Group (SPARG). Ten participants (9 male-1 female; 23-40 years old) with normal hearing are selected to evaluate the separation blocks' *source separation quality*, and these evaluators were asked to rate each algorithm on a scale from 0 (Bad) to 100 (Excellent).

According to the MUSHRA setup rules, we have expected from the participants to identify the reference and anchor from the randomly given voting list and rate 100 for the reference and 0 for the anchor. The anchor is the acoustic signal with a 10 dB distortion of the omnidirectional component of the scene recording, and the reference in each case was obtained by convolving the direct path component of the respective AIR with the anechoic signal. Four scenarios were evaluated in the form of female/-male, violin/violin, male/male, and female/male. Each evaluation consists of 8 audio files (Reference, Anchor, MaxDF, MaxDF+WPF, SPWD-FSF, SPWD-FSF+WPF, SPWD-ASF, SPWD-ASF+WPF) for each signal. The results were collected from the

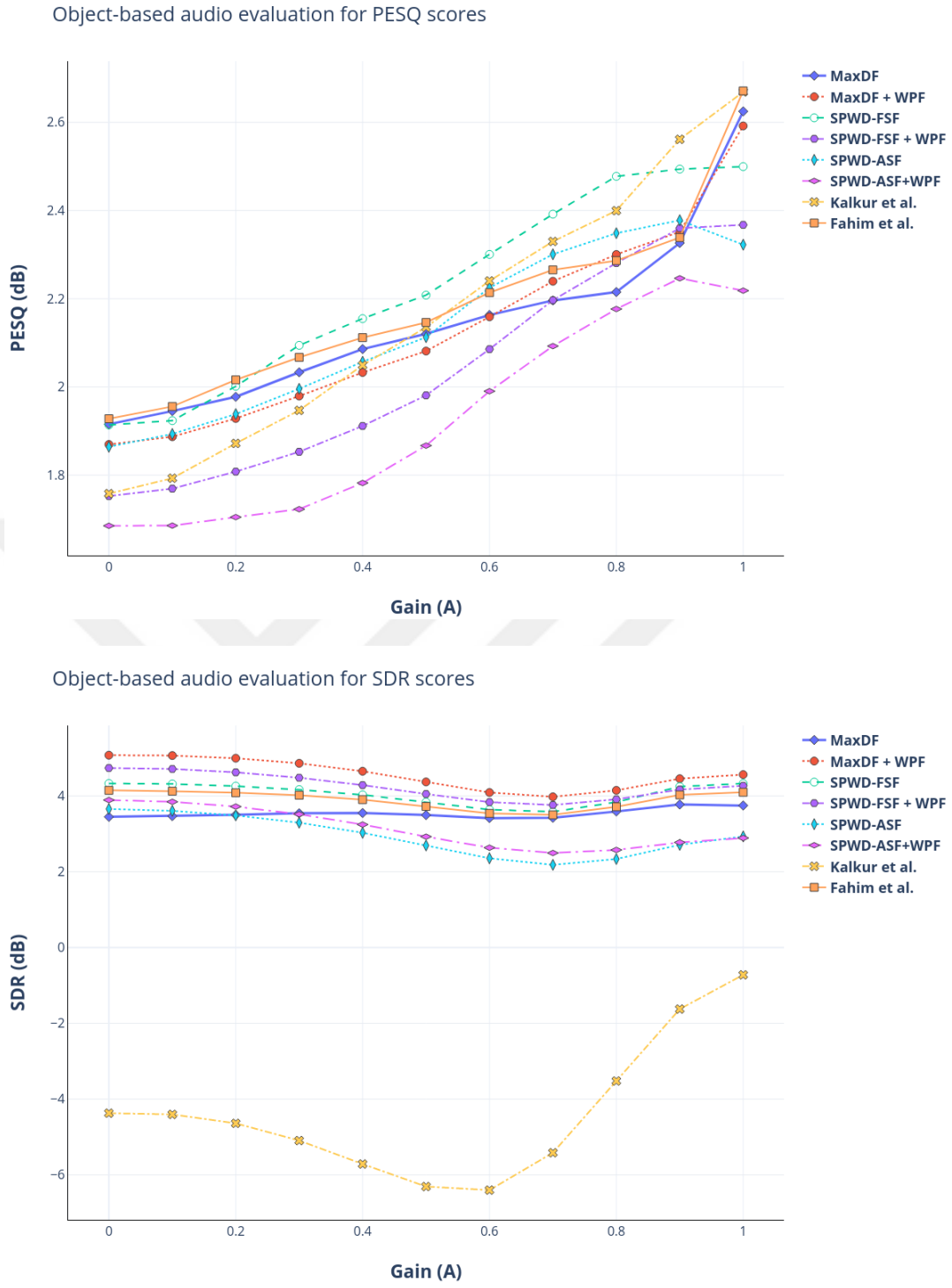


Figure 4.34: Object-based audio source quality (PESQ, SDR) measure at the rendering output by changing the gain in Scenario 3. Before rendering, the first source's gain is set to A and the second source's gain is set to $1 - A$.

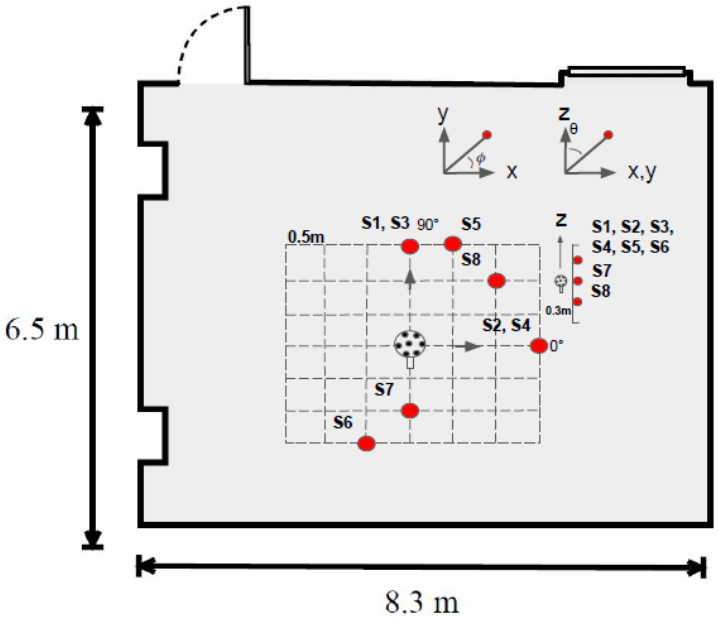


Figure 4.35: Setup generated for webMUSHRA. Scenario 1 (S1,S2), Scenario 2 (S3,S4), Scenario 3 (S5,S6), Scenario 4 (S7,S8) are given and the anechoic signals are female/male, violin/violin, male/male, and female/male respectively.

modified webMUSHRA tool, which is compliant with ITU-R BS.1534 [6]. The separation algorithms' parameters are selected as $\kappa = 4$ and $\varrho = 1$.

The average and standard deviation results obtained from the WebMUSHRA are

	Scn.1(S1,S2)	**Scn.2(S3,S4)**	**Scn.3(S5,S6)**	**Scn.4(S7,S8)**
MaxDF	41.90 ± 27.39	40.90 ± 19.79	43.15 ± 25.39	38.15 ± 23.95
MaxDF + WPF	56.45 ± 20.97	57.6 ± 19.89	60.35 ± 19.09	67.90 ± 19.84
SPWD-FSF	61.35 ± 21.02	48.10 ± 19.47	55.2 ± 18.98	55.30 ± 23.50
SPWD-FSF + WPF	67.90 ± 14.96	62.50 ± 19.27	66.70 ± 16.07	67.40 ± 18.40
SPWD-ASF	69.75 ± 25.75	54.65 ± 22.94	69.25 ± 20.35	68.85 ± 15.75
SPWD-ASF + WPF	68.55 ± 21.81	66.00 ± 26.49	72.75 ± 21.67	78.30 ± 18.49

Table 4.37: Evaluation of the separation methods using webMUSHRA with the help of eleven individual listeners.

given in the Table 4.37. By averaging the rates obtained with the full scenarios given for the algorithms are 41.11 ± 23.89 (MaxDF), 60.58 ± 20.09 (MaxDF + WPF), 54.49 ± 20.97 (SPWD-FSF), 66.13 ± 17.07 (SPWD-FSF+WPF), 65.63 ± 22.05 (SPWD-ASF), 71.40 ± 22.37 (SPWD-ASF+WPF) respectively. The mean and standard errors for the anchor and the reference were 11.96 ± 14.99 and 99.87 ± 1.12, respectively. It is observed that SPWD-ASF with Wiener post-filtering provides better source separation results when compared with the other combinations. While scenario 1 and scenario 2 contain the same AIRs, they contain different anechoic source types (speech and instrument) as near-incoherent and near-coherent sources. According to the rates given for the near-coherent scenario seen from the scores, applying Wiener post-filtering is more robust than without post-filtering for the separation.

One-way ANOVA is also used to analyze the results. *METHOD* was the independent variable which was found to have a significant effect $[F(7, 632) = 140.16,\ p < 0.001,\ \eta^2 = 0.608]$ over the results.

Post-hoc comparisons with Tukey correction indicate that the obtained differences between mean separation quality scores given to tested methods were higher than the baseline MaxDF ($p < 0.001$). MaxDF and SPWD-ASF with Wiener post-filtering improve the scores by at least 13.56 and 30.26 points, respectively. Mean differences between SPWD-FSF and i) SPWD-FSF with Wiener post-filtering $[MD = -11.58, SE = 3.01]$, ii) SPWD-ASF $[MD = -10.97, SE = 3.41]$, and iii) SPWD-ASF with Wiener post-filtering $[MD = -17.01, SE = 3.41]$ were statistically significant at $p = 0.003$, $p = 0.006$ and $p < 0.001$ levels, respectively. No other comparisons among the results revealed statistically significance. In summary, this MUSHRA test analysis supports the proposed modular approach (SPWD-ASF with WPF) in any of its combinations can provide a good level of subjective separation quality.

4.7 Computation cost comparison

4.7.1 Computational cost of HiGRID

$C(n)$ is defined as the cost of a real multiplication with n bits precision. A complex multiplication can be obtained using $3C(n)$ real multiplications. Similarly, a real division can also be obtained using $3C(n)$ multiplications. It is also known that natural logarithm's cost can be investigated as in time $\sim nC(n)$ using the power series method [122]. The cost using eigendecomposition assumption for cross-spatial density matrices can be neglected because $\mathbf{Q}_i$ in (3.7) have been calculated offline and a reduced set of eigenvalues and eigenvectors have been stored in an appropriate dictionary as described in (3.14). In the following analysis, N is the maximum SHD order, and $l = K_{\max}$ is the maximum decomposition level.

HiGRID contains two stages as calculation of: 1) SRPD functional and 2) spatial entropy. SRPD can be calculated in time $\sim 6(N+1)^2C(n)$ at the lowest resolution level (i.e., $l = 0$) and $\sim 3(N+1)^2C(n)$ when $l \geq 1$. The total spatial entropy for the four higher resolution nodes can be calculated in time $\sim 4(n+4)C(n)$.

By evaluating a single given time-frequency bin, HiGRID starts by calculating 12 SRPD values at all the lowest resolution nodes at level, $l = 0$. At the first iteration, SRPD is calculated at 48 higher resolution nodes at level $l = 1$ and spatial entropy is calculated 12 times. Depending on the underlying SRPD functional, $\mathcal{K}_0$ out of 12 nodes are selected for processing at the next iteration.[1] At a given level $1 \leq l < K_{\max}$, SRPD is calculated $4\mathcal{K}_{l-1}$ times and spatial entropy is calculated $\mathcal{K}_{l-1}$ times. Finally, at the highest resolution level (i.e., $l = K_{\max}$) only the SRPD functional is calculated at $\mathcal{K}_{K_{\max}}$ nodes. In summary, HiGRID for a single time-frequency bin can be calculated in time:

$$\text{HiGRID}(n) \sim \left[3\left(72 + \mathcal{K}_{K_{\max}} + 4\sum_{l=1}^{K_{\max}-1}\mathcal{K}_l\right)(N+1)^2 + 4(n+4)\left(12 + \sum_{l=1}^{K_{\max}-1}\mathcal{K}_l\right)\right] M(n)$$

with n digits precision.

[1] Note that at the lowest resolution, the number of multiplications for calculating SRPD is $2(N+1)^2$.

Spatial entropy calculated via SRPD functional will determine the exact number of multiplications. However, if there exists an α such that $0 \leq \alpha \leq 1$ and $\mathcal{K}_l \leq 4\alpha\mathcal{K}_{l-1} \leq (4\alpha)^l\mathcal{K}_0$, an upper bound for the number of multiplications can be obtained. If we assume that IEEE 754 single-precision floating point format with $n = 23$ fraction bits is used in the calculations, the upper bound for the total number of real multiplications can be calculated to be in the order of:

$$\left(216 + 3\mathcal{K}_0\left[16\alpha\frac{1-(4\alpha)^{K_{\max}-1}}{1-4\alpha} + (4\alpha)^{K_{\max}}\right]\right)(N+1)^2 + 108\left(12 + 4\alpha\mathcal{K}_0\frac{1-(4\alpha)^{K_{\max}-1}}{1-4\alpha}\right).$$

A direct comparison is possible with SRP calculated at the same resolution. The number of real multiplications needed for SRP is:

$$36 \cdot 4^{K_{\max}} \cdot (N+1)^2$$

Table 4.38: Upper bounds for the ratio of the number of real multiplications required for HiGRID and SRP

	$K_{\max} = 3$				$K_{\max} = 4$			
α	**0.25**	**0.5**	**0.75**	**1.0**	**0.25**	**0.5**	**0.75**	**1.0**
$\mathcal{K}_0 = 1$				0.29				0.20
$\mathcal{K}_0 = 2$		0.21		0.43		0.08		0.35
$\mathcal{K}_0 = 4$	0.18	0.30	0.47	0.72	0.05	0.13	0.32	0.65
$\mathcal{K}_0 = 8$	0.24	0.47	0.82	1.29	0.07	0.23	0.59	1.25
$\mathcal{K}_0 = 12$	0.30	0.65	1.17	1.86	0.10	0.33	0.87	1.84

Table 4.38 shows the upper bounds of the ratio of the number of multiplications for HiGRID and SRP for different resolution levels $K_{\max} = 3$ and $K_{\max} = 4$ as well as different values of $\mathcal{K}_0$ and α. Cases which are not feasible (e.g. $\mathcal{K}_0 = 1$ and $\alpha = 0.5$ cannot be observed due to the expansion of a node resulting in 4 new nodes at the higher resolution) are not shown. It may be observed that in most cases, HiGRID requires fewer multiplications than SRP. A special case of interest occurs when the DOA of a single prominent source is exactly at the centroid of four nodes. For that case $\mathcal{K}_0 = 4$ and even if entropy cannot eliminate any steering directions apart from

at the lowest resolution (i.e., $\alpha = 1$), HiGRID outperforms SRP in terms of computational cost.

4.7.2 Computational cost of SPWD Algorithms

$C(n)$ is defined as the cost of a real multiplication with n bits precision. A complex multiplication can be obtained using $3C(n)$ real multiplications. Similarly, a real division can be obtained using $3C(n)$ multiplications. We also note that the natural logarithm can be obtained in time $\sim nC(n)$ using the power series method [122]. In the following analysis, N is the maximum SHD order, Q is the number of microphones, and $K_{\max}$ is the maximum decomposition level steering in $M = 12 \times 2^{2K_{\max}}$ pixel locations, and $D = 12 \times 2^{2K_{\max}}$ dictionary items. For a complex multiplication of matrices in the dimensions of $A \times B$ and $B \times D$ matrix, the computational cost can be calculated as $3 \times A \times B \times D \times C(n)$ real multiplications.

RENT consists of two distinct stages for each time-frequency bin: 1) calculation of the SRF and 2) finding the best fitting dictionary atom and the residual. It contains a global histogram clustering technique (sorting the array with quick sort for an interval and check the neighbouring relations) finally. Using real valued Legendre kernels defined for D-DOAs with M steering directions, kernel dictionary is an $M \times D$ matrix which is calculated and stored offline for each spherical harmonics order N. M and D are assumed as equal for the computations. The matrix formulation for the SRF from the microphone recordings is given as:

$$\mathbf{y_N} = \widetilde{\mathbf{Y}}\mathbf{B^{-1}}\mathbf{Y^H}\mathbf{W}\mathbf{p}, \tag{4.16}$$

where $\mathbf{p}$ is a real valued microphone signals in $Q \times 1$, W is real valued quadrature weights in $Q \times Q$, $\mathbf{Y^H}$ is complex valued matrix in $(N+1)^2 \times Q$, $\mathbf{B}$ is the frequency decoupling complex valued component in $(N+1)^2 \times (N+1)^2$, and $\widetilde{\mathbf{Y}}$ is complex valued component in $M \times (N+1)^2$. The computational cost for the SRF from left to right matrix multiplication is $\{[(N+1)^4 + (N+1)^2 \times Q + Q^2 + Q] \times M \times 3C(n)\}$. Note that SRF is a complex matrix in the form of $M \times 1$, where $M = 12 \times 2^{2K_{\max}}$. In order to find the best fitting atom, the dictionary is multiplied with the SRF in the computational cost of $\{D \times M \times 3C(n)\}$ and the residual error is estimated via

$\{M \times 3C(n)\}$ with the division of SRF with the residual is the ratio $\{3C(n)\}$. As a result, computational cost for SRF in a single time-frequency bin is given as :

$$\textit{SRF}_{CC} \approx [[(N+1)^4 + (N+1)^2 \times Q + Q^2 + Q] \times 12 \times 2^{2K_{\max}}] \times 3C(n)$$

Computational cost for RENT in a single time-frequency bin is given as :

$$\textit{RENT}_{CC} \approx \textit{SRF}_{CC} + [(12 \times 2^{2K_{\max}})^2 + (12 \times 2^{2K_{\max}}) + 1] \times 3C(n)$$

SPWD-FSF consists of OMP with L iterations to be executed over the SRF except for threshold ratio computation. Therefore, the computational cost for a single time-frequency bin is given as:

$$\textit{SPWD-FSF}_{CC} \approx \textit{SRF}_{CC} + [L \times [(12 \times 2^{2K_{\max}})^2 + (12 \times 2^{2K_{\max}})]] \times 3C(n)$$

SPWD-ASF consists of OMP with 1 iteration for RENT and $L-1$ iteration to be executed for source separation over the SRF. Therefore, the computational cost for a single time-frequency bin is given as $\textit{SPWD-ASF}_{CC} \approx \textit{SPWD-FSF}_{CC}$:

$$\begin{aligned}\textit{SPWD-ASF}_{CC} \approx \textit{SRF}_{CC} + [((12 \times 2^{2K_{\max}})^2 + (12 \times 2^{2K_{\max}}) + 1) + (L-1) \\ \times [(12 \times 2^{2K_{\max}})^2 + (12 \times 2^{2K_{\max}})]] \times 3C(n)\end{aligned}$$

The computational cost given for the pseudoinverse operation of an $S \times M$ matrix is $M^S \times 3C(n)$, where $M = 12 \times 2^{2K_{\max}}$ is the number of steering directions and S is the total source count. It is calculated for each time-frequency bin.

$$\textit{WPF}\{\textit{pseudoinverse}\}_{CC} \approx [(12 \times 2^{2K_{\max}})^S] \times 3C(n)$$

WPF equation is given in the formulation of $\mathbf{\Phi_a} \approx \mathbf{L}^{\dagger}(\mathbf{\Phi_y} - \mathbf{\Phi}_{\eta})(\mathbf{L}^{\dagger})^H$. The computational cost carried out for the covariance matrix and matrix multiplications of SPWD-ASF + WPF for each time-frequency bins are given as:

$$\begin{aligned}\textit{WPF}_{CC} \approx \textit{WPF}\{\textit{pseudoinverse}\}_{CC} + [(12 \times 2^{2K_{\max}})^2 + (12 \times 2^{2K_{\max}}) \\ \times S^2 + S \times (12 \times 2^{2K_{\max}})^2] \times 3C(n) + \textit{SPWD-ASF}_{CC}\end{aligned}$$

The global DOA histogram sorting (quick sort) cost via RENT is given in the following equation. This operation is executed one time only for each interval, and it can be neglected in the computations because of resulting smaller values.

$$\textit{RENT}\{\textit{sorting}\}_{CC} \approx M\log(M) \times C(n)$$

	$K_{\max}=2$			$K_{\max}=3$			$K_{\max}=4$		
N	2	3	4	2	3	4	2	3	4
SRF_{CC}	0.82	1.05	1.42	3.28	4.2	5.71	13.13	16.80	22.86
$RENT_{CC}$	0.93	1.16	1.54	5.05	5.97	7.48	41.45	45.13	51.18
$SPWD\text{-}FSF_{CC}$	1.93	2.16	2.54	21.00	21.92	23.43	296.34	300.01	306.07
WPF_{CC}	2.38	2.60	2.98	28.08	29.00	30.52	409.62	413.30	419.35

Table 4.39: Computational cost calculated for OMP-based algorithms for 1000000 real multiplication $C(n)$

Table 4.39 represents the computational cost of each algorithm in terms of 1000000 real multiplication $C(n)$. Q is fixed number of microphones as 32, and M is the total source count assumed as 2. While the new generation Intel CPUs have capability of 20 billion cycles and 1 real valued multiplication per 3 cycle in a second [123], $K_{\max}=2$ and $N=4$ can be selectable for both algorithms in real time aiming an approximate number of 3000 time-frequency bin processing for each second.

	$K_{\max}=2$			$K_{\max}=3$			$K_{\max}=4$		
N	2	3	4	2	3	4	2	3	4
SRF_{CC}	0.02	0.02	0.02	0.07	0.07	0.07	0.29	0.29	0.29
$RENT_{CC}$	0.12	0.12	0.12	1.84	1.84	1.84	28.61	28.61	28.61
$SPWD\text{-}FSF_{CC}$	1.13	1.13	1.13	17.79	17.79	17.79	283.50	283.50	283.50
WPF_{CC}	1.57	1.57	1.57	24.87	24.87	24.87	396.78	396.78	396.78

Table 4.40: Computational cost calculated for pursuit based algorithms using hastables for 1000000 real multiplication $C(n)$

It is observed that the cost directly depends over the SRF cost for each time-frequency bin. As an alternative, SRF formula can be split into two parts as offline and online matrices given in the equation (4.17).

$$\mathbf{y_N} = \mathbf{Fp}, \tag{4.17}$$

where $\mathbf{F} = \widetilde{\mathbf{Y}}\mathbf{B^{-1}}\mathbf{Y^H}\mathbf{W}$ is a matrix in $M \times Q$ dimensions and it can be calculated before the calculations and stored in a hashtable to decrease the total number of multiplications. According to this modification, cost of SRF is given as:

$$\textit{SRF}_{CC} \approx [Q \times 12 \times 2^{2K_{\max}}] \times 3C(n).$$

By using the offline SRF multiplicator matrix $\mathbf{F}$, the updated computational costs are displayed in Table 4.40. This cost reduction as removing the cost dependence over N would make the $\textit{WPF}_{CC}$ algorithm to run two times faster for the $K_{max} = 2$ operations. As a conclusion, it should also be noted that hashtable usage can significantly lower the computational cost, if read/write operation speeds from the memory for the stored matrices are fast.

CHAPTER 5

DISCUSSION

Spatial audio has tremendous popularity among electronic device manufacturers, which target headsets and loudspeakers. The main objective is to impress the end-users in 3D immersive audio sensations. The synthesis methods to deliver the recordings in 3D audio format are also available in the market. These methods automatically adapt the audio format based on the number of loudspeakers, their orientations, and positions [10]. Besides, the design of analysis methods to extract the source locations with the purpose of separating them from the real recordings such as theatre performances or concerts is an ongoing research topic [124] [66]. This thesis's main motivation was to develop an acoustic scene analyzer, and the presented results show that a scene encoded in HOA can be analyzed and transcoded into the object-based format. That would enable the end-user devices to render the separated objects based on the loudspeaker orientations and provide them with extended flexibility.

This thesis presented two DOA estimation and three source separation techniques to develop a 3D scene analyzing tool. While these methods are carried out for the rigid spherical microphone array (RSMA) recordings, any HOA signal, real or synthetic, can also be used directly for the analysis. The performance of the proposed methods has been evaluated in Chapter 4 using emulated scenarios with the AIRs from the METU SPARG dataset [7]. As we have observed good performances in the emulations, they also result in significant improvements compared with state of the art. Beyond the emulations, the algorithms provide excellent separation results (given in Fig. 4.26) from the original recordings via Eigenmike em32. In this part, the proposed methods' key findings and limitations for different conditions are discussed to get the optimum solution in the DOA estimation accuracy and separation performance.

There are a variety of methods available for DOA estimation using microphone arrays. They work either in Cartesian or spherical coordinate systems. The problem definition in full elevation and azimuth coverage makes the spherical coordinate system and SHD more convenient to work with. Most of the DOA estimation techniques are useful for signals with non-overlapping short-time spectra. These techniques can not deliver DOA estimations accurately for coherent signals that have overlapping short-time spectra. The source count parameter is required to specify the source directions for the case of coherence occurs. The proposed methods' motivation is to estimate the DOAs and source counts accurately in non-reverberant and reverberant environments with and without sensor noise for both incoherent and near-coherent signals. These methods also have to identify coherent or non-coherent signals by lowering the computational cost.

HiGRID and RENT are the proposed solutions for the DOA estimation from the scene-based recordings. While both of these algorithms provide accurate DOA estimation, HiGRID is the preferred technique when accuracy is concerned. It provides a novel hierarchical grid refinement approach to decrease the computational cost by viewing the sphere in different tessellation levels instead of steering in the full directions. Additionally, RENT is the other solution to satisfy $4^\circ - 6^\circ$ DOA estimation errors, group the time-frequency bins, and estimate the noise PSD. The results also indicate that source count does not significantly affect the DOA estimation error for RENT. The reasons to make a choice among the techniques for a DOA estimation problem differ with the desired achievements. These are the accuracy for the DOA estimation, computational cost, and grouping the time-frequency bins as a single source, valid source, or noise/diffuse field. For example, RENT can be used as an alternative to other state of the art methods as DPD and SSZ, while it is used as a preprocessing stage. PSD or source distribution estimation of the active sources around the array can also be used in generating adaptive filters. RENT is needed for cases where these types of preliminary information are required.

Mills had observed the minimum audible angle to localize the sources having the same frequency tones in an experiment [115]. This angle was examined with a number of participants in an anechoic chamber. According to the low-frequency region comprising voice spectra, the results indicate that the minimum audible angle is be-

tween $4-5$ degrees. This information gives insights into the minimum DOA estimation angle requirement for the source separation in the OBA context. Grouping the sources in more accurate angles does not significantly affect the source separation performance if it is not larger than $4-5$ degrees. Therefore, RENT usage instead of HiGRID does not provide a significant degradation. While using the fixed dictionary at the initialization for RENT increases the computational cost, the atoms in the dictionary would be adaptive to the HiGRID DOA findings in short time intervals. An accurate DOA estimation can be used to create a dynamic dictionary to label the single source TF bins in each time interval. ANCOVA analysis of the randomly generated setups indicates that DOA estimation error does not statistically significantly affect the source separation performance.

The automatic transcoding process performance from the scene-based to object-based audio formats can be affected by the change in reverberation time and sensor noise level. RENT thresholds are selected experimentally by evaluating the source separation performances in the strong reverberant cases. While the algorithm parameters used in RENT are set constant as ($THR_{ss} = 0.5, THR_{df} = 0.1$) for the heavily reverberant case [7] which can be followed as the worst, these parameters can also be tuned for the environments having different reverberation time. However, it is beyond the scope of this thesis. The limitation on the HEALPix tessellation level K_{max} is the algorithm parameter that would directly affect the DOA estimation accuracy and computational cost. The results obtained by using different tessellation levels indicate that the average DOA estimation error does not significantly change via decreasing the level. However, the standard deviation is negatively affected. Whether it is a speech or musical instrument, the source type does not affect the DOA estimation performance. Also, computational cost can be lowered by the SRPD formulation and initial computation of the Legendre kernels via storing them in the memory. The computations presented in Sec. 4.7.1, 4.7.2 indicate that the cost is constant when the stored hashtables are used, while the spherical harmonics maximum order is incremented. However, the maximum tessellation level is a significant parameter that negatively affects the cost.

Rayleigh condition is another limitation to cluster the sources because both HiGRID and RENT use a global histogram clustering for source count and DOA estimation.

If the same source's multipaths from reverberation seen in the corresponding active source angle direction, it can not be separated due to pixel neighboring relation checks. Therefore, the primary assumption for the DOA estimation techniques is the minimum angle difference among the sources. That angle can not be less than π/N, where N is the maximum spherical harmonics order. The maximum order is also limited according to the number of microphones $Q \geq (N+1)^2$ for scene representation.

Acoustic scene decomposition was the other major research topic in this thesis. Separating the sources only from their spectrograms instead of their source directions requires initial source features. The identified features can be the ground truth spectrograms or working frequency regions related to the source type. However, this controlled scenario can not be established for the randomly selected recordings. The spatial data is the critical point for the algorithms to separate the acoustic sources in the conditions where the coherence occurs. Most of the algorithms require initial inputs as the source count and DOAs. An automatic transcoding process comprises the extraction of DOAs first and then uses it to separate the sources. Joint source DOA estimation and separation with a number of source separation algorithm combinations is proposed with this study, and they are evaluated via subjective and objective metrics.

By assuming that the sound emitting objects in the scene are stationary, and the DOA of the sources are fixed, an OMP-based sparse plane wave decomposition fixed spatial filtering (SPWD-FSF) was proposed to obtain perceptually good separation results. While this method gets the spatial information (DOAs of the sources and source count) at the initialization, it would also carry out the separation results in small time intervals via RENT. This would provide continuity while the musicians and instruments make small displacements. The fixed spatial filter's directional center is adapted according to the DOA estimations in short time intervals. The degradation of SPWD-FSF in separation appears for sources positioned in narrow-angle differences. This situation would happen using a smaller fixed filter's symmetric spread parameter κ selection. If the other sources' interference is present in the main lobe of the employed virtual beam (i.e., the von Mises function), worse SIR results would be obtained by using smaller κ. While increasing the κ results in better source separation results, distortion and undesirable artifacts would increase due to the generation

of smaller spatial regions. This problem was addressed by making the virtual beamforming adaptive.

The proposed solution adapts a non-symmetric Kent distribution filter according to the clustered source distributions using MLE. This filter also has a dilation parameter ϱ that affects the spread to prevent the distortion and artifacts happening from the narrower spatial region outputs (given as in Figure A.4). RENT identifies and clusters the source distributions to fit into customized filters. Thus, by decreasing the dilation parameter ϱ, the spatial filter becomes wider in the spatial regions by limiting the interference via spreading over the regions having no sources. Spatial filtering is the perfect solution to carry out the source magnitudes in the defined directions if the spherical harmonics' order is not limited. However, the order limitation N depends on the number of microphones, Q, constituting the array. The spatial distribution of the sources around the array starts to become perfect by increasing the order, as in Fig. 3.10. It should be noted that sampling the acoustic scene with thousands of microphones is impossible. Therefore, the back lobe distributions, in other words, interference from other sources, would occur in the observing propagating directions if the sources are positioned in nearly 180° differences in azimuth.

A modified version of RENT is used to group the time-frequency bins into single-source contributing $\mathcal{S}_{\text{RENT}}$, noise/diffuse field $\aleph_{\text{RENT}}$, and valid source $\mathcal{V}_{\text{RENT}}$. This information is used to produce Wiener post-filters for each active source in the valid time-frequency regions to improve the source separation results. Performance degradation occurs when the angle difference between DOAs of sources is small or when the sources are separated by π. The results obtained in Chapter 4 indicate that the PSD-based WPF approach reduces the described degrading effects of the spatial filters and provides better source separation quality in terms of SIR and MUSHRA results. The results also indicate that the post-filtering approach does not significantly affect the SDR and SAR performances.

Joint source DOA estimation and separation are the main motivations for creating the metadata and obtaining separated sources for object-based audio. The results given in Chapter 4 indicate that the best solution for the source separation is SPWD-ASF + WPF. We have observed that WPF significantly improves the separation performance

in comparison with SPWD-FSF and MaxDF without post-filtering. Besides, WPF provides 1.65 dB SIR improvement to the SPWD-ASF case. The separation results data set indicates that PESQ results are statistically significant with respect to SAR and SDR. From the observations, insertion of WPF negatively affects the PESQ score performances of SPWD-ASF and SPWD-FSF. However, it does not have a significant effect on MaxDF. While the nearest interference angle decreases, SIR and PESQ performances were also reduced.

The source count parameter is also statistically significant over the metrics. It is observed that the metrics were reduced with increasing source count. Besides, while higher D/R has a positive effect over the metrics, increasing DOA estimation error does not statistically significant with the performance.

For the noisy case, SIR is degraded for the Wiener post-filtering method while the noise level increases. It was observed that SPWD-ASF with Wiener post-filtering provides the best SIR performances to noise. The average improvement with post-filtering over MaxDF, SPWD-FSF, SPWD-ASF are 9.12 dB, 4.47 dB, and 2.27 dB, respectively, for the noisy cases. These results should be considered when considering the 30 dB sensor noise case, which is a more realistic scenario. We have observed that SPWD-ASF with Wiener post-filtering is robust to 30 dB noise by achieving 24 dB mean SIR and 3.74° DOA estimation error. Besides, SPWD-FSF had the best performance for the source quality by viewing PESQ scores.

SPWD-ASF + WPF was also compared with the state of the art methods in the source separation performance [66] [70]. While the state of the art algorithms and the proposed algorithm result in similar performances for the non-reverberant environments, SPWD-ASF provides the best interference suppression performance for the heavily reverberant environment [7]. SIR is also improved by $9 - 10$ dB with respect to the algorithms for the reverberant case. While it has superior performance in interference suppression, the algorithm slightly decreases the separation quality observed via PESQ. The subjective measurement test MUSHRA also indicates that all of the proposed algorithms in this thesis satisfies the best performances over the baseline MaxDF method. The participants in this subjective test verified that SPWD-ASF + WPF has the best performance in separation compared to the other methods.

A slight SIR improvement for the noise-free conditions with the WPF insertion to the SPWD-ASF has to be assessed for an additional computational cost of post-filtering. While it provides an average of 6-7 dB SIR improvement for MaxDF and SPWD-FSF, it generates a slight improvement for SPWD-ASF by decreasing PESQ scores with this cost addition. Therefore, it is not recommended to insert WPF in the separation task of SPWD-ASF for the cases where computational cost is an important design constraint. Besides, if the sources to be separated are positioned at opposite angles in azimuth, this condition will degrade the spatial filtering performance. Separation of sensor noise available recordings can also be degraded while WPF was not inserted. It is noted from the statistical analysis that, as expected, WPF is needed to obtain more robust solutions while the sensor noise is available.

The quality of the outputs from SPWD-ASF and WPF techniques is related to RENT's performance providing good TF bin groupings. While single-source TF bin grouping is used to generate adaptive filters, noise/diffuse TF bins are used for noise PSD estimation. In this study, RENT covers 0-9 kHz region for the purposes of DOA estimation as well as separation. It works in lower frequency regions due to redundant costs possible from non-existing sources in high-frequency regions. Therefore, ASF's parameters from the lower frequencies up to 9 kHz are used to filter out the sources in the higher frequencies. The SPWD-based source separation algorithms defined in this thesis cover the full frequency range. The post-processing enhancement technique as WPF is not used over higher frequencies to decrease the cost, and the algorithm outputs are directly used. The effect of increasing the RENT coverage region to the full frequency range can be investigated with a group of signals having higher frequency components as future research.

This work provides an acoustic scene analysis tool for generating audio objects with their associated DOAs and separated signals in time. Each of the proposed separation method's performance is investigated with the change of the sources' spatial information and environment-dependent conditions in the controlled setups and statistical analysis. These experiments provide new insight into the relationship between spatial information and the selected source separation techniques' performance. An adapted technique can also be created by changing the separation techniques in time. For example, while the sources are widely separated at the start, they can move and ap-

proach each other in time. By observing the source DOA angles in small intervals, the hybrid method can start with the SPWD-FSF algorithm and then replace it with SPWD-ASF. On the other hand, the sources can have wide-angle differences at the initial point and would start to become broader in azimuth as $180°$. This case would require Wiener post-filtering to decrease interference. Changing the techniques in time instead of using a single separation method from the start is the demanded solution to get good SIR results in terms of SDR and SAR. The tuning parameters (κ, ϱ) can also be used to control the source quality. Future studies should consider creating this adaptation method that would generate more robust solutions in tracking and separating the sources while obtaining good SIR, SDR, SAR, and PESQ results.

CHAPTER 6

CONCLUSION

The work presented for the acoustic scene analysis is a preliminary research to transfer the real theatre, concert or debate recordings to the headsets or loudspeakers for creating immersive audio around the listener. By this motivation, separating the individual audio objects and creating their meta data with spatial information were the main objectives of this thesis. The limitations and observations of the state of the art before this thesis dissertation were :

- Source count parameter is needed for DOA estimation.
- Computational cost was an important factor for real-time DOA estimation problems.
- DOA estimation is needed for coherent source separation.
- Spatial filtering is effective but it degrades in narrow angle differences and cross-talk.
- Active source and noise PSD were an alternative to spatial filtering.
- Source separation has a trade-off with distortion and artifacts.

Main contributions of this thesis with the results build on existing evidence of the evaluations for the identified techniques are :

- Source count parameter can be estimated via HiGRID or RENT.
- Offline storage of directional terms and lowering the number of steering directions without losing accurate DOA estimation lowers the cost.

- Integration of RENT to the source separation phases is possible.
- An alternative to Fixed Spatial Filtering (FSF) is proposed to perform better in narrow angle differences.
- Wiener-post filtering as a post-processing stage using noise statistics estimated from RENT is proposed to enhance the separation performance.
- Tuning parameters (κ, ϱ) can be used to tune the distortion.

According to the built contributions, these techniques can be applied for the scene decomposition over the rigid spherical microphone array recordings or the spherical harmonic coefficients from the HOA. The final solution presented inside the pink box proposed as a problem formulation can be seen in Fig. 6.1.

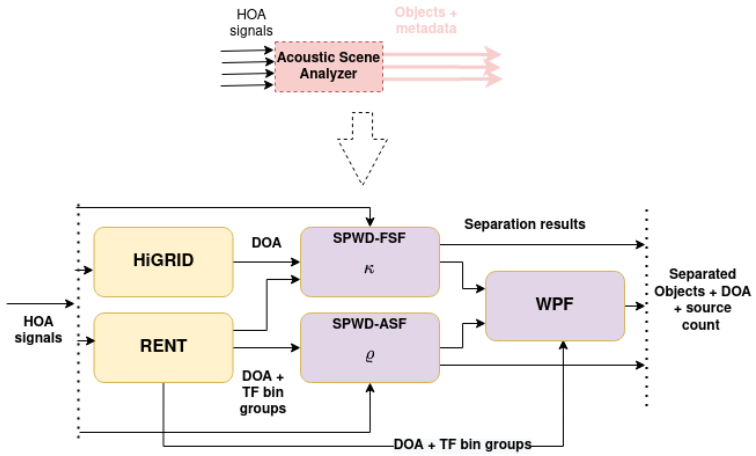


Figure 6.1: Acoustic scene analyzer design for transcoding between scene-based to object-based audio.

If this resulting acoustic scene analyzer is placed after the MPEG-H reference decoder implementation given in Fig. 1.2, the end-user's audio renderer devices can also generate perceptually different audio scenes with respect to the original recorded scene. These scenes would consist of the audio mixtures such as closing a recorded instrument, changing its volume down/up, mixing only a group of objects, or changing the instrument's direction.

Further research is needed to extract the semantic information and distance of the objects from the real recordings. While the clustered source distributions obtained via RENT would provide insights about the relative distance estimations for the sources, this study was beyond the scope of this thesis. Extraction of the semantic information for each source in a separated channel is also possible with the machine learning techniques. That would help to form the audio information into subtitles or identify the audio object type. The acoustic and visual scenes can be integrated using such types of semantic information. As a conclusion, the format change between the scene-based recordings to object-based audio experience will make transferring a real scenario to the end-user devices possible with the ongoing research.

REFERENCES

[1] M. B. Coteli, O. Olgun, and H. Hacihabiboglu, "Multiple sound source localization with steered response power density and hierarchical grid refinement," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2215–2229, 2018.

[2] M. B. Çöteli and H. Hacıhabiboğlu, "Multiple sound source localization with rigid spherical microphone arrays via residual energy test," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 790–794, IEEE, 2019.

[3] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261–270, 2012.

[4] M. B. Coteli and H. Hacihabiboglu, "Acoustic Source Separation Using Rigid Spherical Microphone Arrays Via Spatially Weighted Orthogonal Matching Pursuit," *International Workshop on Acoustic Signal Enhancement*, 2018.

[5] Q. Liu, W. Wang, P. J. Jackson, and T. J. Cox, "A source separation evaluation method in object-based spatial audio," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1088–1092, IEEE, 2015.

[6] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[7] H. H. Orhun Olgun, "*METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0 (Version 0.1.0) [Data set]. Zenodo. @ONLINE*," 2019.

[8] M. Geier, J. Ahrens, and S. Spors, "Object-based audio reproduction and the audio scene description format," *Organised Sound*, vol. 15, no. 3, pp. 219–227, 2010.

[9] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, *et al.*, "Spatial audio object coding (saoc)-the upcoming mpeg standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*, Audio Engineering Society, 2008.

[10] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h audio—the new standard for universal spatial/3d audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830, 2015.

[11] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 135–143, 2004.

[12] H. Kuttruff, *Room Acoustics*. Institut fur Technische Akustik, Aachen, Germany: Spon Press, 5 ed., 2009.

[13] F. Jacobsen, "*Active and reactive sound intensity in a reverberant sound field*," *J. Sound Vib.*, no. 143, pp. 231–240, 1990.

[14] F. Jacobsen, "*A comparison of two different sound intensity measurement principles*," *Acoustical Society of America*, no. 3, pp. 1510–1517, 2005.

[15] M. B. Coteli and H. Hacihabiboglu, "*On the Performance of Acoustic Intensity-based Source Localization with an Open Spherical Microphone Array*," in *Audio Engineering Society 139th Convention*, 2015.

[16] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8. Springer, 2015.

[17] Z. Wang and S. F. Wu, "Helmholtz equation–least-squares method for reconstructing the acoustic pressure field," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2020–2032, 1997.

[18] R. D. Dmitry N. Zotkin and N. A. Gumerov, "*Sound Field Decomposition Using Spherical Microphone Arrays*," in *International Conference on Acoustics, Speech and Signal Processing*, 208.

[19] F. W. Olver and L. C. Maximon, *Bessel functions*. Cambridge University Press, 1960.

[20] C. Müller, *Spherical harmonics*, vol. 17. Springer, 2006.

[21] S. Kazem, S. Abbasbandy, and S. Kumar, "Fractional-order legendre functions for solving fractional-order differential equations," *Applied Mathematical Modelling*, vol. 37, no. 7, pp. 5498–5510, 2013.

[22] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, 2004.

[23] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1781, IEEE, 2002.

[24] mhacoustics, "*em32 Eigenmike microphone array release notes @ONLINE*," October 2013.

[25] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 258–262, IEEE, 2014.

[26] A. H. Moore, C. Evers, P. A. Naylor, A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace pseudointensity vectors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 1, pp. 178–192, 2017.

[27] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization in the spherical harmonic domain using augmented intensity vectors based on grid search," in *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 602–606, IEEE, 2016.

[28] S. Hafezi, A. H. Moore, P. A. Naylor, S. Hafezi, A. H. Moore, and P. A. Naylor, "Augmented intensity vectors for direction of arrival estimation in the spherical harmonic domain," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1956–1968, 2017.

[29] A. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 2296–2300, IEEE, 2015.

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

[31] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[32] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[33] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3d doa estimation of multiple sound sources based on spatially constrained beamforming driven by intensity vectors," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 96–100, IEEE, 2016.

[34] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 702–714, 2007.

[35] H. Do and H. F. Silverman, "A fast microphone array srp-phat source location implementation using coarse-to-fine region contraction (cfrc)," in *2007 IEEE workshop on applications of signal processing to audio and acoustics*, pp. 295–298, IEEE, 2007.

[36] Z. Chu, S. Zhao, Y. Yang, and Y. Yang, "Deconvolution using clean-sc for acoustic source identification with spherical microphone arrays," *Journal of Sound and Vibration*, vol. 440, pp. 161–173, 2019.

[37] S. Zhao, Z. Chu, Y. Yang, and Y. Zhang, "High-resolution clean-sc for acoustic source identification with spherical microphone arrays," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. EL598–EL603, 2019.

[38] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[39] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[40] J. P. Burg, "The relationship between maximum entropy spectra and maximum likelihood spectra," *Geophysics*, vol. 37, no. 2, pp. 375–376, 1972.

[41] O. Olgun and H. Hacihabiboglu, "Localization of multiple sources in the spherical harmonic domain with hierarchical grid refinement and eb-music," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 101–105, IEEE, 2018.

[42] L. Kumar and R. M. Hegde, "Near-field acoustic source localization and beamforming in spherical harmonics domain," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3351–3361, 2016.

[43] Y. Xiao, J. Yin, H. Qi, H. Yin, and G. Hua, "Mvdr algorithm based on estimated diagonal loading for beamforming," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[44] J. Benesty, J. Chen, and Y. Huang, "A generalized mvdr spectrum," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 827–830, 2005.

[45] P. Indyk and M. Ruzic, "Near-optimal sparse recovery in the l1 norm," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 199–207, IEEE, 2008.

[46] D. Xu, N. Hu, Z. Ye, and M. Bao, "The estimate for doas of signals using sparse recovery method," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2573–2576, IEEE, 2012.

[47] J. Li, Z. Li, and X. Zhang, "Partial angular sparse representation based doa estimation using sparse separate nested acoustic vector sensor array," *Sensors*, vol. 18, no. 12, p. 4465, 2018.

[48] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[49] Z. Tan and A. Nehorai, "Sparse direction of arrival estimation using co-prime arrays with off-grid targets," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 26–29, 2013.

[50] T. Noohi, N. Epain, and C. T. Jin, "Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 346–349, IEEE, 2013.

[51] Q. Huang, G. Zhang, and Y. Fang, "Real-valued doa estimation for spherical arrays using sparse bayesian learning," *Signal Processing*, vol. 125, pp. 79–86, 2016.

[52] K. Singhal and R. M. Hegde, "A sparse reconstruction method for speech source localization using partial dictionaries over a spherical microphone array," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[53] G. Ping, E. Fernandez-Grande, P. Gerstoft, and Z. Chu, "Three-dimensional source localization using sparse bayesian learning on a spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 147, no. 6, pp. 3895–3904, 2020.

[54] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Transactions on Audio, speech, and language processing*, vol. 14, no. 2, pp. 666–678, 2006.

[55] T. Virtanen, "Monaural sound source separation by nonnegative matrix factor-

ization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[56] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative nmf and its application to single-channel source separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[57] J. Meyer and T. Agnello, "Spherical microphone array for spatial sound recording," in *Audio Engineering Society Convention 115*, Audio Engineering Society, 2003.

[58] S. Yan, H. Sun, U. P. Svensson, X. Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 361–371, 2010.

[59] N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 238–247, 2013.

[60] Y. Peled and B. Rafaely, "Linearly-constrained minimum-variance method for spherical microphone arrays based on plane-wave decomposition of the sound field," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 12, pp. 2532–2540, 2013.

[61] D. P. Jarrett, E. A. Habets, J. Benesty, and P. A. Naylor, "A tradeoff beamformer for noise reduction in the spherical harmonic domain," in *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*, pp. 1–4, VDE, 2012.

[62] D. P. Jarrett, M. Taseska, E. A. Habets, and P. A. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband doa estimates," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 967–978, 2014.

[63] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an mvdr beamformer and doa-based second-order statistics estimation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 654–658, IEEE, 2013.

[64] Y. Yamamoto and Y. Haneda, "Spherical microphone array post-filtering for reverberation suppression using isotropic beamformings," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2016.

[65] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[66] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Psd estimation of multiple sound sources in a reverberant room using a spherical microphone array," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 76–80, IEEE, 2017.

[67] V. Abolghasemi, S. Ferdowsi, and S. Sanei, "Sparse multichannel source separation using incoherent k-svd method," in *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 477–480, IEEE, 2011.

[68] A. Deleforge and W. Kellermann, "Phase-optimized k-svd for signal extraction from underdetermined multichannel sparse mixtures," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 355–359, IEEE, 2015.

[69] V. Varanasi and R. Hegde, "Stochastic online dictionary learning for speech source localization and separation in spherical harmonic domain," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70, IEEE, 2018.

[70] S. N. Kalkur, S. Reddy C, and R. M. Hegde, "Joint source localization and separation in spherical harmonic domain using a sparsity based method," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[71] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 748–756, 2008.

[72] N. Souviraà-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Multi-channel audio source separation using multiple deformed references," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 1775–1787, 2015.

[73] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[74] M. E. Davies and C. J. James, "Source separation using single channel ica," *Signal Processing*, vol. 87, no. 8, pp. 1819–1832, 2007.

[75] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[76] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 356–360, IEEE, 2019.

[77] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 86–90, IEEE, 2019.

[78] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE signal processing Letters*, vol. 12, no. 10, pp. 713–716, 2005.

[79] D. P. Jarrett and E. A. Habets, "On the noise reduction performance of a spherical harmonic domain tradeoff beamformer," *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 773–776, 2012.

[80] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 36–40, IEEE, 2018.

[81] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: Spherical harmonics domain approach," *Proc. Interspeech 2020*, pp. 5076–5080, 2020.

[82] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1949, IEEE, 2002.

[83] R. Lai and S. Osher, "A splitting method for orthogonality constrained problems," *Journal of Scientific Computing*, vol. 58, no. 2, pp. 431–449, 2014.

[84] R. A. Horn, "The hadamard product," in *Proc. Symp. Appl. Math*, vol. 40, pp. 87–169, 1990.

[85] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, vol. 7, 2013.

[86] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann, "Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere," *The Astrophysical Journal*, vol. 622, no. 2, p. 759, 2005.

[87] M. Batty, "Spatial entropy," *Geographical analysis*, vol. 6, no. 1, pp. 1–31, 1974.

[88] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[89] P. Soille, *Morphological image analysis: principles and applications*. Springer Science & Business Media, 2013.

[90] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007.

[91] S. Mallat, "Wavelets for a vision," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 604–614, 1996.

[92] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*, pp. 40–44, IEEE, 1993.

[93] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

[94] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 255–261, 2005.

[95] A. Koretz and B. Rafaely, "Dolph–Chebyshev beampattern design for spherical arrays," *IEEE Trans. Signal Process.*, vol. 57, pp. 2417–2420, June 2009.

[96] K. V. Mardia, *Statistics of directional data*. London, UK: Academic Press, 2014.

[97] J. T. Kent, "The fisher-bingham distribution on the sphere," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 1, pp. 71–80, 1982.

[98] P. Kasarapu, "Modelling of directional data using kent distributions," *arXiv preprint arXiv:1506.08105*, 2015.

[99] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[100] K. E. Hild, D. Erdogmus, and J. C. Principe, "Experimental upper bound for the performance of convolutive source separation methods," *IEEE transactions on signal processing*, vol. 54, no. 2, pp. 627–635, 2006.

[101] M. Schoeffler, F.-R. Stöter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st Web Audio Conference*, pp. 1–6, 2015.

[102] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752, IEEE, 2001.

[103] Python, "Bss score," 2020.

[104] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[105] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[106] Python, "Pesq score," 2020.

[107] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*, pp. 4214–4217, IEEE, 2010.

[108] Matlab, "Awgn channel," 2020.

[109] B. Olufsen, "Music for archimedes," *Compact disc CD B&O 101*, 1992.

[110] P. ITU-T, "Objective measurement of active speech level," *ITU-T Recommendation*, 1993.

[111] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, vol. 47, 1997.

[112] J. Pätynen, V. Pulkki, and T. Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[113] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[114] "Sound Quality Assessment Material. Recordings for Subjective Tests, CD and User's Handbook for the EBU-SQAM Compact Disc, Tech. 3253-E. European Broadcasting Union (EBU) Technical Centre, Brussels, April 1988., howpublished = http://tech.ebu.ch/publications/sqamcd, note = Accessed: 2010-09-30."

[115] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[116] J. Allan and J. Berg, "Audio level alignment—evaluation method and performance of ebu r 128 by analyzing fader movements," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.

[117] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using wigner-d weighting," *IEEE Signal Processing Letters*, vol. 15, pp. 417–420, 2008.

[118] Eigenmike, "showdown showcase recordings," 2020.

[119] JASP Team, "JASP (Version 0.14)[Computer software]," 2020.

[120] M. A. Shah, I. A. Shah, D.-G. Lee, and S. Hur, "Design approaches of mems microphones for enhanced performance," *Journal of Sensors*, vol. 2019, 2019.

[121] S. Hafezi, A. H. Moore, and P. A. Naylor, "Multiple source localization using estimation consistency in the time-frequency domain," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 516–520, IEEE, 2017.

[122] R. P. Brent, "Fast multiple-precision evaluation of elementary functions," *Journal of the ACM (JACM)*, vol. 23, no. 2, pp. 242–251, 1976.

[123] Intel, "*INTEL software developer manual @ONLINE*," November 2020.

[124] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. De Campos, R. J. Hughes, D. Menzies, M. F. S. Gálvez, Y. Tang, J. Woodcock, *et al.*, "An audio-visual system for object-based audio: from recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1919–1931, 2018.

Appendix A

CONTROLLED SETUPS FOR THE EVALUATION OF SPWD-BASED ALGORITHMS

A.1 Sparse Plane Wave Decomposition with Fixed Spatial Filtering (SPWD-FSF)

The first evaluation setup is the measurement of the effect for the angle difference in azimuth between the source directions to consider the possible limitations of the fixed spatial filtering. Therefore, the sources positioned in 3 scenarios as narrow ($\Delta\phi = 53°$), medium ($\Delta\phi = 105°$) and wide ($\Delta\phi = 180°$) angle differences in azimuth are separated with the SPWD-FSF method. Each source is at 0.3 m above the microphone array in the z-axis with an approximate inclination angle of $\theta = 80°$. The effect of the spread parameter over the source separation performance is checked with three different $\kappa = 4, 6,$ and 10 values. Fig. A.2 presents the example von Mises function distribution used in the algorithm according to different κ values. It is observed that the symmetric distribution becomes narrower as κ increases.

From Table A.1, it is observed that increasing the κ value results in superior SIR performance for both cases except for the sources having an azimuth angle difference of $180°$ while SIR for one source increases and the other source's SIR decreases. The origin of this unexpected result is due to the back-lobe of the MaxDF beam used while obtaining the SRF, as shown in Fig. 3.10. The back lobe reduces the separation efficiency as it allows more leakage from the interfering sources. From Tables A.2 and A.3, the quality of the separation results in the evaluation of SDR and SAR slightly decreases with the increasing κ value.

The second evaluation setup is the measurement of the effect for the angle dif-

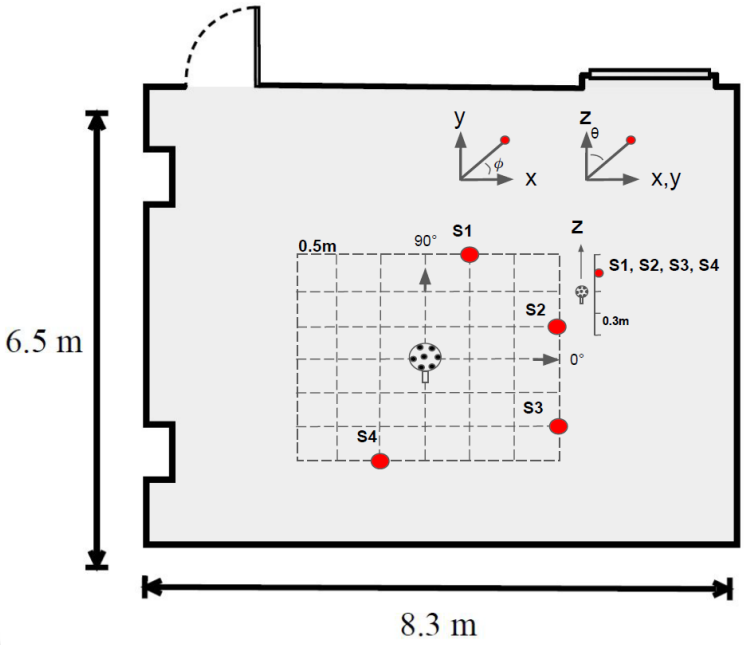


Figure A.1: Scenarios for the evaluation of source separation performances with the change of angle differences. Scenario 1 (S1,S2), Scenario 2 (S1,S3), Scenario 3 (S1,S4)

	MaxDF		SPWD$\kappa = 4$		SPWD$\kappa = 6$		SPWD$\kappa = 10$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	3.01	1.78	9.36	7.21	11.53	9.53	14.25	12.97
Scn. 2 ($\Delta\phi = 105°$)	11.97	8.91	15.88	15.54	16.08	17.60	16.22	20.47
Scn. 3 ($\Delta\phi = 180°$)	9.09	7.11	18.78	9.89	19.64	9.85	20.51	9.38
Avg.±Stdev	$\mathbf{6.98 \pm 3.89}$		$\mathbf{12.78 \pm 4.56}$		$\mathbf{14.04 \pm 4.29}$		$\mathbf{15.63 \pm 4.37}$	

Table A.1: Source Separation SIR (dB) Measurements for SPWD-FSF in different DOA differences (azimuth).

ference in elevation between the source directions to consider the limitations of the fixed spatial filtering. Therefore, the sources positioned in 2 scenarios as narrow ($\Delta\theta \approx 30°$) and wide ($\Delta\theta \approx 60°$) angle differences in elevation are separated with the SPWD-FSF. Each source is 1.0 m away from the microphone array in the x-axis with an azimuth angle of $\phi = 0°$. We have checked the fixed spatial filtering effect using three different $\kappa = 4, 6, 10$ values. The setup designed for the evaluation con-

	MaxDF		SPWD $\kappa = 4$		SPWD $\kappa = 6$		SPWD $\kappa = 10$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	-2.00	-2.35	-0.36	-0.16	-0.41	0.17	-0.61	0.56
Scn. 2 ($\Delta\phi = 105°$)	-0.34	0.0	0.54	1.86	0.22	1.87	-0.21	1.81
Scn. 3 ($\Delta\phi = 180°$)	-1.75	-2.09	0.10	-0.41	0.08	-0.64	0.05	-1.09
Avg.±Stdev	$\mathbf{-1.42 \pm 0.99}$		$\mathbf{0.26 \pm 0.85}$		$\mathbf{0.22 \pm 0.88}$		$\mathbf{0.09 \pm 1.01}$	

Table A.2: Source Separation SDR (dB) Measurements for SPWD-FSF in different DOA differences (azimuth).

	MaxDF		SPWD $\kappa = 4$		SPWD $\kappa = 6$		SPWD $\kappa = 10$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn.1 ($\Delta\phi = 53°$)	1.39	1.97	0.59	1.46	0.17	1.17	-0.30	1.03
Scn.2 ($\Delta\phi = 105°$)	0.18	1.12	0.78	2.17	0.44	2.07	-0.01	1.91
Scn.3 ($\Delta\phi = 180°$)	-0.93	-0.76	0.22	0.42	0.18	0.18	0.13	-0.21
Avg±Stdev	$\mathbf{0.50 \pm 1.18}$		$\mathbf{0.94 \pm 0.73}$		$\mathbf{0.70 \pm 0.77}$		$\mathbf{0.43 \pm 0.86}$	

Table A.3: Source Separation SAR (dB) Measurements for SPWD-FSF in different DOA differences (azimuth).

tains two scenarios given in Fig. A.3. SIR, SDR and SAR values measured after the SPWD-FSF is processed for the separation are given in Tables A.4, A.5, A.6.

	MaxDF		SPWD $\kappa = 4$		SPWD $\kappa = 6$		SPWD $\kappa = 10$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	1.96	0.76	5.93	2.72	7.04	3.55	8.57	4.97
Scn. 2 ($\Delta\theta \approx 60°$)	2.69	4.03	7.80	10.02	9.29	11.69	11.27	13.82
Avg.±Stdev	$\mathbf{2.36 \pm 1.36}$		$\mathbf{6.61 \pm 3.08}$		$\mathbf{7.89 \pm 3.46}$		$\mathbf{9.65 \pm 3.78}$	

Table A.4: Source Separation SIR (dB) Measurements for SPWD-FSF in different DOA differences (elevation).

It is observed that SIR is significantly improved via SPWD-FSF, in comparison with MaxDF. Besides, SIR in the wide-angle case ($\Delta\theta \approx 60°$) is improved by $5 - 6$ dB on

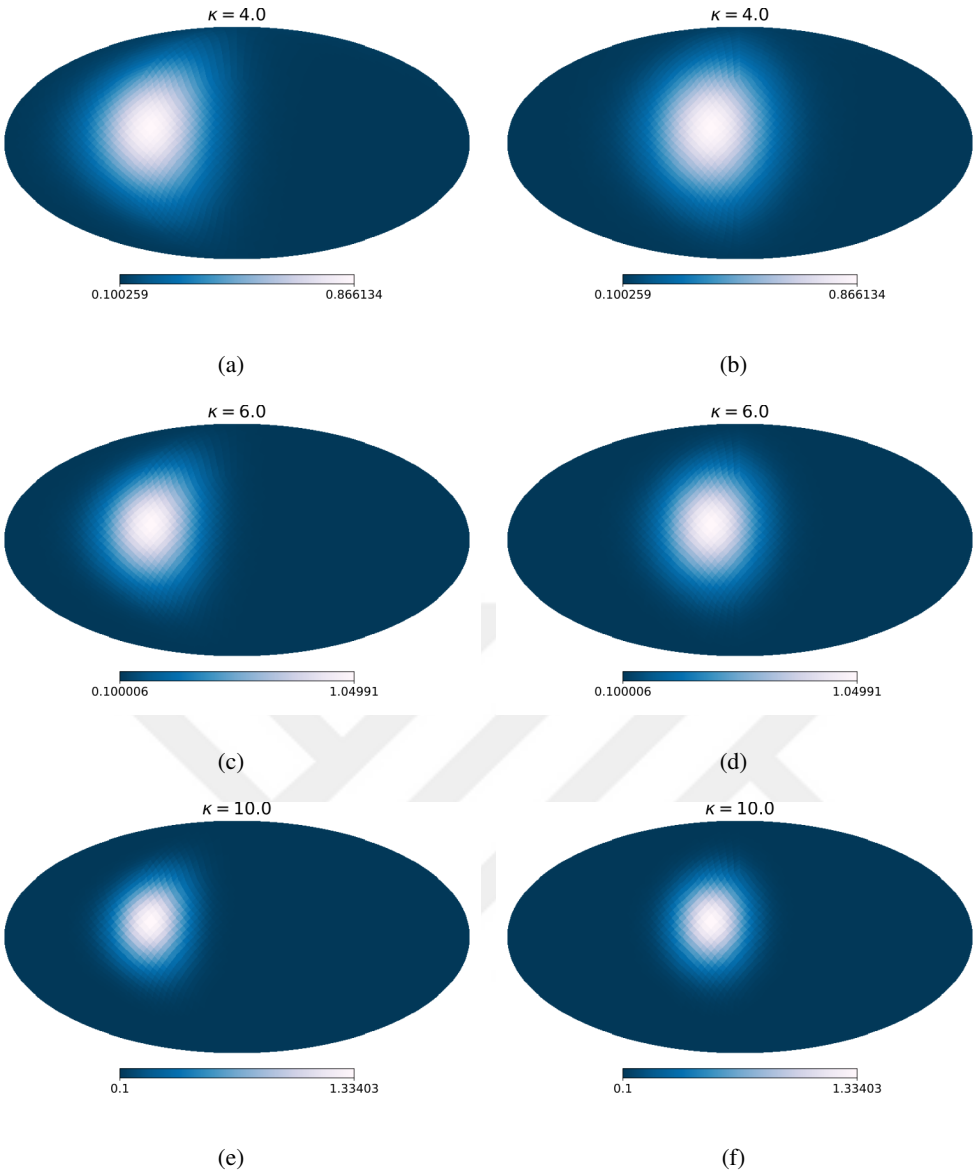


Figure A.2: von Mises spatial filters calculated for two different sources located in the angle differences of $(\Delta\phi = 53°)$. Scenario 1 : (S1, S2) at $\kappa = 4$, $\kappa = 6$, $\kappa = 10$ from top to bottom.

average when compared to narrow separation case $(\Delta\theta \approx 30°)$. A degradation in SIR similar to what was observed in the case of emulating a narrow angular separation is also seen for sources at different elevations. The algorithm also satisfies better SDR and SAR performances in comparison with MaxDF. Increasing the κ value also provides better SIR values, but it slightly decreases SDR and SAR values.

	MaxDF		SPWD $\kappa = 4$		SPWD $\kappa = 6$		SPWD $\kappa = 10$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	-2.29	-2.82	-0.44	-1.37	-0.43	-1.05	-0.69	-0.66
Scn. 2 ($\Delta\theta \approx 60°$)	-2.31	-1.26	-0.18	1.00	-0.12	1.16	-0.30	1.12
Avg$\pm$Stdev	$\mathbf{-2.17 \pm 0.65}$		$\mathbf{-0.24 \pm 0.97}$		$\mathbf{-0.12 \pm 0.93}$		$\mathbf{-0.13 \pm 0.85}$	

Table A.5: Source Separation SDR (dB) Measurements for SPWD-FSF in different DOA differences (elevation).

	MaxDF		SPWD $\kappa = 4$		SPWD $\kappa = 6$		SPWD $\kappa = 10$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	1.89	2.30	1.68	2.62	1.19	2.38	0.41	1.92
Scn. 2 ($\Delta\theta \approx 60°$)	1.19	1.69	1.23	1.99	0.88	1.85	0.32	0.79
Avg$\pm$Stdev	$\mathbf{1.76 \pm 0.46}$		$\mathbf{1.88 \pm 0.58}$		$\mathbf{1.57 \pm 0.67}$		$\mathbf{0.86 \pm 0.73}$	

Table A.6: Source Separation SAR (dB) Measurements for SPWD-FSF in different DOA differences (elevation).

The algorithm provides a better SIR, SDR, and SAR performance than the MaxDF algorithm for both cases. However, an adaptive spatial filter is required to estimate the source distribution for the conditions of narrow-angle differences and significant D/R ratio differences.

A.2 Sparse Plane Wave Decomposition with Adaptive Spatial Filtering (SPWD-ASF)

The first evaluation setup consists of three different scenarios as the previously identified scenarios given in Fig. A.1. The source DOA estimation is obtained using RENT. The frequency range and threshold selected for RENT analysis are between $0 - 9000$ Hz and 0.5, respectively. OMP-based signal extraction covers the whole frequency region as $0 - 24$ kHz, where the sampling rate is selected as 48 kHz. FFT window size and overlap are 2048 and 50%, respectively. Two sources identified in consecutive analysis intervals are associated if their estimated DOAs have less

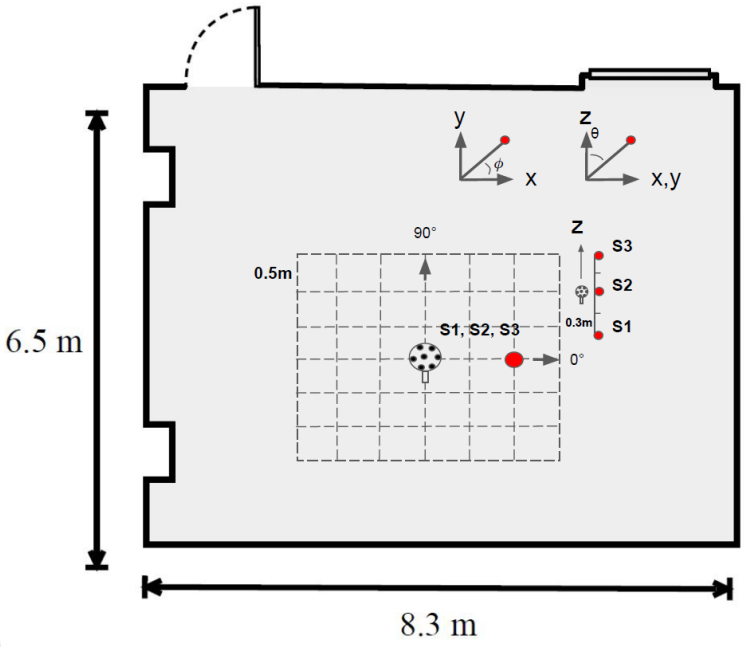


Figure A.3: Top views of the two scenarios used in the evaluations. Red circles represent source positions. RSMA is positioned at the center. Scenario 1 : (S1, S2) in $(\Delta\theta \approx 30°)$ angle difference and Scenario 1 : (S1, S3) $(\Delta\theta \approx 60°)$ angle difference.

than $\Psi = 18°$ separation. Reference signals for the measurement of SIR, SDR, and SAR [99] are calculated via obtaining the omni-directional component of the acoustic scene where there is only one single source active in the same position.

It was noted that the SPWD-FSF algorithm has lower SIR for the condition where sources are positioned in narrow-angle differences (seen in Table A.1). The modified solution's motivation is to propose an improvement for these types of conditions to generate an adaptive spatial filter. A general comparison is obtained to investigate SIR, SDR, and SAR change for the same scenarios. Tables A.7, A.8 and A.9 represent the algorithm's performances in the specified angle differences with the comparison among the changes in the dilation parameter (ϱ). The parameter defined for Kent distribution and fixed $\kappa = 4$ in SPWD-FSF are compared.

From Tables A.7, A.8 and A.9, it is seen that adaptive spatial filtering with different values of dilation parameters ($\varrho = 0.5, 1.0, 2.0$) outperforms in the condition where acoustic sources are positioned in narrow angles when compared to SPWD-

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	9.36	7.21	14.54	9.10	17.12	17.53	19.57	19.87
Scn. 2 ($\Delta\phi = 105°$)	15.88	15.54	17.04	17.31	18.92	21.42	21.35	23.68
Scn. 3 ($\Delta\phi = 180°$)	18.78	9.89	21.83	9.92	22.79	8.82	23.28	7.73
Avg.$\pm$Stdev	$\mathbf{12.78 \pm 4.56}$		$\mathbf{14.96 \pm 4.83}$		$\mathbf{17.77 \pm 4.90}$		$\mathbf{19.30 \pm 5.91}$	

Table A.7: SIR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences via evaluating the spread parameter (ϱ).

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	-0.36	-0.16	-0.74	-1.26	-1.12	1.06	-1.56	1.44
Scn. 2 ($\Delta\phi = 105°$)	0.54	1.86	-0.44	1.75	-0.85	1.46	-1.21	0.98
Scn. 3 ($\Delta\phi = 180°$)	0.10	-0.41	-0.06	-1.44	-0.09	-2.27	-0.22	-3.70
Avg.$\pm$Stdev	$\mathbf{0.26 \pm 0.85}$		$\mathbf{-0.37 \pm 1.15}$		$\mathbf{-0.30 \pm 1.40}$		$\mathbf{-0.71 \pm 1.87}$	

Table A.8: SDR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences via evaluating the spread parameter (ϱ).

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1($\Delta\phi = 53°$)	0.59	1.46	-0.46	-0.33	-0.97	1.24	-1.16	1.02
Scn. 2 ($\Delta\phi = 105°$)	0.78	2.17	-0.28	1.95	-0.75	1.53	-1.48	1.55
Scn. 3 ($\Delta\phi = 180°$)	0.22	0.42	-0.01	-0.69	-0.05	-1.38	-0.18	-2.70
Avg.$\pm$Stdev	$\mathbf{0.94 \pm 0.73}$		$\mathbf{0.03 \pm 0.96}$		$\mathbf{-0.06 \pm 1.20}$		$\mathbf{-0.49 \pm 1.60}$	

Table A.9: SAR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences via evaluating the spread parameter (ϱ).

FSF. However, SDR and SAR slightly decrease in 0-1 dB. Adaptive spatial filters fitted with different acoustic source distribution with their corresponding angles are shown in Figure A.4. ASF also degrades in the condition where sources are positioned

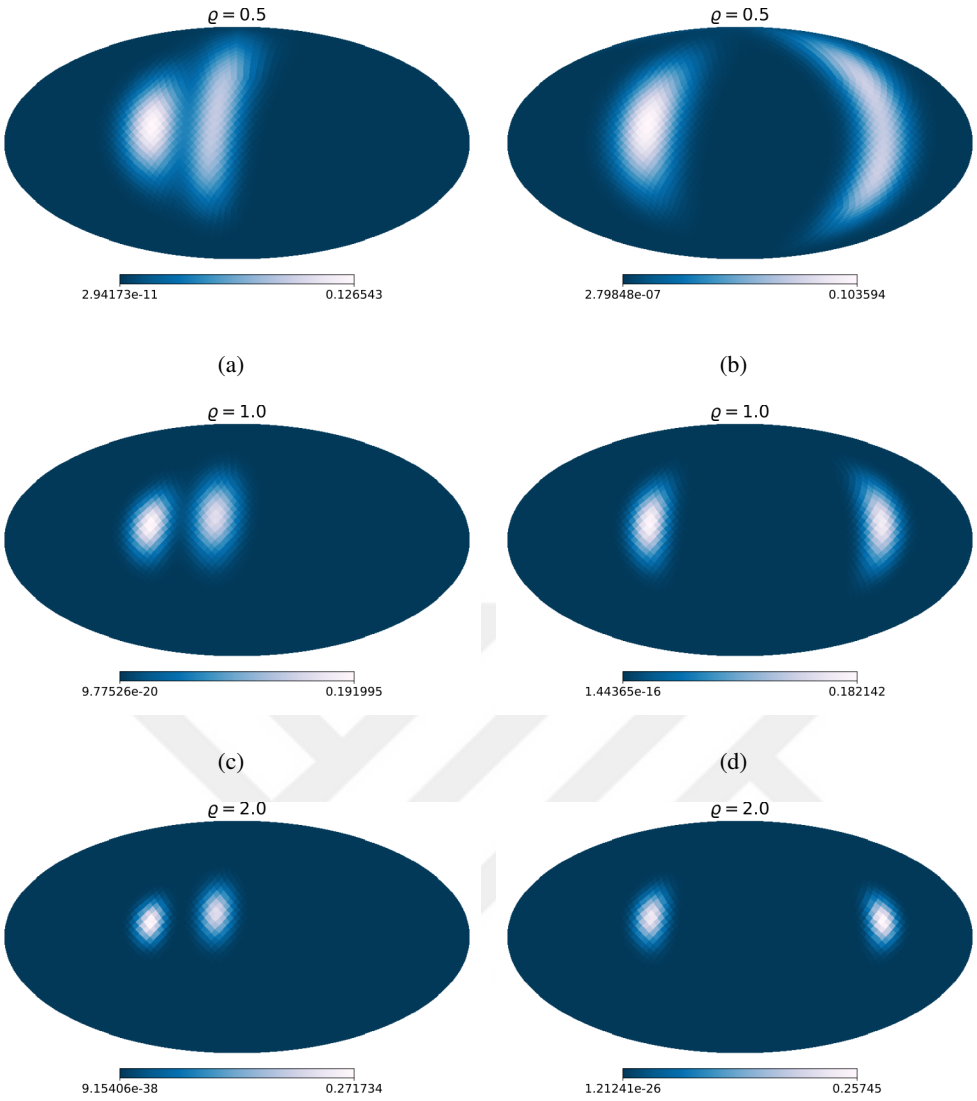


Figure A.4: Adaptive spatial filters ($\varrho = 0.5, \varrho = 1.0, \varrho = 2.0$) calculated for two sources in the angle differences of ($\Delta\phi = 53°$) and ($\Delta\phi = 180°$) from left to right. Scenario 1 : (S1, S2), Scenario 2 : (S1, S4)

in wide azimuth angles of $180°$ because of cross-talk occurring due to the sidelobes of the employed Legendre kernels given in Fig. 3.10.

The second evaluation setup is the measurement of the effect for the angle difference in elevation. Therefore, the sources positioned in 2 scenarios as narrow ($\Delta\theta \approx 30°$) and wide ($\Delta\theta \approx 60°$) angle differences in elevation are separated with the SPWD-

ASF. Each source is 1.0 m away from the microphone array in the x-axes with an angle of $\phi = 0°$. We have checked the effect of the spread parameter for adaptive spatial filtering with three different $\varrho = 0.5, 1.0, 2.0$ values. The setup designed for the evaluation contains two scenarios given in Fig. A.3. SIR, SDR and SAR values measured after the SPWD-ASF processing for the separation are given in Tables A.10, A.11, A.12.

It may be observed that SIR is significantly improved by using SPWD-ASF, according to SPWD-FSF. Besides, SIR in the narrower angle case ($\Delta\theta \approx 30°$) is improved by $3 - 4$ dB in average when compared to narrower ($\Delta\theta \approx 30°$) of SPWD-FSF. The adaptive spatial filtering solution outperforms in the narrower and wider angle difference cases. However, the algorithm does not satisfy better SDR and SAR performances according to the SPWD-FSF. Measured SIR in the increasing ϱ value also provides better SIR values but slightly decreases SDR and SAR values.

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	5.93	2.72	7.54	8.87	8.97	9.62	10.25	11.62
Scn. 2 ($\Delta\theta \approx 60°$)	7.80	10.02	7.27	15.22	11.50	18.22	14.39	19.82
Avg.$\pm$Stdev	$\mathbf{6.61 \pm 3.08}$		$\mathbf{9.72 \pm 3.72}$		$\mathbf{12.07 \pm 4.23}$		$\mathbf{14.02 \pm 4.23}$	

Table A.10: Source Separation SIR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences (in elevation) via evaluating the spread parameter (ϱ).

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	-0.44	-1.37	-1.10	0.04	-2.06	-0.54	-3.02	-1.04
Scn. 2 ($\Delta\theta \approx 60°$)	-0.18	1.00	-0.63	1.06	-1.32	0.29	-2.78	-0.22
Avg.$\pm$Stdev	$\mathbf{-0.24 \pm 0.97}$		$\mathbf{-0.15 \pm 0.93}$		$\mathbf{-0.90 \pm 1.01}$		$\mathbf{-1.76 \pm 1.35}$	

Table A.11: Source Separation SDR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences (in elevation) via evaluating the spread parameter (ϱ).

	SPWD-F$\kappa = 4$		SPWD-A$_{0.5}$		SPWD-A$_{1.0}$		SPWD-A$_{2.0}$	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	1.68	2.62	0.23	1.17	-1.19	0.34	-2.42	0.51
Scn. 2 ($\Delta\theta \approx 60°$)	1.23	1.99	0.87	1.36	-0.79	0.43	-2.54	-0.13
Avg.$\pm$ Stdev	$\mathbf{1.88 \pm 0.58}$		$\mathbf{0.90 \pm 0.49}$		$\mathbf{-0.30 \pm 0.81}$		$\mathbf{-1.14 \pm 1.56}$	

Table A.12: Source Separation SAR (dB) Measurements for SPWD-FSF and SPWD-ASF in different DOA differences (in elevation) via evaluating the spread parameter (ϱ).

A.3 Wiener-post filtering (WPF)

The first evaluation setup consists of three different scenarios as the previously identified scenarios given in Fig. A.1. The algorithm parameters are set as in the previously identified scenario (given in SPWD-FSF and SPWD-ASF).

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	9.36	7.21	15.47	10.57	17.12	17.53	20.20	15.93
Scn. 2 ($\Delta\phi = 105°$)	15.88	15.54	24.48	17.16	18.92	21.42	27.25	21.77
Scn. 3 ($\Delta\phi = 180°$)	18.78	9.89	21.78	17.70	22.79	8.82	22.94	19.80
Av.$\pm$Stdev	$\mathbf{12.78 \pm 4.56}$		$\mathbf{17.86 \pm 4.86}$		$\mathbf{17.77 \pm 4.90}$		$\mathbf{21.32 \pm 3.75}$	

Table A.13: SIR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (azimuth).

From Tables A.13, A.14 and A.15, it can be observed that Wiener post-filtering outputs good separation results when compared to SPWD-FSF and SPWD-ASF, while the acoustic sources are positioned in narrow angles, and they are at the opposite directions, which would create interference. However, SDR and SAR slightly decrease in $0 - 1$ dB. The WPF's main contribution is generating a weighting function based on the signals' PSD. Beyond the spatial information, weighting the sources based on their power can create more robust separation results without thinking about the effect of spatial locations.

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	-0.36	-0.16	0.15	-0.44	-1.12	1.06	-0.70	-0.89
Scn. 2 ($\Delta\phi = 105°$)	0.54	1.86	1.18	1.74	-0.85	1.46	-0.20	0.20
Scn. 3 ($\Delta\phi = 180°$)	0.10	-0.41	-0.49	1.02	-0.09	-2.27	-2.20	-0.97
Avg.$\pm$Stdev	$\mathbf{0.26 \pm 0.85}$		$\mathbf{0.53 \pm 0.92}$		$\mathbf{-0.30 \pm 1.40}$		$\mathbf{-0.76 \pm 0.82}$	

Table A.14: SDR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (azimuth).

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-4	S1	S2-4	S1	S2-4	S1	S2-4
Scn. 1 ($\Delta\phi = 53°$)	0.59	1.46	0.40	0.27	-0.97	1.24	-0.62	-0.69
Scn. 2 ($\Delta\phi = 105°$)	0.78	2.17	1.22	1.95	-0.75	1.53	-0.18	0.26
Scn. 3 ($\Delta\phi = 180°$)	0.22	0.42	-0.44	1.19	-0.05	-1.38	-2.16	-0.89
Avg.$\pm$Stdev	$\mathbf{0.94 \pm 0.73}$		$\mathbf{0.77 \pm 0.85}$		$\mathbf{-0.06 \pm 1.20}$		$\mathbf{-0.71 \pm 0.82}$	

Table A.15: SAR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (azimuth).

The second evaluation setup is the measurement of the effect for the angle difference in elevation among the source directions to investigate the impacts of WPF. Therefore, the setup designed for the evaluation contains two scenarios given in Fig. A.3. The scenarios are separated with the SPWD-FSF, SPWD-FSF + WPF, SPWD-ASF, SPWD-ASF + WPF algorithms. SIR, SDR, and SAR values measured after the WPF application for the separation are given in Tables A.16, A.17, A.18.

It can be observed that using WPF slightly improves SIR in comparison with SPWD-ASF by $1-2$ dB. Besides, SIR is significantly improved via WPF addition in comparison with SPWD-FSF by $4-5$ dB. However, the post-filtering does not provide better SDR and SAR performances than the SPWD-FSF without WPF. Measured SIR in the increasing angle differences also results in better SIR values and SDR and SAR.

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	5.93	2.72	9.14	5.47	8.97	9.62	11.11	9.43
Scn. 2 ($\Delta\theta \approx 60°$)	7.80	10.02	12.03	13.01	11.50	18.22	14.72	17.60
Av.$\pm$ Stdev	$\mathbf{6.61 \pm 3.08}$		$\mathbf{9.91 \pm 3.38}$		$\mathbf{12.07 \pm 4.23}$		$\mathbf{13.21 \pm 3.66}$	

Table A.16: Source Separation SIR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (inclination).

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	-0.44	-1.37	-0.65	-0.67	-2.06	-0.54	-2.08	-2.06
Scn. 2 ($\Delta\theta \approx 60°$)	-0.18	1.00	0.32	0.88	-1.32	0.29	0.89	1.37
Av.$\pm$Stdev	$\mathbf{-0.24 \pm 0.97}$		$\mathbf{-0.03 \pm 0.76}$		$\mathbf{-0.90 \pm 1.01}$		$\mathbf{-0.47 \pm 1.85}$	

Table A.17: Source Separation SDR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (inclination).

	(1) SPWD$\kappa = 4$		(1) w. WPF		(2) SPWD$_{1.0}$		(2) w. WPF	
	S1	S2-3	S1	S2-3	S1	S2-3	S1	S2-3
Scn. 1 ($\Delta\theta \approx 30°$)	1.68	2.62	0.32	1.62	-1.19	0.34	-1.54	-1.28
Scn. 2 ($\Delta\theta \approx 60°$)	1.23	1.99	0.89	1.37	-0.79	0.43	-0.94	-0.95
Av.$\pm$Stdev	$\mathbf{1.88 \pm 0.58}$		$\mathbf{1.05 \pm 0.57}$		$\mathbf{-0.30 \pm 0.81}$		$\mathbf{-1.17 \pm 0.28}$	

Table A.18: Source Separation SAR (dB) Measurements for SPWD-FSF$\kappa = 4$ and SPWD-ASF$_{1.0}$ with and without WPF in different DOA differences (inclination).