Sound Source Localization and Modeling: Spherical Harmonics Domain Approaches

Yonggang Hu

B.Sc. Engineering, University of Science and Technology, Nanjing, China. 2013

January 2021

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University

Research School of Engineering
College of Engineering and Computer Science
The Australian National University

This thesis is dedicated to all the friends, teachers, family members and my wife.

Declaration

The contents of this thesis are the results of original research and have not been submitted for a higher degree to any other university or institution. Much of this work has either been published or submitted for publications as journal papers and conference proceedings. Following is a list of these papers.

Journal Publications

- Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 28, pp. 3108-3123, 2020.

- Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe, "Multiple source direction of arrival estimations using relative sound pressure based MUSIC", *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 29, pp. 253-264, 2021.

Conference Proceedings

- Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling the interior response of real loudspeakers with finite measurements", in 2018 *IEEE 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 16-20.

- Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-

domain approaches", in 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561-565.

- Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using relative harmonic coefficients in modal domain", in 2019 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 348-352.

- Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using realtive harmonic coefficients", in 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571-575.

- Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, and S. Gannot, "Decoupled direction-of-arrival estimations using relative harmonic coefficients", in 2020 *IEEE 28th European Signal Processing Conference (EUSIPCO)*, 246-250.

- Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: spherical harmonics domain approach", in 2020 *INTERSPEECH*, 5076-5080.

The following papers are also results from my Ph.D. study, but not included in this thesis:

Journal Publications

- Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Direct-path relative harmonic coefficients identification from multiple simultaneous speakers", *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, to be submitted, 2021.

- Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Decoupled Multiple Speaker Direction-of-Arrival Estimations Under Reverberant Environments", *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, to be submitted, 2021.

Conference Proceedings

- Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Evaluation and comparison of three source direction-of-arrival estimators using relative harmonic coefficients", in 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, accepted.

The research work presented in this thesis has been performed jointly with Dr. P. N. Samarasinghe, Prof. Thushara D. Abhayapala, Prof. Sharon Gannot and Dr. Glenn Dickins. Approximately 80% of this work is my own.

Yonggang Hu
Audio and Acoustic Signal Processing Group
Research School of Engineering
Australian National University
Canberra ACT 2600

Acknowledgments

Without the support of the many faces in my life, this work would not have been completed. Here, I would like to say thanks very much to each of them as follows:

First and foremost, my supervisors, Dr. Prasanga Samarasinghe and Prof. Thushara Abhayapala, for their professional guidance and true friendship during the research at Australian National University. Special thanks go to Prasanga. Being my chair, she ever gave me sufficient support and care, with great patience, throughout the years. Her help and support make my life easier and brighter. Thanks very much for Thushara. He ever led me to join meaningful projects, trained me to an independent researcher, and instructed me on how to be a good writer. His great help benefits me forever.

Many thanks go to Prof. Sharon Gannot for his professional guidance and detailed comments on our academic publications. The experience, learned during the continuous collaborations with him, greatly motivates me in my future academic career.

Prof. Wen Zhang recruited me to this Ph.D. program and gave me great support when applying for the fund. She also gave me sufficient guidance and support at the beginning of my research at the Australian National University.

Thanks very much to Dr. Glenn Dickins who led me into a commercial project in Dolby Sydney. The experience learned from this engineering project greatly enriched my Ph.D. career.

The Australian National University, for giving me this Ph.D. opportunity, HDR Fee Remission Merit Scholarship, and all the administrative and IT supports.

The China Scholarship Council who supports me sufficient fund for a living and the necessary administrative help.

My friends and partners in the audio and acoustic signal processing group,

including Hanchi, Xiang, Fahim, Aimee, Fei, Huanyu, Noman, Lachlan, Huiyuan, Wageesha, and Felix.

My parents, for their unconditional love, care, and encouragement.

My hardworking days and nights over the years in the Australian Nation University. They are the time I will remember forever.

Finally, to my wife Jiajia, for her love, care, patience, and waiting every day over so many years.

Abstract

Sound source localization has been an important research topic in the acoustic signal processing community because of its wide use in many acoustic applications, including speech separation, speech enhancement, sound event detection, automatic speech recognition, automated camera steering, and virtual reality. In the recent decade, there is a growing interest in the research of sound source localization using higher-order microphone arrays, which are capable of recording and analyzing the soundfield over a target spatial area. This thesis studies a novel source feature called the *relative harmonic coefficient*, that easily estimated from the higher-order microphone measurements. This source feature has direct applications for sound source localization due to its sole dependence on the source position.

This thesis proposes two novel sound source localization algorithms using the relative harmonic coefficients: (i) a low-complexity single source localization approach that localizes the source' elevation and azimuth separately. This approach is also appliable to acoustic enhancement for the higher-order microphone array recordings; (ii) a semi-supervised multi-source localization algorithm in a noisy and reverberant environment. Although this approach uses a learning schema, it still has a strong potential to be implemented in practice because only a limited number of labeled measurements are required. However, this algorithm has an inherent limitation as it requires the availability of single-source components. Thus, it is unusable in scenarios where the original recordings have limited single-source components (e.g., multiple sources simultaneously active). To address this issue, we develop a novel MUSIC framework based approach that directly uses simultaneous multi-source recordings. This developed MUSIC approach uses robust measurements of relative sound pressure from the higher-order microphone and is shown to be more suitable in noisy environments than the traditional MUSIC method.

While the proposed approaches address the source localization problems, in practice, the broader problem of source localization has some more common challenges, which have received less attention. One such challenge is the common assumption of the sound sources being omnidirectional, which is hardly the case with a typical commercial loudspeaker. Therefore, in this thesis, we analyze the broader problem of analyzing directional characteristics of the commercial loudspeakers by deriving equivalent theoretical acoustic models. Several acoustic models are investigated, including plane waves decomposition, point source decomposition, and mixed source decomposition. We finally conduct extensive experimental examinations to see which acoustic model has more similar characteristics with commercial loudspeakers.

List of Acronyms

AIR	Acoustic Impulse Response
ATF	Acoustic Transfer Function
AWGN	Additive White Gaussian Noise
BSS	Blind Source Separation
CNN	Convolutional Neural Networks
CRNN	Convolutional Recurrent Neural Network
CPSD	Cross Power Spectral Density
DOA	Direction of Arrival
DNN	Deep Neural Network
EM	Expectation Maximization
ESPRIT	Estimation of Signal Parameter Via Rotational Invariance Technique
GPR	Gaussian Process Regression
GCC-PHAT	Generalized Cross-Correlation with Phase Transform
HRTF	Head-Related Transfer Function
ICA	Independent Component Analysis
ISTFT	Inverse Short-time Fourier Transform
IRLS	Iteratively Reweighted Least Squares
LASSO	Least Absolute Shrinkage and Selection Operator
MUSIC	Multiple Signal Classification
MP	Matching Pursuit
MVDR	Minimum Variance Distortionless Response
MAEE	Mean Absolute Estimated Error
MMGP	Multi-Mode Gaussian Process

PSD	Power Spectral Density
ReTF	Relative Transfer Function
RIR	Room Impulse Response
RHC	Relative Harmonic Coefficients
RSP	Relative Sound Pressure
RMUSIC	Relative Sound Pressure Based MUSIC
SNR	Signal to Noise Ratio
SHD	Spherical Harmonics Domain
SRP	Steered Response Power
SHD-RMUSIC	Spherical Harmonics Domain RMUSIC
SVD	Singular Eigenvalue Decomposition
SRP-PHAT	Steered Response Power Phase Transform
STFT	Short-time Fourier Transform
TDOA	Time Difference of Arrival
WDO	W-Disjoint Orthogonality

Notations and Symbols

i	$\sqrt{-1}$
$\mathrm{Re}\{\cdot\}$	Real part
$\mathrm{Im}\{\cdot\}$	Imaginary part
$\delta\{\cdot\}$	Dirac delta function
$\delta_{nm}\{\cdot\}$	Kronecker delta function
N	Order of soundfield
$E\{\cdot\}$	Expectation operator
$\lceil\cdot\rceil$	Ceiling operator
$[\cdot]^T$	Transpose of a matrix
$[\cdot]^*$	Complex conjugate of a matrix
$\lvert\cdot\rvert$	Euclidean norm of a vector
$[\cdot]^H$	Complex conjugate transpose of a matrix
$\boldsymbol{x}\cdot\boldsymbol{y}$	Dot product between two vectors
$\mathbb{R}^\ell$	ℓ dimensional real number space
$\boldsymbol{A}^\dagger$	Matrix psuedoinverse: $\boldsymbol{A}^\dagger = [\boldsymbol{A}^H\boldsymbol{A}]^{-1}\boldsymbol{A}^H$
$j_n(\cdot)$	Spherical Bessel functions of the first kind
$h_n(\cdot)$	Spherical Hankel functions of the first kind
$Y_{nm}(\cdot)$	Spherical harmonics function with order n and degree m
$P_{nm}(\cdot)$	Associated Legendre function with order n and degree m
$j_n^{'}(\cdot)$	Partial derivative of spherical Bessel function of the first kind
$h_n^{'}(\cdot)$	Partial derivative of spherical Hankel function of the first kind

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Scope

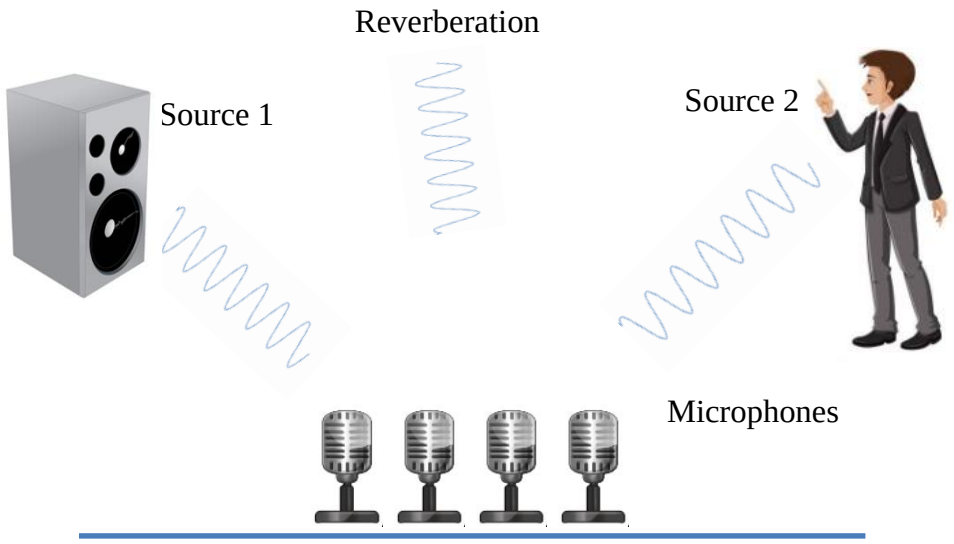


Figure 1.1: Sound source localization using a set of microphones.

Sound source localization is the task to use several sensors/microphones to accurately estimate the unknown spatial positions, e.g., the direction of arrival (DOA), of all the sound sources presented in the environment (e.g., see the acoustic event in Fig. 1.1). For a couple of decades, it has been an important research topic

in the acoustic signal processing community, due to its wide usage in many spatial acoustic techniques and applications [3], such as speech separation [4], speech enhancement [5], sound event detection [6], acoustic beamforming [7], automatic speech recognition [8], and automated camera steering [9]. Recent applications of sound source localization include teleconferencing systems, mobile devices, and virtual reality systems [10]. A recent challenge on acoustic source localization and tracking (LOCATA) [11] endorsed by the IEEE Audio and Acoustic Signal Processing technical committee, is proof of the academic interest on this topic across the world.

(a) (b)

Figure 1.2: (a): a commercial spherical microphone array called EigenMike and (b) a planar microphone array [2].

In the recent decade, there is a growing interest in the research community to use higher-order microphone arrays (e.g., a spherical and planar microphone array in Fig. 1.2) to address the sound source localization challenges [12, 13, 14]. The higher-order microphone arrays have the advantage that they are capable of fully capturing the soundfield over a whole three-dimensional area. The multi-channel measurements of the higher-order microphone arrays can be decomposed into the spherical harmonics domain (i.e., the modal domain in [15]) using a set of orthogonal spatial basis functions [16]. A spherical harmonic decomposition of the measured soundfield has several advantages, such as the decoupling of frequency-dependent and angular-dependent components [17]. Lots of early sound source localization methods, such as the MUltiple SIgnal Classification (MUSIC) [18] and

estimation of signal parameters via rotational invariance techniques (ESPRIT) [19], have been decomposed into the spherical harmonics domain for improved performance [17,20,21]. From the recent literature, we have witnessed significant progress of the spherical harmonics domain source localization. However, we see that there remain some challenging problems with sufficient space for further improvement, and in this thesis, we address the following remaining issues:

- *Spherical harmonics domain source features*: typical source localization approaches generally comprise of three steps below (i.e., see Fig. 1.3): (i) recording the soundfield of interest using the microphone array; (ii) estimating a source feature from the measurements of the microphone array; and (iii) feeding the estimated source feature to the localization algorithm and searching over the directional space to recover the unknown source location(s). Intuitively, the source features, taken as the inputs of the algorithms, are vital to the localization accuracy as they contain relevant characteristics of the sound source(s) to be localized. Although some features in the spherical harmonics domain are available, such as spatial correlation matrix of the measured spherical harmonic coefficients [17], first-order ambisonics [22] and modal coherence patterns [23], there is still of great interest to develop the source feature with remarkable properties for sound source localization, such as a direct/close relation to the source position and easy estimations in complex acoustic environments.

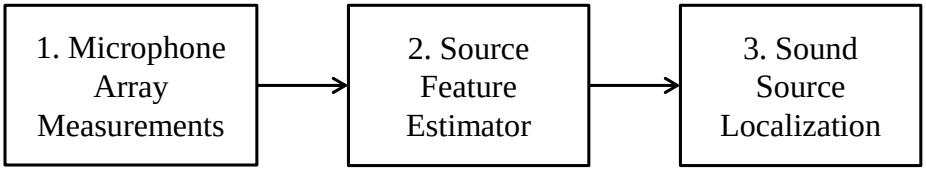


Figure 1.3: General steps of sound source localization algorithms.

- *single source localization*: localization of a single sound source is a fundamental but vital task due to the reasons: (i) it is widely used under common scenarios where there is only a single source active in the environment; and

(ii) many multiple source localization algorithms are accomplished by simplifying them into repetitive single source localization problems [24,25]. Current single source localization methods generally achieve satisfying localization accuracy in typical environments, but at the cost of significant computational complexity. It is because they require a two-dimensional grid searching over all possible directions over the space. This inherent drawback limits the practical use where a fast response time is required (e.g., sound source tracking). It is of great interest to develop two-dimensional DOA estimators in the spherical harmonics domain, which has a reduced complexity while achieving competitive localization accuracy.

- *Multiple source localization in complex environments*: it is still of great challenge to achieve accurate localization of multiple sources in complex acoustic environments: (i) acoustic scenes characterized by strong noise: the environmental, thermal, and other forms of interfering noise; (ii) strong reverberant environments: the original recordings are contaminated by the multi-path reverberation resulting from strong reflections from objects in the enclosure/room. Both issues are common environmental factors that hinder an accurate acquisition of the original recordings, which in return cause inaccurate estimations of the desired source feature. As a consequence, the accuracy of multiple source localization becomes severally degraded due to the interfering noise and reverberation. Another difficulty, suffered by most of the multi-source localization approaches, is the overlapping component due to the simultaneous sources. It is still of significant value to provide a promising solution to this problem in complex environments, which can simplify the challenging multi-source localization problems into easier single source localization problems.

- *Practical factors that degrade the localization performance*: most of the existing source localization algorithms often make some assumptions that are hardly true when implemented with commercial hardware. One typical example is the assumption of an omnidirectional behavior for commercial loudspeakers. This assumption holds for theoretical acoustic models but hardly conforms to reality because a real-life loudspeaker has a unique directivity

pattern. To the author's best knowledge, very few published research considers such issues when proposing source localization methods. This thesis takes such problems into an investigation by examining the real characteristics of real-life sources/loudspeakers to calibrate the theoretical acoustic models for improved localization performance.

To address the above issues, the key question which drives this thesis is as follows:

How can we achieve improved sound source localization performance in diverse acoustic environments by using a novel and appropriate source feature that is estimated from higher-order microphone array recordings?

1.2 Problems and Solutions

We elaborate the formulated problem raised by the thesis into three further questions and provide some intended solutions to each of the questions:

Problem (i): Is the spherical harmonics domain sound source feature, studied by this thesis, suitable for sound source localization in diverse environments?

Solutions: The newly spherical harmonics domain source feature investigated by this thesis is called *relative harmonic coefficients*, which has several remarkable properties desired for sound source localization: (1) its independence from the time-varying source signal and sole dependence on the source position even in a reverberant environment; (2) easy estimations in noisy environments from the higher-order microphone array recordings; and (3) a significant spatial resolution due to its unique directivity pattern over space, making it efficient to distinguish/localize the sound source(s) propagating from different directions. The unique properties of relative harmonic coefficients are also exploitable by an overlapped frame detector that preserves the single-source frames so that the challenging localization of multiple sources is simplified into single source localization issues.

Problem (ii): How to achieve improved sound source localization

performance in diverse acoustic environments using spherical harmonics domain approaches?

Solutions: This thesis intends to develop several novel spherical harmonics domain approaches to address sound source localization challenges under different acoustic scenarios, respectively: (1) a single source localization and tracking algorithm in a typical noisy environment using a decoupled localization method, which is highlighted by a dramatically reduced computational complexity while achieving sufficient localization accuracy; (2) a semi-supervised algorithm to localize multiple overlapped sources in a severely noisy and reverberant environment, which only requires a limited number of labeled training samples and performs with improved localization accuracy compared with the state-of-art methods; and (3) an unsupervised subspace method to localize multiple simultaneous sources, which is more suitable in noisy environments than the traditional subspace methods. The performance of the localization algorithms developed by this thesis will be validated using extensive experiments in both simulated and real-life environments.

Problem (iii): How to evaluate/examine the characteristics of real-life sound sources, such as the directivity pattern of the commercial loudspeakers?

Solutions: This thesis addresses this issue by presenting a compact framework to evaluate/examine the real directivity pattern of commercial loudspeakers. We intend to use several equivalent theoretical acoustic models (i.e., plane waves, point sources, and mixed sources) to see which one performs with the most similar characteristics compared to the commercial loudspeakers. We exploit the spatial sparsity of the loudspeaker to propose several solutions addressing this formated problem. Besides, we provide two different metrics to evaluate the performance of the developed algorithms.

1.3 Thesis Outline

Motivated by the above problems, this thesis aims at sound source localization and source modeling using spherical harmonics domain approaches. Figure 1.4 presents the thesis outline.

Chapter 2: Literature Review and Background Theory:

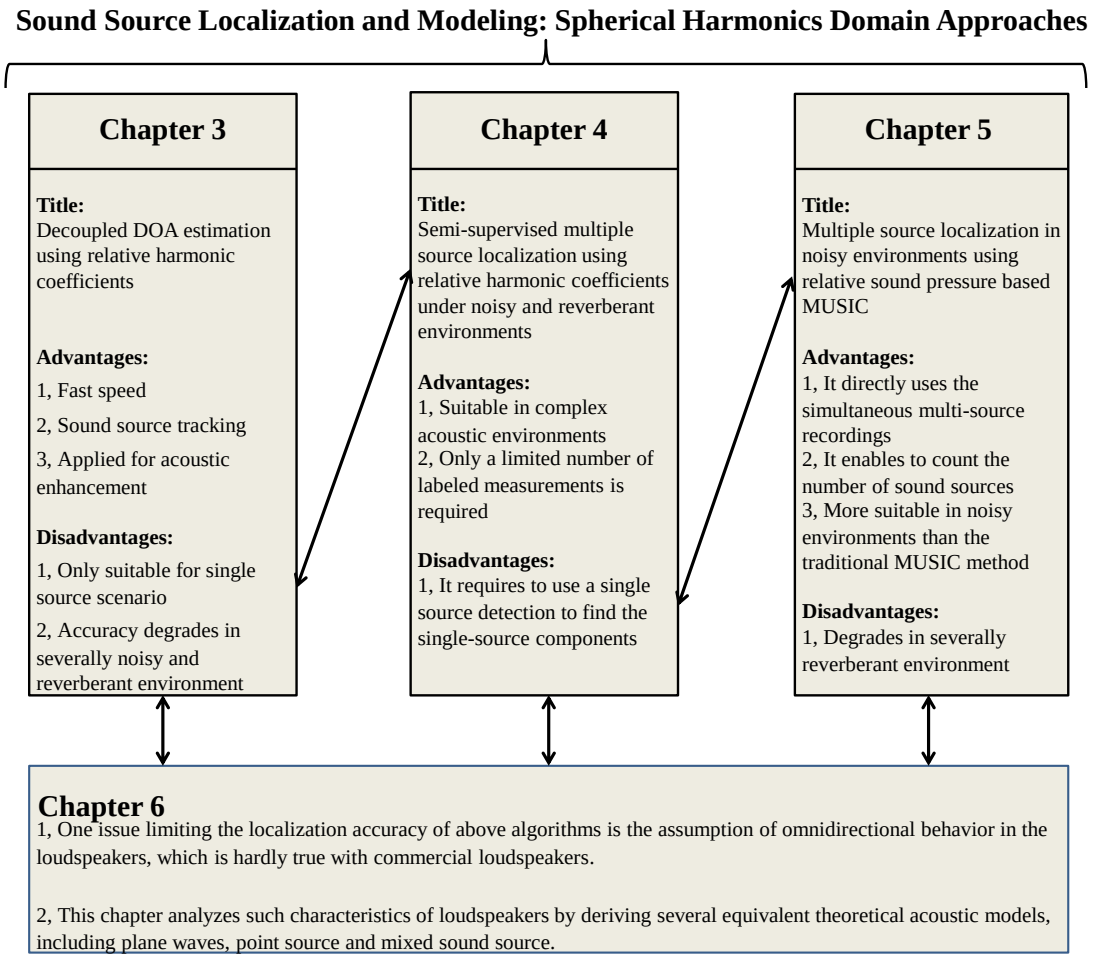


Figure 1.4: Thesis outline.

This chapter first presents an extensive literature review of past and present sound source localization methods. After that, we introduce the background knowledge about the decomposition of a soundfield measured using a higher-order microphone array into the spherical harmonics domain. Finally, we review a feature called relative transfer function (ReTF), which motivates a novel spherical harmonics domain source feature to be studied by this thesis. Overall, the preliminary theory and literature review in this chapter lay the foundation for the novel source localization algorithms developed in the following chapters.

Chapter 3: Decoupled Direction-of-Arrival Estimation Using Rela-

tive Harmonic Coefficients:

This chapter first presents a novel feature called *relative harmonic coefficients* by decomposing the higher-order microphone array measurements into the spherical harmonics domain. We then introduce the means to estimate this source feature in noisy environments. The properties of this feature enable us to develop two decoupled single source DOA estimators. They are capable of recovering the source's elevation and azimuth information separately. Compared to traditional single source localization methods, the proposed algorithms in this chapter significantly reduces the computational complexity. Hence, they are more applicable for computationally complex applications like sound source tracking. Additionally, we show that this developed source feature has a secondary usage in enhancing raw higher-order microphone recordings corrupted by noise.

Chapter 4: Semi-Supervised Multiple Source Localization Using Relative Harmonic Coefficients Under Noisy and Reverberant Environments:

The developed approach in the above chapter addresses single source localization, which is not suitable under multi-source scenarios. This chapter uses the relative harmonic coefficients to propose a semi-supervised algorithm to address the challenging multi-source localization problem in reverberant environments. A full investigation of this source feature in reverberant environments is presented, including (i) an illustration confirming its sole dependence on the source position in reverberant environments; (ii) a feature selector exploiting its inherent directivity over space. Source features at varied spherical harmonic modes, representing unique characterization of the spatial soundfield, are merged/fused by a unifying model. Based on the model, we then formulate a mapping function revealing the underlying relationship between the source feature(s) and position(s) using a Bayesian inference approach. The mixed measurements due to the multiple sources inevitably comprise overlapped components, which hinder accurate localization of the sound sources. To address this issue, we propose a pre-processing technique to detect the overlapped frames, reducing this challenging multi-source localization problem to a single source localization issue. Unlike most data-driven localization methods, this proposed method is highlighted with a strong potential to be implemented in practice as it only requires a limited number of labeled measurements.

Chapter 5: Multiple Source Localization in Noisy Environments Using Relative Sound Pressure Based MUSIC:

Although the approach developed in the above chapter addresses the multi-source localization, it requires a pre-processing tool to detect the single-source components. Hence, the method above does not suit the cases where the original recordings have a limited number of single-source frames/bins. By contrast, this chapter provides another solution to address the multi-source localization problem by directly using the overlapped multi-source recordings. The proposed method implements a MUSIC algorithm framework while being more suitable in noisy environments than the traditional MUSIC approach. After that, we decompose the proposed MUSIC approach into the spherical harmonics domain, where a frequency smoothing technique is allowed to de-correlate the coherent source signals for improved localization accuracy. The proposed algorithm enables us to estimate the number of active sound sources, which is pre-requisite knowledge for the traditional MUSIC approach.

Chapter 6: Modeling Characteristics of Real Loudspeakers Using Various Acoustic Models:

The performance of the localization algorithms above is affected by some practical factors, such as the inherent directivity pattern of the commercial loudspeakers. This chapter analyzes such characteristics of loudspeakers by deriving equivalent theoretical models, including plane waves decomposition, point source decomposition, and mixed source decomposition. Each proposed model employs three sparse decomposition algorithms for optimized solutions, including iteratively reweighted least squares (IRLS), matching pursuit (MP), and least absolute shrinkage and selection operator (LASSO). A successful model shall enable the prediction of the soundfield outside the original recording region. Therefore, we validate the effectiveness of the models by comparing the simulated soundfield with secondary measurements obtained beyond the recording area. Experimental results have confirmed that both the plane wave and mixed source models achieve promising performance.

Chapter 7: Summary:

Chapter 7 presents concluding remarks and contributions of the work in this thesis and some promising research directions for future work.

Chapter 2

Literature Review and Background Theory

Overview: *The primary goal of this thesis is to address sound source localization challenges using a novel source feature in the spherical harmonics domain. This chapter first presents a comprehensive literature review of sound source localization, where both the single and multiple source localization techniques are covered. After that, we give a detailed introduction of the background knowledge on decomposing a soundfield measured using a higher-order microphone array into the spherical harmonics domain. By the end, we introduce an existing feature called relative transfer function (ReTF), which has been widely used by current sound source localization algorithms. The wide applications of ReTF motivate us to develop the novel spherical harmonics domain source feature in the next chapter.*

2.1 Literature Review: Sound Source Localization

Sound source localization algorithms in the literature can be broadly divided into single and multiple source localization, depending on the number of sound sources present in the environment. Single source localization refers to the scenario where only a single sound source is active in the environment. By contrast, multi-source localization addresses the scenarios where multiple sources are active.

2.1.1 Single Source Localization

Early single source localization approaches exploit the time difference of arrival (TDOA) between microphone pairs [26]. The time delay estimation effectively maximizes the 'synchrony' between time-shifted microphone outputs to identify the source position. These methods are dual-step approaches [27] because they require two stages to accomplish the source position estimation. The first stage mainly estimates the TDOA using the measurements of the pair(s) of microphones. For example, the TDOA was estimated by identifying the peaks in the cross-correlation between the microphones, such as the generalized cross-correlation phase transform (GCC-PHAT) [28]. However, the GCC-PHAT method is inaccurate in reverberant conditions because it assumes free-field wave propagation [28]. Researchers in [29] overcame this issue by proposing an alternative TDOA estimator by reformulating the original problem as a linear regression problem. After that, the work in [26] developed another TDOA extraction method, considering not only room reverberations but also spatially correlated noise. The estimated TDOA from the first stage is then applied to sound source localization in the second stage. Early application in [30] localized the sound source using the spatial-temporal information via three localization schemes, i.e., a recursive form of the Gauss method, the extended Kalman filter, and the unscented Kalman filter. A recent application in [31] combined the traditional TDOA based tracking schemes with a learning-based approach to estimate the trajectory of a single speaker in noisy and reverberant environments.

The other popular type of approaches to address single source localization prob-

lem is beamforming based methods. A typical example is the steered response power (SRP) based method [32], which explores all possible directions over the two-dimensional directional space to search for areas with higher response power. The corresponding improved method called steered response power phase transform (SRP-PHAT) enhanced the SRP method's localization robustness to the noisy and reverberant environments [33]. However, the SRP-PHAT method depends on a costly spatial grid-search to find a global maximum, making the computational cost a heavy burden. Some papers intended to reduce the computational cost to make it more practical. For example, the method in [34] used a stochastic region contraction to reduce the computational complexity; and the other approach in [35] performed a full exploration of the sampled space rather than the continuous space.

Another type of source localization, mainly used under single-source scenarios, is binaural source localization. This technique estimates the source location concerning the human ears. Most binaural source localization approaches use the Head-Related Transfer Function (HRTF), which can be viewed as a set of acoustic filters from the sound source to a listener's eardrums [36]. The HRTFs contain all the listening cues and spatial information of the sound source to be localized, such as interaural level differences (ILDs), interaural phase differences (IPDs), and interaural time differences (ITDs) [37, 38, 39]. However, those source features are only valid at lower frequencies and are insufficient to localize the source elevation over the three dimensions due to the 'cone-of-confusion' [40]. One recent research in [41] overcame this limitation by introducing a new feature vector combining the interaural phase and magnitude features present in the HRTF. After that, the work in [42] adopted the combined feature vector to a learning-based approach for robust binaural source localization in complex environments. However, the HRTF based localization methods suffer a major drawback, i.e., they can be hardly generalized to different speakers because the HRTFs strongly depend on the listener's anatomy (every listener's anatomy is unique). Currently, there are two solutions to this problem: (i) measuring the HRTFs of different people in an anechoic chamber to study the transformation characteristics of the external ear and to synthesize virtual reality over headphones [43]; and (ii) modeling the HRTF using a small number of parameters estimated from a limited set of practical measurements [44].

Most of the aforementioned single source localization methods implement in

Table 2.1: Summary of single source localization approaches using different types of source features.

Approach	Input feature	Sources	Unsupervised
[26]	Time difference of arrival	Single	Yes
[28]	Generalized cross-correlation	Single	Yes
[32]	Steered response power	Single	Yes
[38]	Interaural time differences	Single	Yes
[42]	Interaural phase and magnitude	Single	No
[45]	Relative transfer function	Single	No
[46]	Model-based features	Single	No
[47]	Power-ratios	Single	No

an unsupervised schema where no prior information and recordings are required. However, their localization performance degrades severely in a complex acoustic environment with the interference of the multi-path acoustic reverberations and the noise with low signal-to-noise ratios. In the recent decade, many machine learning-based (also called 'data-driven') approaches performed with improved localization performance in those unfavorable environments. These techniques typically learn the patterns between a pre-defined feature and the source position from a training dataset measured in advance, and then utilize these learned patterns to predict the source location for the unknown testing sources [48]. Next, we review some recently proposed learning-based methods addressing the single source localization problems. Single-source DOA estimations in [49,50,51] were realized using a multi-layer neural network, which used different variations of generalized cross-correlation (GCC) as the algorithm inputs. In a special issue on "Acoustic Source Localization and Tracking in Dynamic Real-Life Scenes" in the IEEE Journal on Selected Topics in Signal Processing, some variants of neural networks for data-driven source localization were introduced [10]. Up to the present, deep learning-based localization approaches have adopted several different architectures of neural networks, such as convolutional neural networks (CNNs) [52,53], deep residual networks (DNN) [54], and convolutional and recurrent networks (CRNNs) [55]. And, different source features have been used as the inputs of those learning-based methods, such as

binaural features [42], the eigenvectors of the spatial covariance matrix [56], and raw short-time Fourier transform (STFT) of signals [52, 53]. For clarity, we summarize some source features frequently used by current single source localization techniques (i.e., see Table 2.1). The above deep learning-based methods address the source localization problem by transferring them into a classification problem. As a result, they suit more to localize the discrete source positions (e.g., source DOAs). By contrast, the regression schemes are more favorable when localizing continuous variables of the source positions. For example, [45] adopted and fused a Bayesian inference approach of Gaussian Process Regression (GPR) over multiple nodes to estimate the Cartesian coordinates of the sound source. Overall, data-driven source localization is often criticized as a cumbersome task because it requires a large training set of labeled measurements. To overcome this drawback, the work in [45,57,58] adopted the semi-supervised paradigm, which required only a small number of labeled samples with known positions. Besides, they also need a large set of unlabeled samples, whose corresponding source locations were unknown. More recent research in [59] presented a weakly-labeled learning-based localization paradigm, using a significantly reduced number of labeled samples.

2.1.2 Multiple Source Localization

Localization of multiple sources in the environment is inherently more challenging than single source localization. Next, we review some multiple source localization methods under different acoustic environments as well as some approaches in the spherical harmonics domain.

Environments with low noise and low reverberation

We first review the approaches which are suitable in environments that have light reverberations and low signal-to-noise ratios.

(i) *Single single-source component assumption/detection based methods*: These types of multi-source localization techniques [22,25,60] rely on the hypothesis that, given the simultaneous multi-source recordings, the speech components have a sparse distribution over the STFT domain. Hence, only a single source is active or dominant over the others at an arbitrary time-frequency bin. This property is also

referred to as the W-disjoint orthogonality (WDO) assumption [61]. The assumption of WDO simplifies the challenging multiple source localization problems into repetitive single source localization problems. However, such an assumption hardly conforms to all STFT bins as some bins contain overlapped multiple sources. By contrast, some recently proposed algorithms, such as [24, 62], do not require the sound sources to strictly follow the WDO, as they use a pre-processing technique to detect the areas whose contributions are only from one significant sound source. For example, Nadiri *et al.* adopted the coherence test, initially developed in [63], to identify the single-source components in the first stage. After that, the algorithm in [24] accomplished the multi-source localization by implementing repetitive single source localization using the detected single-source components. Since each single-source component has a corresponding source position, all estimated single source positions are then collected and clustered using clustering algorithms, such as the K-means algorithm [64]. They assign each estimated source position to the closest subset and update each cluster center as the average of all estimates in each cluster.

(ii) *Blind source separation based methods*: Instead of detecting the single-source components, the other type of multi-source localization algorithms uses the separated source signal by blind source separation (BSS) algorithms. They divide the mixed multiple source recordings into individual single source recordings, using no prior information about the sources. When multi-source signals are accurately separated, the identified de-mixing system contains the relative transfer functions from each source to each microphone, which transfers the challenging multi-source localization problem into repeated single source localization problems. Hence, the BSS based methods also belong to dual-step approaches, i.e., separating the mixed measurements in the first stage and then implementing single source localization using the separated signals in the second stage. Researchers in [65] first used independent component analysis (ICA) to separate the mixed signal in the frequency domain, addressing both near-field and far-field DOA estimations in the second stage. However, there exists a limitation that the number of multiple sources cannot exceed the number of microphones on the array. To relax this limitation, the work in [66] exploited the sparsity and statistical models of speech signals in the STFT domain to develop an improved blind source separation method, allowing for

more sound sources than microphones in the array. However, similar to the WDO property, the sparsity assumption hardly conforms to the scenarios where there are many sources, especially in reverberant rooms [67]. In the recent decade, many source separation algorithms have been developed in the literature for improved performance in a more reverberant environment [5, 68]. A promising solution, used by lots of recent algorithms, is to use the binary masking for associating sources in the STFT domain to achieve accurate source separation [69, 70].

(iii) *Subspace methods*: The above-mentioned multi-source localization techniques require the availability of single-source components. By contrast, subspace methods use simultaneous recordings directly [18, 21, 71]. The most popular subspace method shall be the multiple signal classification (MUSIC) [18] and its improved versions, attracting great attention due to their easy implementation with satisfying localization performance [72, 73, 74, 75, 76, 77, 78]. The fundamental theory is to decompose the recording's spatial covariance matrix to compute the noise subspace, which is orthogonal to the steering plane wave vectors towards the source DOA. In practice, they utilize a singular value decomposition (SVD) of the spatial covariance matrix and then search the significant peaks over the pseudo-spectrum in the two-dimensional directional space. Another well-known subspace technique is the high-resolution estimation of signal parameters via rotational invariance technique (ESPRIT) that uses parametric methods such as least-squares estimation [19]. One key advantage of the ESPRIT method, in comparison with the MUSIC method, is that it computes the source DOAs using a closed-form expression, not requiring a costly search over the two-dimensional directional space. Most of the existing subspace methods mainly assume far-field sound sources. By contrast, a recent publication in [79] adjusted the subspace method to localize and track multiple near-field sources, where both the source DOAs and ranges are localized. Finally, we point out that the subspace methods are vulnerable to interference from room reverberations and noise, which severely degrade their localization accuracy.

Environments with strong noise and high reverberation

Localization of multiple sources in noisy and reverberant environments is a much more challenging task. Researchers in the community have proposed a vast number

of algorithms to address this challenging task. For clarity, this thesis reviews the following three types of approaches in the literature:

(i) *Learning-based approaches*: Data-driven algorithms have become a promising solution to address multiple source localization in noisy and reverberant environments. Overall, most learning-based multi-source localization methods also belong to dual-step approaches. The first stage mainly addresses the estimations of the source features from the single-source areas. In the second stage, the estimated features are fed into the learning algorithms to recover the source positions based on the learning relation/model. Based on the WDO assumption, the approach in [22] adopted a CRNN to estimate the DOAs of multiple sound sources using a first-order Ambisonics source feature. Chakrabarty *et al.* used the phase component of the STFT coefficients of the received microphone signals as the input feature to a CNN for supervised multi-speaker DOA estimations [80]. Those deep learning-based algorithms accomplish the goal by classifying the desired source DOAs into some of the candidates' directions over the 2-D space. Although localization accuracy is improved, most learning-based solutions require many labeled training samples in advance. The other limitation is the tedious requirement to re-measure the training samples and re-learn the models in different reverberant enclosures.

(ii) *Reverberations modeling based approaches*: Unlike traditional conventions to neglect the reverberations by relying on some assumptions such as the WDO property, the other intuitive solution is to exploit the reverberations for source localization in reverberant environments [81]. For example, Ribeiro *et al.* used an approximate model of the reverberate enclosure to extract related information from early acoustic reflections [82]. However, this method has not yet made full use of the room reverberations. By contrast, research in [83,84] adopted a full acoustic reverberation model using the reverberant room impulse response [85]. However, the methods failed to address multiple source localization because it cannot provide related clues to localize the sound sources propagating from different locations. More recently, Birnie *et al.* in [86, 87] used a complete model of environmental reverberation characteristics to develop an improved version of SHD-MUSIC. This approach was more suitable for reverberant environments compared to the traditional SHD-MUSIC method. However, this method requires measured or simulated room coupling coefficients between the recording and source regions in the rever-

berant room [88, 89]. Similar to the learning-based approaches, the reverberation based localization approaches also suffer a weak generalization into a different reverberant enclosure. It is because the characteristics of room reverberations can be changed significantly in a new environment so that the original reverberant model becomes invalid.

(iii) *Direct-path recordings based approaches*: Different from the room-dependent methods above, there are some localization methods in the literature that are more independent from the reverberant rooms. They detect the direct-path recordings within the mixed reverberant measurements, which are only dominant by the components due to a direct speaker/source. Recently, Li *et al.* used microphone pairs to detect the direct-path ReTF to address multi-source localization in noisy reverberant environments [90, 91]. To the authors' best knowledge, several other direct-path detectors were available. For example, a coherence test identifies the direct-path components by checking whether the correlation matrix has a unit rank or not [24, 62, 92]. Additional research in [93, 94] measured the cross-correlation coefficients between the reverberant observations. The key advantages of the direct-path based localization methods are: (i) they are more independent of the reverberant rooms, thus can be applied to different reverberant enclosures; and (ii) they implement in an unsupervised scheme where no prior information is required. Both advantages enhance the practicality of the algorithms under real-life scenarios. One potential limitation is that their performance severely relies on the accuracy of the direct-path recording detection, where some user-defined thresholds may be required for the detection.

Spherical harmonics domain approaches

In the recent decade, higher-order microphone arrays are now widely applied to sound source localization [12,13,14]. The multi-channel measurements using higher-order microphone array recordings can be decomposed into the spherical harmonics domain (SHD) using a set of orthogonal spatial basis functions [16]. One of the main advantages of SHD techniques is the decoupling of frequency-dependent and angular-dependent components. Beamforming approaches, such as the SRP based localization method, can be implemented in the spherical harmonics domain (i.e.,

Table 2.2: Summary of multiple source localization approaches using different types of source features. Note the column named as SHD shows whether the method performs in the spherical harmonics domain or not.

Approach	Input feature	Sources	SHD	Unsupervised
[18]	Spatial correlation matrix	Multiple	No	Yes
[22]	First-order Ambisonics	Multiple	Yes	No
[23]	Modal coherence patterns	Multiple	Yes	No
[55]	Magnitudes and phases	Multiple	No	No
[80]	STFT coefficients	Multiple	No	No
[95]	Time difference of arrival	Multiple	No	Yes
[96]	Steered response power	Multiple	Yes	No
[97]	Phase difference	Multiple	No	No
[25, 98]	Intensity/pseudointensity vectors	Multiple	Yes	Yes

SHD-SRP) [99]. Another example of the beamformer based localization approach is the spherical harmonics domain minimum variance distortionless response (SHD-MVDR) to localize near-field sources [100].

Subspace methods mentioned above have mostly been re-defined in the SHD framework. For example, early research in [101] first proposed the spherical harmonics domain MUSIC (SHD-MUSIC). After that, the work in [17] improved the localization accuracy of SHD-MUSIC using frequency smoothing to de-correlate the coherent source signal. The MUSIC method mentioned above, which is more suitable to reverberate environments, was also implemented in the spherical harmonics domain [86, 87]. The ESPRIT method in the spherical harmonics domain is called EB-ESPRIT, which provides an elegant closed-form solution for three-dimensional source localization [20, 21]. The acoustic intensity, pseudointensity, and subspace pseudointensity based localization approaches were also implemented in the spherical harmonics domain [25, 98, 102]. Recently, data-driven localization techniques in the spherical harmonics domain were also available. For example, Fahim *et al.* achieved multi-source DOA using a CNN by learning the modal coherence patterns of an incident soundfield through the measured spherical harmonic coefficients [23]. Another approach in [22] used CRNN to estimate the DOAs of multiple sound sources using a first-order Ambisonics source feature.

The multi-source localization approaches above use several different types of source features, such as generalized cross-correlation, first-order Ambisonics, STFT coefficients, and modal coherence patterns. Table 2.2 summarizes several different types of source features used by current multi-source localization techniques. As analyzed in Chapter 1.1, the source features are vital to the localization accuracy as they contain relevant characteristics of the sound source to be localized. A promising source feature suitable for source localization shall better convey/represent the source position, and be less dependent on the time-varying source signal. As introduced in subsection 2.3, the source feature of relative transfer function (ReTF) is proved to have such properties. Before introducing the ReTF, the next subsection presents a background introduction about the spherical harmonics representation of a measured soundfield using higher-order microphone arrays.

2.2 Background: Spherical Harmonics Representation of a Soundfield

Higher-order microphone arrays (e.g., spherical microphone array) have become popular in the field of array signal processing as they are capable of recording and analyzing the desired soundfield over a sizable spatial region. The spherical harmonics representation, based on a spherical Fourier transform, is a now commonly used tool by the higher-order microphone arrays. A set of orthonormal basis functions, called the spherical harmonics functions, are adopted to decompose the multi-channel recordings on the microphone array into the spherical harmonics domain, where the soundfield is represented by the spherical harmonic coefficients. Overall, the spherical harmonics decomposition naturally provides a more insightful/compact representation of the spatial sound field than that using a distributed set of target points on the microphone array. Up to present, spherical harmonics domain techniques have been widely applied to many spatial signal processing applications and techniques, including spatial sound field reproduction [103, 104, 105, 106, 107, 108, 109], active noise control [110, 111, 112, 113, 114, 115], sound source localization [20, 102, 116, 117, 118, 119], room acoustic modeling [120, 121, 88, 122], spatial filtering and beamforming [123, 124, 125, 126], echo cancellation [127, 128],

soundfield separation [129,130], power spectral densities estimation [4,131] and the design of higher-order loudspeakers [132,133].

The algorithms to be developed by this thesis are implemented in the spherical harmonics domain using higher-order microphone array recordings. Hence, this subsection introduces detailed background knowledge about the spherical harmonics representation of a soundfield using the higher-order microphone arrays. In the following, we first review the classical wave equation, characterizing the sound waves propagating over space, and discuss the wave equation solution using a finite and compact expansion of spherical harmonics functions. Then, we introduce the spherical harmonics representation of acoustic models, including both the plane waves and point sources propagating over free space. Finally, we show the method to measure the incoming soundfield using a spherical microphone array.

2.2.1 Coordinate System

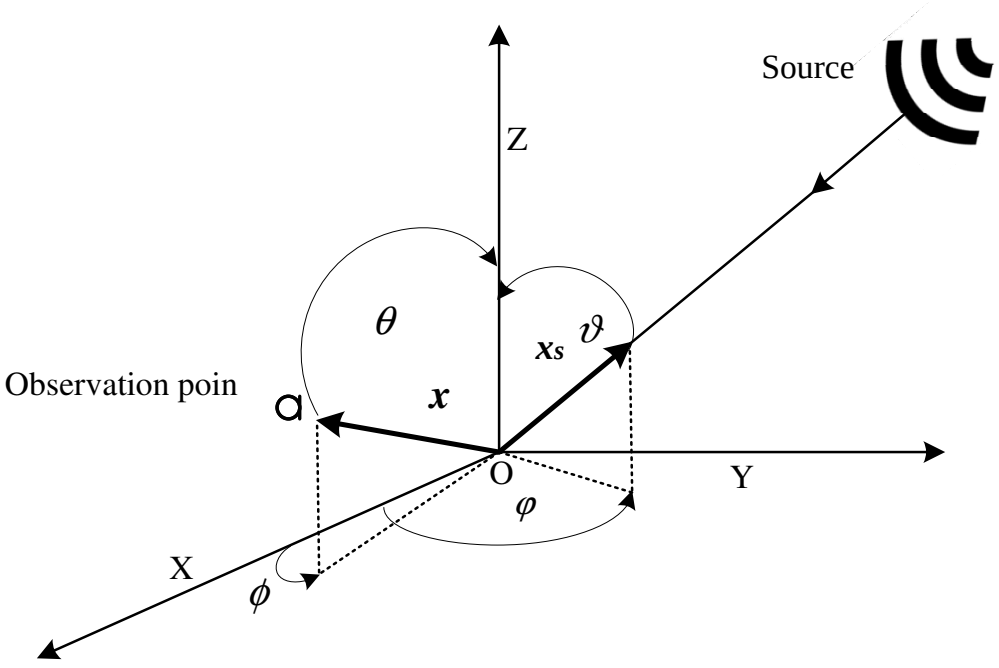


Figure 2.1: A typical soundfield.

Figure 2.1 presents a typical soundfield where a single sound source is active in

the three-dimensional space. The sound source, generating the waves over space, is located at $\boldsymbol{x}_s = (r_s, \theta_s, \phi_s)$ in the spherical polar coordinates. The coordinates of the observation point is $\boldsymbol{x} = (r, \theta, \phi)$ where $r = \|\boldsymbol{x}\|$ denotes the distance from the origin O with $\|\cdot\|$ implying the Euclidean distance, θ denotes the polar coordinate from the vertical axis with $0 \leq \theta \leq \pi$, and ϕ denotes the azimuthal coordinate in the horizontal plane containing the origin with $0 \leq \phi \leq 2\pi$. The microphone position in the right-handed Cartesian coordinates (x, y, z) is transformed using the spherical polar coordinates through,

$$x = r \sin\theta \cos\phi \tag{2.1a}$$

$$y = r \sin\theta \sin\phi \tag{2.1b}$$

$$z = r \cos\theta. \tag{2.1c}$$

Note that, Figure 2.1 only shows a single microphone for notational convenience. Actually, spatial acoustic processing techniques generally use a set of microphones, distributed over the microphone array, to measure the soundfield.

In the next subsection, we discuss the representation of the theoretical sound pressure at the arbitrary observation point $\boldsymbol{x}$.

2.2.2 Helmholtz Equation

This subsection introduces the wave equation derived based on a combination of the Euler equation, the continuity equation, and the state equation. The equation describes the sound pressure at any arbitrary point within a homogenous medium, i.e., the sound pressure $P(\boldsymbol{x}, t)$ at an arbitrary position $\boldsymbol{x}$ and time t is represented by,

$$\nabla^2 P(\boldsymbol{x}, t) = \frac{1}{c^2} \frac{\partial^2 P(\boldsymbol{x}, t)}{\partial t^2} \tag{2.2}$$

where c denotes the speed of sound and ∇^2 denotes the Laplacian operator. Note that ∇^2 has varied representations in different coordinate systems. For example, ∇^2 in Cartesian coordinates follows as,

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.3}$$

By contrast, ∇^2 in spherical coordinates is,

$$\nabla^2(\cdot) = \frac{1}{r^2}\frac{\partial}{\partial r}[r^2\frac{\partial}{\partial r}(\cdot)] + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}[\sin\theta\frac{\partial}{\partial\theta}(\cdot)] + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}(\cdot). \tag{2.4}$$

The time-domain wave equation in (2.2) is transformed into the frequency domain using a Fourier transform

$$\mathcal{F}\{\frac{\partial P(t)}{\partial t}\} = -i\omega P(\omega) \tag{2.5}$$

to become

$$\nabla^2 P(\boldsymbol{x},\omega) + k^2 P(\boldsymbol{x},\omega) = 0 \tag{2.6}$$

which is called the Helmholtz equation, where ω denotes the angular frequency and $k = \omega/c$ denotes the wavenumber.

2.2.3 General Solution

The solution of equation (2.6) is solvable using separation of variables,

$$P(\boldsymbol{x},\omega) = X(r,\omega)\Theta(\theta,\omega)\Phi(\phi,\omega) \tag{2.7}$$

which produces three ordinary differential equations [16] as,

$$\frac{d^2\Phi}{d\phi^2} + m^2\Phi = 0 \tag{2.8}$$

$$\frac{1}{\sin\theta}\frac{d}{d\theta}(\sin\theta\frac{d\Theta}{d\theta}) + [n(n+1) - \frac{m^2}{\sin^2\theta}]\Theta = 0 \tag{2.9}$$

$$\frac{1}{r^2}\frac{d}{dr}(r^2\frac{dX}{dr}) + k^2X - \frac{n(n+1)}{r^2}X = 0 \tag{2.10}$$

where both n and m are integers.
The solution to the differential equation in (2.8) is

$$\Phi(\phi) = \Phi_1 e^{im\Phi} + \Phi_2 e^{-im\phi} \tag{2.11}$$

where Φ_1 and Φ_2 denote arbitrary constants, and $i = \sqrt{-1}$ denotes the imaginary

number.

The solution to the differential equation in (2.9) is

$$\Theta(\theta) = \Theta_1 P_{nm}(\cos\theta) \tag{2.12}$$

where Θ_1 denotes an arbitrary constant and $P_{nm}(\cdot)$ denotes the associated Legendre function of the first kind. Note that the definition of $P_{nm}(\cdot)$ may differ in different books. This thesis uses the Ferrers' definition [134]

$$P_{nm}(t) = \frac{(1-t^2)^{m/2}}{2^n n!}\frac{d^{m+n}}{dt^{m+n}}(t^2-1)^n \tag{2.13}$$

which is only valid for $n = 0, 1, 2, \cdots$, $m = 0, 1, \cdots, n$, and zero for $m > n$ with $-1 \leq t \leq 1$.

The solution to the differential equation in (2.10) is

$$X(r) = X_1 j_n(kr) + X_2 y_n(kr) \tag{2.14}$$

where X_1 and X_2 denote two arbitrary constants, and $j_n(\cdot)$ and $y_n(\cdot)$ denote the spherical Bessel functions of the first and second kind, respectively. Equation (2.10) has an alternate solution below,

$$X(r) = X_3 h_n(kr) + X_4 h_n^{(2)}(kr) \tag{2.15}$$

where $h_n(\cdot)$ and $h_n^{(2)}(\cdot)$ denote the spherical Hankel functions of the first and second kinds, respectively.

For notational convenience, the angle functions, i.e., (2.11) and (2.12), are conveniently combined into a single compact function called spherical harmonics function [16],

$$Y_{nm}(\theta,\phi) = (-1)^m \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{im\phi} \tag{2.16}$$

which is orthonormal over the two-dimensional directional space, i.e.,

$$\int_{\mathbb{S}^2} Y_{nm}(\widehat{\boldsymbol{\phi}}) Y^*_{m'n'}(\widehat{\boldsymbol{\phi}}) ds(\widehat{\boldsymbol{\phi}}) = \delta_{mm'}\delta_{nn'} \tag{2.17}$$

where $[\cdot]^*$ denotes the complex conjugate operator and $\delta_{nn'}$ denotes the Kronecker delta function

$$\delta_{nn'} = \begin{cases} 1, & n = n' \\ 0, & n \neq n'. \end{cases} \tag{2.18}$$

Figure 2.2 presents the spherical harmonics function $Y_{nm}(\cdot)$ in (2.16) up to the second order. Given an arbitrary function, it has a unique directivity pattern over the three-dimensional space.

Based on (2.14) and (2.16), both $j_n(kr)Y_{nm}(\theta,\phi)$ and $y_n(kr)Y_{nm}(\theta,\phi)$ are the solutions to (2.6). Therefore, we write a general solution to the Helmholtz equation as follows,

$$P(\boldsymbol{x},k) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}[\alpha_{nm}(k)j_n(kr) + \widehat{\alpha}_{nm}(k)y_n(kr)]Y_{nm}(\theta,\phi) \tag{2.19}$$

where $\alpha_{nm}(k)$ and $\widehat{\alpha}_{nm}(k)$ denote the spherical harmonic coefficients, which contain all the information of the measured soundfield in the spherical harmonis domain. Similarly, both $h_n(kr)Y_{nm}(\theta,\phi)$ and $h_n^{(2)}(kr)Y_{nm}(\theta,\phi)$ denote the solutions to (2.6). Hence, an alternate solution to the Helmholtz equation is

$$P(\boldsymbol{x},k) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}[\beta_{nm}(k)h_n(kr) + \widehat{\beta}_{nm}(k)h_n^{(2)}(kr)]Y_{nm}(\theta,\phi) \tag{2.20}$$

where β_{nm} and $\widehat{\beta}_{nm}$ are their spherical harmonic coefficients.

Note that the solution to the given soundfield also depends on whether it is an interior or exterior soundfield. An incoming soundfield due to the sound sources located outside its outer boundary is called an *interior field* (e.g., see Fig 2.3). For an interior soundfield, the sound pressure at the origin O should have a finite value. Hence, the solution in (2.20) is unusable because the output of the spherical Hankel function is infinity at the origin O. By contrast, the solution in (2.19) is applicable because the values of $j_{nm}(\cdot)$ are finite over the whole recording area. Since the incoming soundfield is more common in practice, this thesis mainly uses the solution in equation (2.19), so that the case with the solution in (2.20) is not used. However, the $y_{nm}(\cdot)$ in (2.19) is also infinite at the origin O. To address the

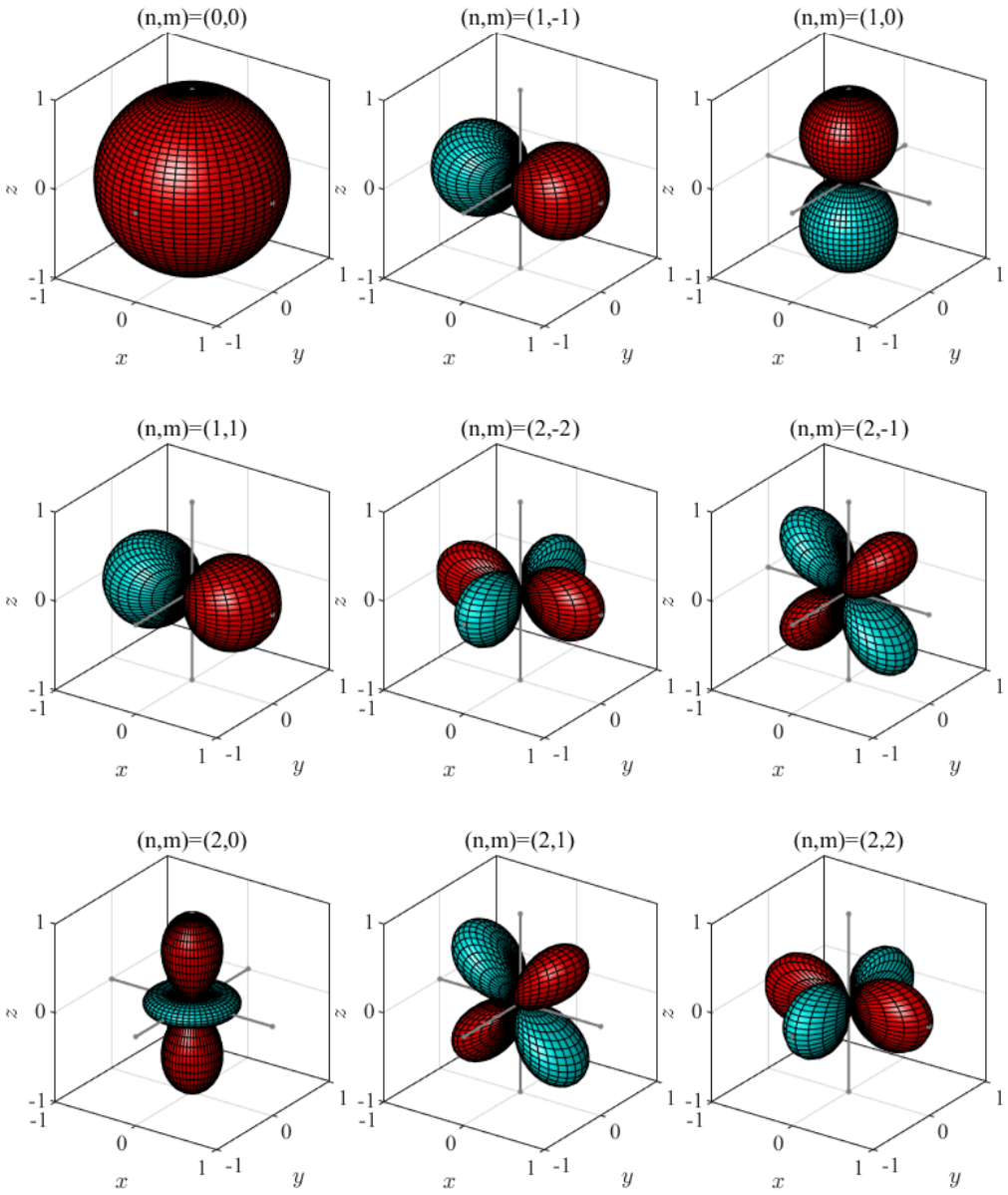


Figure 2.2: Two-dimensional spherical harmonics functions $Y_{nm}(\cdot)$ up to the second soundfield order. The (n, m) denotes the index of the functions. Red and cyan portions denote regions where the values are positive and negative, respectively. The distance of the surface from the origin indicates the absolute value of the features over that direction.

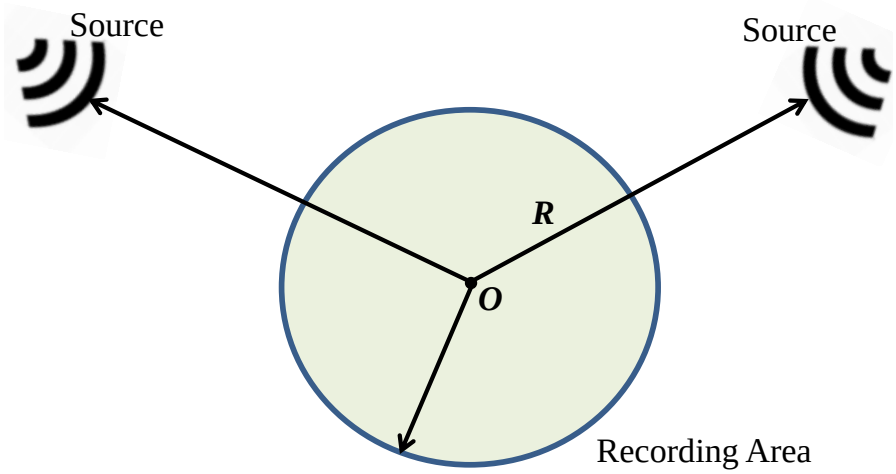


Figure 2.3: An interior field where all the sources are located outside of the recording area's outer boundary with radius R.

problem, we set $\widehat{\alpha}_{nm}(k) = 0$ for all n so that the general solution for a homogenous interior soundfield follows as,

$$P(\boldsymbol{x}, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \alpha_{nm}(k) j_n(kr) Y_{nm}(\theta, \phi). \tag{2.21}$$

The spherical harmonic representation in (2.21) is a fundamental tool in many spatial acoustic signal processing applications and techniques because of its efficient parameterization of any continuous source-free soundfield. Note that the spherical harmonic coefficients $\alpha_{nm}(\cdot)$ in (2.21) are independent of the individual observation points over the recording area. Hence, if the spherical harmonic coefficients of a soundfield are measured, they can be used to characterize/describe the entire continuous soundfield over the sizeable microphone array. Theoretically, the spherical harmonic representation in (2.21) requires an infinite number of the soundfield orders for an error-free decomposition. The Bessel function $j_n(\cdot)$ in (2.21) has a high pass behavior, whose output remains close to zero at higher soundfield orders. Hence, the soundfield in (2.21) can be truncated using a finite number of soundfield orders at the cost of insignificant errors. Researchers in [103] provided a practical

setting of soundfield order by $N = \lceil kr \rceil$, resulting in truncation error around 4% for a far-field soundfield (e.g., see Fig. 2.4 (a) and (b)). In this case, the soundfield decomposition in (2.21) can be rewritten as,

$$P(\boldsymbol{x}, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k) j_n(kr) Y_{nm}(\theta, \phi). \tag{2.22}$$

2.2.4 Point Source Decomposition

The inhomogeneous Helmholtz equation due to a point source located at the position $\boldsymbol{x}_s$ is defined by,

$$\nabla^2 g(\boldsymbol{x}|\boldsymbol{x}_s, k) + k^2 g(\boldsymbol{x}|\boldsymbol{x}_s, k) = -\delta(\boldsymbol{x} - \boldsymbol{x}_s) \tag{2.23}$$

where $\boldsymbol{x}$ denotes the observation position and $g(\boldsymbol{x}|\boldsymbol{x}_s, k)$ denotes the Green's function. Equation (2.23) implies that the homogeneous Helmholtz equation is satisfied everywhere except at the position $\boldsymbol{x} = \boldsymbol{x}_s$. Additionally, the solution to (2.23) requires to satisfy the Sommerfeld radiation condition [135]. A solution to (2.23), satisfying the Sommerfeld radiation condition, is given in [15],

$$g(\boldsymbol{x}|\boldsymbol{x}_s, k) = \frac{e^{ik\|\boldsymbol{x}-\boldsymbol{x}_s\|}}{4\pi\|\boldsymbol{x} - \boldsymbol{x}_s\|} \tag{2.24}$$

which can be decomposed into the spherical harmonics domain as [136, 137]

$$\frac{e^{ik\|\boldsymbol{x}-\boldsymbol{x}_s\|}}{4\pi\|\boldsymbol{x} - \boldsymbol{x}_s\|} = ik \sum_{n=0}^{N} \sum_{m=-n}^{n} h_n(kr_s) Y_{nm}^*(\theta_s, \phi_s) j_n(kr) Y_{nm}(\theta, \phi), \quad r_s > r \tag{2.25}$$

which suits for an interior soundfield where the sound source locates beyond the area boundary. In this case, the corresponding spherical harmonic coefficients follow as,

$$\alpha_{nm}(k) = ikh_n(kr_s) Y_{nm}^*(\theta_s, \phi_s). \tag{2.26}$$

On the contrary, when the source locates within the area of interest, i.e, an exterior soundfield, the sound pressure decomposed into the spherical harmonics domain is

represented as,

$$\frac{e^{ik\|\boldsymbol{x}-\boldsymbol{x}_s\|}}{4\pi\|\boldsymbol{x}-\boldsymbol{x}_s\|} = ik\sum_{n=0}^{N}\sum_{m=-n}^{n} j_n(kr_s)Y_{nm}^*(\theta_s,\phi_s)h_n(kr)Y_{nm}(\theta,\phi), \quad r_s < r \tag{2.27}$$

whose spherical harmonic coefficients are,

$$\beta_{nm}(k) = ikj_n(kr_s)Y_{nm}^*(\theta_s,\phi_s). \tag{2.28}$$

Next, we set a point source at infinity distance ($r_s \to \infty$) [137] (also called plane wave). Hence, the sound pressure due to a plane wave is computed by substituting $r_s = \infty$ and performing the far-field transition to (2.24),

$$P(\boldsymbol{x}|\boldsymbol{x}_s, k) = e^{i\boldsymbol{k}^T\cdot\boldsymbol{x}} \tag{2.29}$$

where $[\cdot]^T$ denotes the matrix transpose operator. The spherical harmonics decomposition of the sound pressure in (2.29) follows as,

$$e^{i\boldsymbol{k}^T\cdot\boldsymbol{x}} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi i^n Y_{nm}^*(\theta_s,\phi_s)j_n(kr)Y_{nm}(\theta,\phi). \tag{2.30}$$

which satisfies the general solution for an interior soundfield (2.21) with the spherical harmonics coefficients below,

$$\alpha_{nm}(k) = 4\pi i^n Y_{nm}^*(\theta_s,\phi_s). \tag{2.31}$$

Figure 2.4 exhibits the examples of the soundfield due to a plane wave and point source, respectively. Subfigure (a) and (c) denote the original soundfield with an infinite soundfield order N. By contrast, subfigure (b) and (d) denote the soundfield with a truncated order at $N = \lceil kr \rceil$. As mentioned in (2.22), the truncated error is acceptable at around 4% [103].

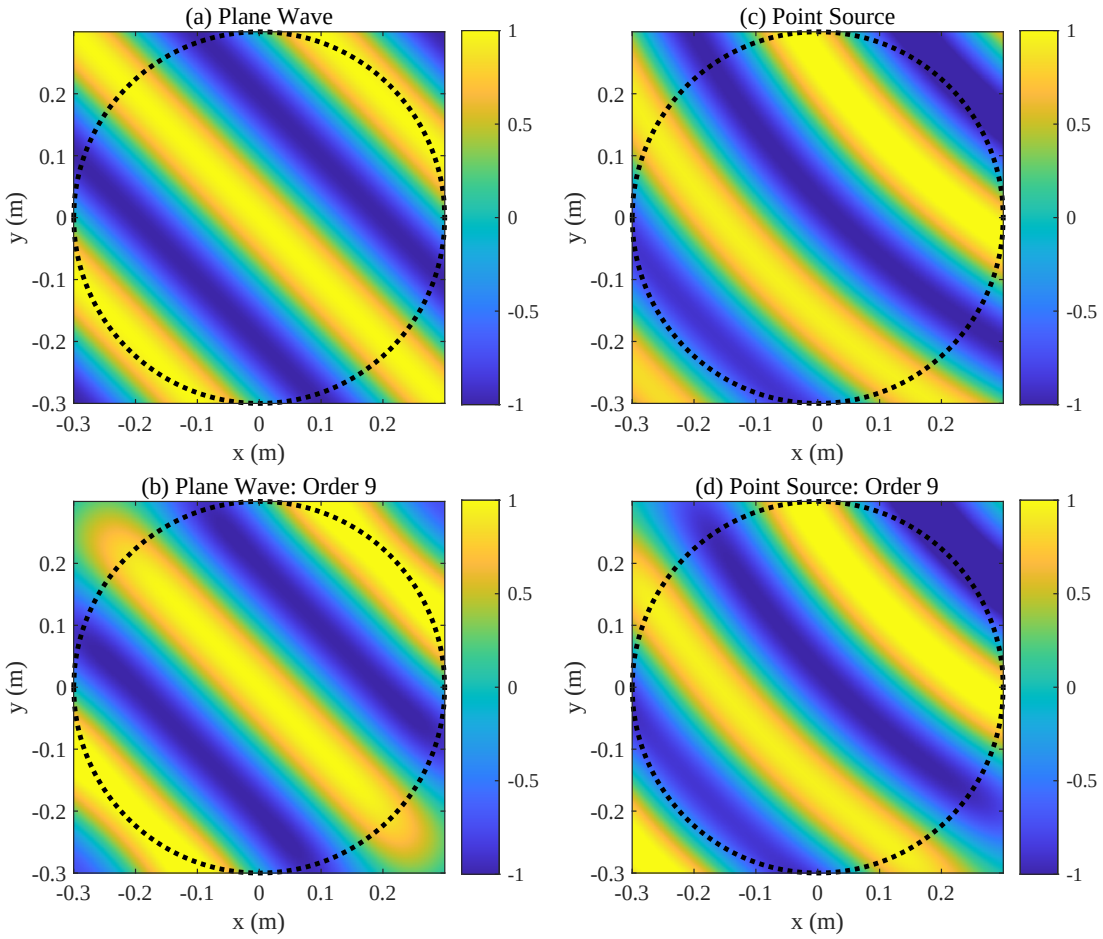


Figure 2.4: Four spatial soundfield representations at frequency 1500 Hz, including original plane wave (a), truncated plane wave (b), point source (c), and truncated point source (d). Note that the truncated soundfield order is 9 as $N = \lceil kr \rceil$. The sound source is located at $\boldsymbol{x}_s = (0.95, 0.78, 1.73)$ in polar coordinates.

2.2.5 Soundfield Recording Using Spherical Microphone Arrays

In the past decade, soundfield recording using different structured higher-order microphone arrays has been developed, including a spherical microphone array [138], multiple circular microphone arrays [139], and a planar microphone array [140]. Samarasinghe *et al.* used a set of higher-order microphones (circular microphone

arrays in [105], spherical microphone arrays in [141]) to develop the measuring techniques that are more suitable for large spatial regions. These measuring techniques generally use a spherical Fourier transform [16] to decompose the multi-channel recordings into the spherical harmonics domain and then achieve a full measurement of the spherical harmonic coefficients. Due to space limitations, this subsection mainly reviews sound field recording using a typical spherical microphone array, which is used by this thesis in the following chapters.

As discussed, an arbitrary point sound pressure, within the recording area, can be represented in the spherical harmonics domain using the spherical harmonics expansion [16],

$$P(\boldsymbol{x}, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k) j_n(kr) Y_{nm}(\theta, \phi). \tag{2.32}$$

We exploit the property that the $Y_{nm}(\cdot)$ is orthonormal over the space, multiply $Y_{nm}^*(\cdot)$ at both the sides of (2.21) and then integrate it over the two-dimensional space to yield,

$$\alpha_{nm}(k) j_n(kr) = \int_0^{\pi} \int_0^{2\pi} P(\boldsymbol{x}, k) Y_{mn}^*(\theta, \phi) d\theta d\phi \tag{2.33}$$

which requires an infinite number of microphones in theory. However, in practice, only a limited number of microphones are available. In this case, we can approximate the continuous sampling in (2.33) by uniformly distributing the limited number of microphones over the surface of the spherical microphone array,

$$\alpha_{nm}(k) = \frac{1}{j_n(kr)} \sum_{j=1}^{M} a_j P(\boldsymbol{x}_j, k)\, Y_{nm}^*(\theta_j, \phi_j) \tag{2.34}$$

in which M denotes the number of microphones, $*$ signifies the conjugate transpose operation, and a_j works as the weight of each microphone (known in advance) to ensure the error between the measured and theoretical estimations is as small as possible. The above analysis mainly uses an open-sphere microphone array. However, a practical microphone array can be rigid (e.g., the commercial Eigenmike), whose acoustic reflections caused by the surface are non-negligible. In this case, the recordings of spherical harmonic coefficients in (3.11) require a slight modifica-

tion to account for the influence caused by the microphone array's surface acoustic reflections [142],

$$\alpha_{nm}(k) = \frac{1}{b_n(kr)} \sum_{j=1}^{M} a_j P(\boldsymbol{x}_j, k)\, Y_{nm}^{*}(\theta_j, \phi_j) \tag{2.35}$$

where

$$b_n(kr) = j_n(kr) - \frac{j_n^{'}(kr_o)}{h_n^{'}(kr_o)} h_n(kr) \tag{2.36}$$

where r_o denotes the radius of the spherical microphone array, $j_n^{'}(\cdot)$ and $h_n^{'}(\cdot)$ represent the partial derivative of spherical Bessel function $j_n(\cdot)$ and spherical Hankel function $h_n(\cdot)$, respectively.

2.3 Relative Transfer Function (ReTF)

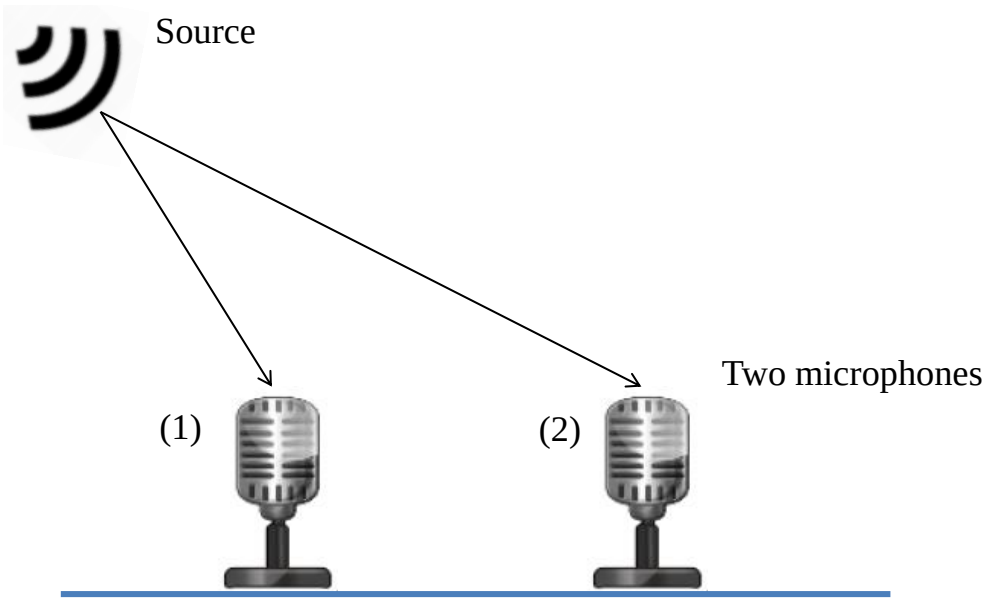


Figure 2.5: Recording using a microphone pair.

This subsection reviews the relative transfer function (ReTF)[1], between a mi-

[1]Note that the room transfer function is denoted as 'RTF'. Hence, this thesis abbreviates the

crophone pair, which has been widely used to address sound source localization in diverse acoustic environments [143, 30, 144].

2.3.1 Definition of ReTF

Figure 2.5 presents an acoustic event where a pair of microphones measures the incoming soundfield due to a single sound source. The recording in the time domain, measured by the microphone pair in a noiseless environment, is represented as,

$$\begin{aligned} p(\boldsymbol{x}_1, n) &= s(n) * a(\boldsymbol{x}_1, n) \\ p(\boldsymbol{x}_2, n) &= s(n) * a(\boldsymbol{x}_2, n) \end{aligned} \tag{2.37}$$

where $s(n)$ is the source signal, $a(\boldsymbol{x}_j, n)$ is the room impulse response between the sound source to the j-th microphone ($j = 1, 2$). This expression can be represented in the frequency domain using a Fourier transform as follows,

$$\begin{aligned} P(\boldsymbol{x}_1, k) &= S(k)A(\boldsymbol{x}_1, k) \\ P(\boldsymbol{x}_2, k) &= S(k)A(\boldsymbol{x}_2, k) \end{aligned} \tag{2.38}$$

where $k = 2\pi f/c$ is the wavenumber, f is the frequency bin, c is the speed of sound, $P(\boldsymbol{x}_j, k)$, $S(k)$ and $A(\boldsymbol{x}_j, k)$ denote the Fourier transforms of $p(\boldsymbol{x}_j, n)$, $s(n)$ and $a(\boldsymbol{x}_j, n)$, respectively.

The ReTF is formally defined as the ratio between the acoustic transfer functions of the two microphones in (2.38) [143, 30],

$$Q(\boldsymbol{x}_1, k) = \frac{P(\boldsymbol{x}_1, k)}{P(\boldsymbol{x}_2, k)} = \frac{A(\boldsymbol{x}_1, k)}{A(\boldsymbol{x}_2, k)} \tag{2.39}$$

where the second microphone is taken as the reference microphone. Early research in [145, 45] revealed several unique properties of ReTF as follows: (i) it is independent of the time-varying source signal; (ii) in a static environment, its only varying degree of freedom is the sound source position; and (iii) it can be easily estimated in a noisy environment (e.g., see an estimator of ReTF in the next subsection).

relative transfer function into 'ReTF' to avoid confusion.

2.3.2 Estimation of ReTF

In practice, the microphone recordings inevitably contain some additive noise,

$$\begin{aligned} P(\boldsymbol{x}_1, k) &= S(k)A(\boldsymbol{x}_1, k) + V(\boldsymbol{x}_1, k) \\ P(\boldsymbol{x}_2, k) &= S(k)A(\boldsymbol{x}_2, k) + V(\boldsymbol{x}_2, k) \end{aligned} \tag{2.40}$$

in which $V(\boldsymbol{x}_1, k)$ and $V(\boldsymbol{x}_2, k)$ denote the additive noise at the two microphones, respectively. To relieve the negative effects caused by noise when calculating the ReTF, it is of the common approach [45, 57] to exploit a biased estimator of ReTF using the power spectral density (PSD) and cross PSD (CPSD) of the measured signals,

$$\begin{aligned} Q(\boldsymbol{x}_1, k) &= \frac{S_{p_1p_2}(k)}{S_{p_2p_2}(k) - S_{v_2v_2}(k)} \\ &= \frac{S_{ss}(k)A(\boldsymbol{x}_1, k)A^*(\boldsymbol{x}_2, k)}{S_{ss}(k)|A(x_2, k)|^2} = \frac{A(\boldsymbol{x}_1, k)}{A(\boldsymbol{x}_2, k)} \end{aligned} \tag{2.41}$$

where

$$\begin{aligned} S_{p_1p_2}(k) &= \mathbb{E}\left\{P(\boldsymbol{x}_1, k)P^*(\boldsymbol{x}_2, k)\right\} \\ S_{p_2p_2}(k) &= \mathbb{E}\left\{P(\boldsymbol{x}_2, k)P^*(\boldsymbol{x}_2, k)\right\} \\ S_{v_2v_2}(k) &= \mathbb{E}\left\{V(\boldsymbol{x}_2, k)\,V^*(\boldsymbol{x}_2, k)\right\} \\ S_{ss}(k) &= \mathbb{E}\left\{S(k)S^*(k)\right\} \end{aligned} \tag{2.42}$$

where $\mathbb{E}[\cdot]$ denotes statistical expectation and $*$ denotes the conjugate transpose, $S_{p_1p_2}(k)$ denotes the CPSD between the received signal at the two microphones, $S_{ss}(k)$ denotes the PSD of the sound source signal, $S_{p_2p_2}(k)$ and $S_{v_2v_2}(k)$ denote the PSD of the received signal and noise signal at the second channel, respectively. Note that the estimator in (2.41) exploits the assumption that the sound pressure of sound source and additive noise are not correlated. However, the noise PSD of $S_{v_2v_2}(k)$ is unknown in practice. One potential solution is to use some state-of-art power spectral density techniques, such as [146, 147], to estimate the noise PSD. After that, the estimated noise PSD is substituted into the denominator in (2.41)

to compute the ReTF. For simplicity, the methods in [45, 57] estimated the ReTF by neglecting the noise PSD $S_{v_2v_2}(k)$ in the denominator of (2.41), i.e.,

$$Q(\boldsymbol{x}_1, k) \approx \frac{S_{p_1p_2}(k)}{S_{p_2p_2}(k)} \tag{2.43}$$

which is named as a biased ReTF estimator.

2.3.3 Application into Sound Source Localization

Early use of ReTF for source localization in [26] was to extract the TDOA of source signal in the first stage, which was applied to address single source localization in the second stage. An investigation in [145] reveals that the ReTF intrinsically embeds in a low-dimensional manifold, which is solely governed by its source position. Even in a static reverberant environment, the sound source position is the only varying degree-of-freedom of the ReTFs in the enclosure. Hence, the ReTF is capable of recovering the unknown source position in reverberant environments. An online scheme using ReTFs to track multiple moving speakers in a reverberant environment was presented in [148]. More recently, Brendel *et al.* exploited the ReTF to propose an expectation-maximization (EM) based algorithm, achieving a joint speaker number counting and localization in adverse acoustic conditions [149]. With several pairs of microphones, Laufer-Goldshtein *et al.* have exploited the ReTF for both semi-supervised single source localization [45, 57, 81, 58] and source tracking [31], respectively. With a binaural setup of microphones, Li *et al.* achieved a supervised multiple source localization using the direct-path ReTF where only a single source is active [90]. Opochinsky *et al.* then fed the ReTF into a deep-learning network for weakly-supervised ranking-based source localization [59].

Motivated by the extensive applications of ReTF, this thesis will investigate the other type of sound source feature called the *relative harmonic coefficients* in the spherical harmonics domain. As mentioned in Chapter 1.2, this new source feature has several significant properties suitable to address sound source localization problems, including (i) its independence from the time-varying source signal and sole dependence on the source position even in a reverberant environment; (ii) easy estimations in noisy environments from the higher-order microphone array

recordings; (iii) a significant spatial resolution due to its unique directivity pattern over space; and (iv) exploitable by an overlapped frame detector to simplify the challenging localization of multiple sources into single source localization issues. This thesis intends to apply this source feature to develop some novel localization algorithms under different acoustic scenarios. In the next chapter, we will formally define the relative harmonic coefficients in the spherical harmonics domain and use it to develop a decoupled DOA estimator in diverse environments.

2.4 Summary

Since sound source localization has been a hot topic over decades, this chapter first presents an extensive update of the localization algorithms and up-to-date research progress in the literature review. After that, we introduce the spherical harmonics representation of a measured soundfield, which acts as a fundamental tool for any spherical harmonic domain-based signal processing technique and application. Finally, we give a detailed introduction of a widely used feature called relative transfer function (ReTF), including its definition, estimator, and unique properties. The ReTF introduced in this chapter motivates us to develop a newly spherical harmonics domain source feature called *relative harmonic coefficients*, which are exploited by several novel source localization algorithms in the following chapters.

Chapter 3

Decoupled Direction-of-arrival Estimation Using Relative Harmonic Coefficient

***Overview**: This chapter presents a source feature called the relative harmonic coefficients based on the spherical harmonics representation of a measured sound-field. This source feature is shown to be easily estimated from the noisy higher-order microphone array recordings. We derive a closed-form expression of this feature, where the elevation and azimuth components are decoupled. Hence, the relative harmonic coefficients relate to the source elevation and azimuth independently, which enables them to act as features capable of recovering unknown source elevation and azimuth separately. Based on this property, we develop two decoupled source direction of arrival estimation algorithms. The proposed algorithms are highlighted by a large reduction of computational complexity, thus enabling a direct application for sound source tracking. Simulation results, using both a static and moving sound source, confirm the proposed methods are computationally efficient while achieving competitive localization accuracy. Additionally, this chapter also shows that this developed approach has another advantage to be applied as an acoustic enhancement tool for the noisy higher-order microphone array recordings.*

3.1 Introduction

As reviewed in Chapter 2, sound source localization methods can be divided into different types depending on the source features used as the inputs of the algorithms, such as the generalized cross-correlation [28], first-order Ambisonics [22], phase difference [97], STFT coefficients [80], modal coherence patterns [23] and intensity/pseudointensity vectors [25, 98] (see more about the features in the Table 2.2). Intuitively, the source features have a great influence on the algorithm's performance as they contain relevant clues of the sound source(s) to be localized. For example, if a given source feature depends on the time-varying source signal, such as the STFT coefficients, it is of greater difficulty to use this source feature for accurate source localization. Hence, source localization algorithms require features that have less dependency on the time-varying source signal but only depend on the source position. The ReTF, introduced in Chapter 2.3, is one of the promising features. An investigation in [145] reveals that the ReTF is solely governed by its source position, even in a static reverberant environment. As a result, the source feature of ReTF has been widely used by recently proposed sound source localization/tracking techniques [31, 45, 57, 81, 90, 59].

Motivated by the wide applications of ReTF, this chapter presents a spherical harmonics domain source feature called the *relative harmonic coefficients*, as well as its estimators from the noisy higher-order microphone array recordings. The derived expression of the relative harmonic coefficients confirms that it is solely dependent on the source DOA. Note that traditional DOA estimation methods, such as the popular generalized cross-correlation phase transform (GCC-PHAT) [28], steered response power (SRP) [32] and SRP-phase transforms (SRP-PHAT) [33] based techniques, achieve satisfying localization accuracy while associating with a significant computational complexity since they require a 2-D grid searching over all possible directions. Generally, a higher grid resolution to sample the directional space increases the accuracy of the algorithms at the cost of an additional higher computational expense. However, the source's elevation and azimuth angles are decoupled in the relative harmonic coefficients, thus it enables us to estimate the source's elevation and azimuth in two separate stages. Based on the above properties, this chapter presents two decoupled solutions to achieve accurate two-

dimensional DOA estimates which are attributed by a dramatically reduced computational complexity. The proposed methods are finally validated using static source localization as well as a direct application for tracking a moving sound source. Additionally, we also show that the developed method by this chapter has a direct application for acoustic enhancement in the spherical harmonics domain. Extensive simulations, using the spherical microphone array measurements from a far-field speaker, confirm the effective denoising performance of this method in noisy environments.

The remainder of the chapter is structured as follows. Firstly, we introduce the definition of the relative harmonic coefficients, the feature's estimator, and closed-form expression. Then, we propose two decoupled DOA estimators using the estimated relative harmonic coefficients. Thereafter, we show how to apply the estimated source feature into an acoustic enhancement schema. Finally, extensive simulation results are presented to validate the proposed methods.

3.2 Problem Formulation

Assume an active sound source propagating from an arbitrary DOA over the 2-D space, e.g., (ϑ_s, φ_s) where $0 < \vartheta_s < \pi$, $0 < \varphi_s < 2\pi$. The sound wave is impinging on a higher-order microphone array (see Fig. 3.1). The array comprises of M microphones whose polar coordinates are $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$ $(j = 1, \ldots, M)$, with respect to its local origin O. The sound pressure, recorded by the array is represented in the frequency domain by

$$\begin{aligned} \overline{P}(\boldsymbol{x}_j, k) =& P(\boldsymbol{x}_j, k) + V(\boldsymbol{x}_j, k) \\ =& S(k)A(\boldsymbol{x}_j, k) + V(\boldsymbol{x}_j, k), \quad j = 1, \cdots, M \end{aligned} \tag{3.1}$$

where $k = 2\pi f/c$ is the wavenumber, with f the frequency bin and c the speed of sound, $S(k)$ is the source signal, $A(\boldsymbol{x}_j, k)$ is the acoustic transfer function from the source to the j-th microphone, $\overline{P}(\boldsymbol{x}_j, k)$ and $P(\boldsymbol{x}_j, k)$ denote the noisy and clean sound pressure, $V(\boldsymbol{x}_j, k)$ denotes the additive noise signal.

This chapter aims to use a spherical harmonics domain source feature of relative harmonic coefficients, estimated from the original noisy recordings, to develop

low-complexity DOA estimations with competitive localization accuracy. In the meanwhile, an acoustic enhancement is also proposed for denoising the noisy measurements in the spherical harmonics domain. In the next section, we introduce the source feature called the relative harmonic coefficients which are to be used by the developed approaches.

3.3 Relative Harmonic Coefficients (RHC)

This section presents a detailed introduction of the spherical harmonics domain source feature called the relative harmonic coefficients, whose properties motivate the decoupled localization approaches developed in the next section.

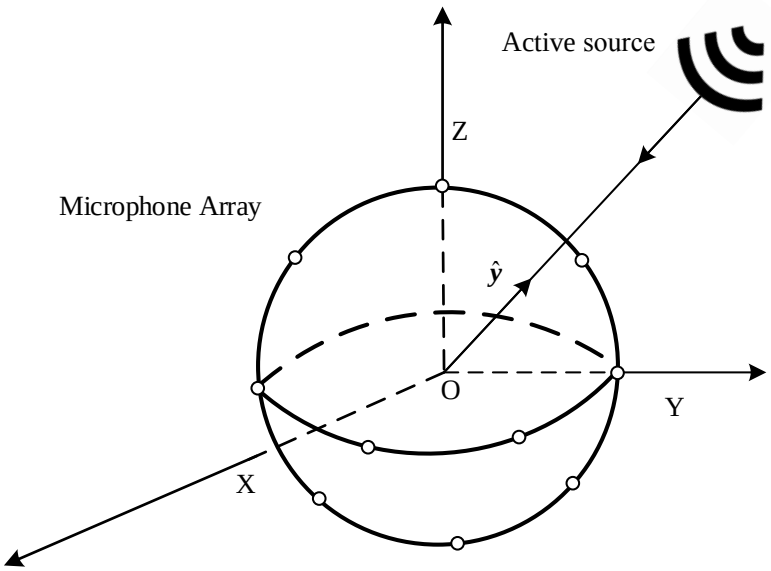


Figure 3.1: Soundfield recording using a spherical microphone array.

3.3.1 Definition of RHC

Based on the spherical harmonics decomposition of a soundfield, the sound pressure at an arbitrary point microphone $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$, $j = 1, \cdots, M$ within a recording

area (see Figure 3.1) can be expressed as [16],

$$P(\boldsymbol{x}_j, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k)\, j_n(kr)\, Y_{nm}(\theta_j, \phi_j) \tag{3.2}$$

where $n(\geqslant 0)$ and m are integers, $N = \lceil kr \rceil$ is the truncated order of the soundfield [103], $\alpha_{nm}(k)$ is the spherical harmonic coefficient, $j_n(\cdot)$ is the spherical Bessel function of the first kind, $Y_{nm}(\theta_j, \phi_j)$ is the spherical harmonics function.

Assuming the soundfield decomposed by (3.2) is generated by a single sound source, the relative harmonic coefficients (RHC) of order n and mode m is formally defined as the ratio between $\alpha_{nm}(k)$ and $\alpha_{00}(k)$ [1, 150, 151],

$$\beta_{nm}(k) = \frac{\alpha_{nm}(k)}{\alpha_{00}(k)}. \tag{3.3}$$

Let the frequency band of interest be $[k_{\min}, k_{\max}]$. Then, we propose a $k \times 1$ feature vector for each (n, m) mode,

$$\boldsymbol{\beta}_{nm} = \Big[\beta_{nm}(k_1), \beta_{nm}(k_2), \cdots, \beta_{nm}(k_F)\Big]^T \tag{3.4}$$

where $k_{\min} \leqslant k_1, \cdots, k_F \leqslant k_{\max}$. We combine feature vectors of all the spherical harmonic modes to obtain $F \times (N+1)^2$ matrix of relative harmonic coefficients as,

$$\boldsymbol{B} = \Big[\boldsymbol{\beta}_{00}, \boldsymbol{\beta}_{1,-1}, \cdot\cdot, \boldsymbol{\beta}_{nm}, \cdot\cdot, \boldsymbol{\beta}_{NN}\Big]. \tag{3.5}$$

3.3.2 Estimation of RHC

This subsection proposes two methods to estimate the relative harmonic coefficients in the presence of noise, one using the point-to-point relative transfer function concept, and another using the spherical harmonic decomposition. We focus on the estimation at a single frequency bin as that of a wide frequency band follows a similar process.

Estimation of RHC using the Point-to-Point Relative Transfer Function

Let us define a spatial function at $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$ using the vector of $[\beta_{00}(k), \cdots, \beta_{NN}(k)]^T$,

$$\begin{aligned} Q(\boldsymbol{x}_j, k) &= \sum_{n=0}^{N} \sum_{m=-n}^{n} \beta_{nm}(k)\, j_n(kr)\, Y_{nm}(\theta_j, \phi_j) \\ &= \frac{\sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k)\, j_n(kr)\, Y_{nm}(\theta_j, \phi_j)}{\alpha_{00}(k)} \\ &= \frac{P(\boldsymbol{x}_j, k)}{\alpha_{00}(k)} \end{aligned} \tag{3.6}$$

which is derived using representation of the sound pressure of $P(\boldsymbol{x}_j, k)$ implied by (3.2). The coefficient $\alpha_{00}(k)$ at the denominator represents the sound pressure for the point microphone located at $x_o = (0, 0, 0)$,

$$\begin{aligned} P(\boldsymbol{x}_o, k) &= \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k)\, j_n(0)\, Y_{nm}(0, 0) \\ &= \rho \alpha_{00}(k) \end{aligned} \tag{3.7}$$

where $\rho = 1/\sqrt{4\pi}$ is a fixed constant so that we omit it in the following for notational convenience. Substitute (3.7) to (3.6), the defined $Q(\boldsymbol{x}_j, k)$ is rewritten as,

$$Q(\boldsymbol{x}_j, k) = \frac{P(\boldsymbol{x}_j, k)}{P(\boldsymbol{x}_o, k)}. \tag{3.8}$$

Since the relative harmonic coefficients are defined at the case of a single sound source, (3.6) is further simplified into,

$$Q(\boldsymbol{x}_j, k) = \frac{S(k) A(\boldsymbol{x}_j, k)}{S(k) A(\boldsymbol{x}_o, k)} = \frac{A(\boldsymbol{x}_j, k)}{A(\boldsymbol{x}_o, k)} \tag{3.9}$$

where $A(\boldsymbol{x}_o, k)$ represents the acoustic transfer function from the source to the microphone located at $\boldsymbol{x}_o$. Above inference by (3.9) implies $Q(\boldsymbol{x}_j, k)$ coincides with the ReTF between the pair of microphones located at $\boldsymbol{x}_j$ and x_o respectively. Therefore, the relative harmonic coefficients can be calculated using a spherical

harmonics decomposition of the measured ReTF.

Using a higher-order microphone array, we first approximate sound pressure at the origin of the array by the addition of all the recordings on the array. Then, we use the estimator of (2.43) [45,57], given in the Section 2.3, to calculate the ReTFs of all the microphones on the array. Finally, the relative harmonic coefficients can be estimated by decomposing the M estimations of the ReTFs as,

$$\beta_{nm}(k) = \frac{1}{j_n(kr)} \sum_{j=1}^{M} a_j \, Q(\boldsymbol{x}_j, k) \, Y_{nm}^*(\theta_j, \phi_j) \tag{3.10}$$

in which a_i works as the weight of each microphone to ensure the right side in (3.10) equals to the left side. Note that the decomposition in (3.10) is suffered by the "Bessel zero problem" at low frequencies, causing erroneous estimations of the desired spherical harmonics coefficients because the noise signal can be easily amplified [141].

Estimating RHC using Spherical Harmonic Coefficients

Let us directly decompose the measured noisy soundfield into the spherical harmonics domain [103],

$$\overline{\alpha}_{nm}(k) = \frac{1}{j_n(kr)} \sum_{j=1}^{M} a_j \overline{P}(\boldsymbol{x}_j, k) \, Y_{nm}^*(\theta_j, \phi_j) \tag{3.11}$$

where $\overline{P}(\boldsymbol{x}_j, k)$ denotes the noisy sound pressure at the j-th microphone. Since there is only a single sound source presented in the recordings, we can rewrite $\overline{\alpha}_{nm}(k)$ in (3.11) as,

$$\begin{aligned} \overline{\alpha}_{nm}(k) &= \alpha_{nm}(k) + \gamma_{nm}(k) \\ &= \beta_{nm}(k)\alpha_{00}(k) + \gamma_{nm}(k) \end{aligned} \tag{3.12}$$

where $\alpha_{nm}(k)$ and $\gamma_{nm}(k)$ represents the spherical harmonic coefficients due to the source and noise signal, respectively. The $\beta_{nm}(k)$ in (3.12) are the already defined relative harmonic coefficients that relate to the $\alpha_{00}(k)$. However, in practice, only

the noisy relative harmonic coefficients are available,

$$\overline{\alpha}_{00}(k) = \alpha_{00}(k) + \gamma_{00}(k). \tag{3.13}$$

Note that $\beta_{nm}(k)$ is independent of the source signal, thus it is constant over the time-varying signal. In order to alleviate the negative effects caused by the noise when calculating the $\beta_{nm}(k)$, we exploit the power spectral density (PSD) and CPSD (cross PSD) of the measured signals,

$$\frac{S_{\overline{\alpha}_{nm}\overline{\alpha}_{00}}(k)}{S_{\overline{\alpha}_{00}\overline{\alpha}_{00}}(k) - S_{\gamma_{00}\gamma_{00}}(k)} = \frac{\beta_{nm}(k) S_{\alpha_{00}\alpha_{00}}(k)}{S_{\alpha_{00}\alpha_{00}}(k)} = \beta_{nm}(k) \tag{3.14}$$

where

$$\begin{aligned} S_{\overline{\alpha}_{nm}\overline{\alpha}_{00}}(k) &= \mathbb{E}\left\{\overline{\alpha}_{nm}(k)\overline{\alpha}^{*}_{00}(k)\right\} \\ S_{\overline{\alpha}_{00}\overline{\alpha}_{00}}(k) &= \mathbb{E}\left\{\overline{\alpha}_{00}(k)\overline{\alpha}^{*}_{00}(k)\right\} \\ S_{\gamma_{00}\gamma_{00}}(k) &= \mathbb{E}\left\{\gamma_{00}(k)\gamma^{*}_{00}(k)\right\} \\ S_{\alpha_{00}\alpha_{00}}(k) &= \mathbb{E}\left\{\alpha_{00}(k)\alpha^{*}_{00}(k)\right\} \end{aligned} \tag{3.15}$$

with $\mathbb{E}[\cdot]$ denoting the statistical expectation over the time-varying signal. Note that (3.14) exploits the fact that the spherical harmonic coefficients of source signal and noise signal are uncorrelated because their corresponding sound pressure are assumed to be uncorrelated. However, the noise PSD of $S_{\gamma_{00}\gamma_{00}}(k)$ at the denominator of (3.14) is still unknown. Some state-of-art power spectral density techniques are available to update the $S_{\gamma_{00}\gamma_{00}}(k)$ [146, 147]. For simplicity in practice, we adopt a biased feature estimator by neglecting it, so that the source feature can be represented using,

$$\beta_{nm}(k) \approx \frac{S_{\overline{\alpha}_{nm}\overline{\alpha}_{00}}(k)}{S_{\overline{\alpha}_{00}\overline{\alpha}_{00}}(k)}. \tag{3.16}$$

3.3.3 A Theoretical Expression for RHC

With the intention to analyze the properties of RHCs and their suitability as a feature for DOA applications, here, we derive a theoretical expression for RHCs assuming free field propagation. Let us follow the common assumption to represent

observations over the recording area using plane wave modeling [120,121], because the aperture of the recording area is much smaller when compared to its distance to the sound source. Following the spherical harmonics decomposition of the plane waves [103], sound pressure at the j-th microphone, due to direct-path recording from the sound source, can be represented as,

$$P(x_j,k)=\sum_{n=0}^{N}\sum_{m=-n}^{n} S(k)4\pi i^n Y_{nm}^*(\vartheta_s,\varphi_s)\, j_n(kr)\, Y_{nm}(\theta_j,\phi_j) \tag{3.17}$$

whose spherical harmonic coefficient due to the sound source is,

$$\alpha_{nm}(k)=S(k)4\pi i^n Y_{nm}^*(\vartheta_s,\varphi_s). \tag{3.18}$$

Using (3.3) and (3.18), we derive the expression of the feature with order n and mode m in a free-field environment as:

$$\beta_{nm}(k)=2\sqrt{\pi} i^n Y_{nm}^*(\vartheta_s,\varphi_s). \tag{3.19}$$

For the N-th order microphone array, our feature vector for the source from direction of (ϑ_s,φ_s) is,

$$\boldsymbol{\beta}(\vartheta_s,\varphi_s)=\left[1, 2\sqrt{\pi} i Y_{1,-1}^*(\vartheta_s,\varphi_s),\cdots, 2\sqrt{\pi} i^N Y_{NN}^*(\vartheta_s,\varphi_s)\right]^T \tag{3.20}$$

whose properties are briefly summarized as follows:

- Source signal independent: its only degree-of-freedom coincides with the source DOA.
- Frequency independent: in practice, even when recording the same sound source impinging the array from (ϑ_s,φ_s), the features may slightly differ from the expected frequency-independence. In this case, a frequency smoothing is then suggested:

  $$\overline{\boldsymbol{\beta}}(\vartheta_s,\varphi_s)=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\beta}(k,\vartheta_s,\varphi_s) \tag{3.21}$$

  where $\overline{\boldsymbol{\beta}}(\vartheta_s,\varphi_s)$ denotes the smoothed feature vector over the K frequency bins of interest. This frequency smoothed version also reduces the computational

complexity as it avoids repetitive processing at each frequency bin respectively.

- Uniqueness over space: the feature vector of (3.20) is unique over the space.

- Calculated feature set: assume our defined sound source area for source localization is $\mathbf{\Phi} = \Big\{(\vartheta_s, \varphi_s) : 0 < \vartheta_s \leq \pi, 0 < \varphi_s \leq 2\pi\Big\}$, we can calculate a feature set using (3.20):

$$\mathcal{H} = \{\boldsymbol{\beta}(\vartheta_1, \varphi_1), \boldsymbol{\beta}(\vartheta_2, \varphi_2), \cdots, \boldsymbol{\beta}(\vartheta_S, \varphi_S)\} \tag{3.22}$$

  which comprises all possible candidate feature vectors over the defined source area (S denotes the total number of discrete DOA samples of $\mathbf{\Phi}$). Note that the feature set of $\mathcal{H}$ is calculated in advance without any prior recordings.

Assuming the relative harmonic coefficients due to the desired sound source are practically estimated (e.g., $\overline{\beta}_{nm}$), we can compare it to the calculated set of $\mathcal{H}$ to recover its original DOA,

$$\arg \min_{(\vartheta_s, \varphi_s)} \sum_{n=0}^{N} \sum_{m=-n}^{n} \mid \overline{\beta}_{nm} - 2\sqrt{\pi} i^n Y_{nm}^*(\vartheta_s, \varphi_s) \mid^2 \tag{3.23}$$

which uses a distance-based metric. However, this approach requires an exhaustive search over the 2-D directional set of $\mathcal{H}$. As explained in the next section, we show the relative harmonic coefficients are capable for decoupled DOA estimations, while achieving a large reduction of the computational complexity.

3.4 Decoupled DOA Estimation

This section exploits the relative harmonic coefficients to develop two decoupled source DOA estimations, which avoid the exhaustive search over the 2-D space.

We review the expression of the spherical harmonics function,

$$Y_{nm}(\theta, \phi) = (-1)^m \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_{nm}(\cos\theta) e^{im\phi}. \tag{3.24}$$

Substituting (3.24) into (3.19), we have the detailed expression of the relative

harmonic coefficients,

$$\beta_{nm}(\vartheta_s, \varphi_s) = \sqrt{\frac{(2n+1)(n-m)!}{(n+m)!}} P_{nm}(\cos\vartheta_s) i^n e^{-im\varphi_s} \tag{3.25}$$

where the associated Legendre function, i.e., $P_{nm}(\cdot)$, is a real-valued function. Note that the dependence on frequency in (3.25) is omitted for convenience. Table 3.1 lists exact expressions of (3.25) for the spherical harmonic modes up to $n = 2$ and $m = 1$. The specifications of the relative harmonic coefficients in Table 3.1 imply that the components of source elevation ϑ_s and azimuth φ_s are decoupled. Hence, assuming the relative harmonic coefficients are accurately estimated, we can recover the source's elevation and azimuth in two separate stages. In the following two subsections, we detail two methods that exploit this property to perform DOA estimations.

Table 3.1: Relative harmonic coefficients up to the 2nd order.

(n, m)	$\beta_{nm}(k)$	(n, m)	$\beta_{nm}(k)$
(0, 0)	1	(2,-2)	$\sqrt{\frac{15}{8}}\sin^2(\vartheta_s)e^{2i\varphi_s}$
(1,-1)	$i\sqrt{\frac{3}{2}}\sin(\vartheta_s)e^{i\varphi_s}$	(2,-1)	$\sqrt{\frac{15}{8}}\sin(2\vartheta_s)e^{i\varphi_s}$
(1, 0)	$i\sqrt{3}\cos(\vartheta_s)$	(2, 0)	$\sqrt{\frac{5}{4}}\left(3\cos^2(\vartheta_s)-1\right)$
(1, 1)	$-i\sqrt{\frac{3}{2}}\sin(\vartheta_s)e^{-i\varphi_s}$	(2, 1)	$-\sqrt{\frac{15}{8}}\sin(2\vartheta_s)e^{-i\varphi_s}$

3.4.1 The First Method

This method estimates the source elevation in the first stage, which is then used to recover the azimuth in the second stage.

The First Stage: Elevation Estimation

The magnitude of the relative harmonic coefficients in (3.25) is,

$$|\beta_{nm}(\vartheta_s)| = \sqrt{\frac{(2n+1)(n-m)!}{(n+m)!}} |P_{nm}(\cos\vartheta_s)| \tag{3.26}$$

which only depends on the elevation ϑ_s. Figure 3.2 demonstrates an example of the $|\beta_{1,-1}(\vartheta_s)|$ where the $0 < \vartheta_s < \pi$. Combining (3.26) up to the N-th order, we have a vector of the magnitude,

$$|\boldsymbol{\beta}(\vartheta_s)| = [1, |\beta_{1,-1}(\vartheta_s)|, \cdots, |\beta_{NN}(\vartheta_s)|] \tag{3.27}$$

whose properties are summarized as follows:

- Unique mapping: the vector of (3.27) has a unique mapping to the elevation when $0 < \vartheta_s \leq \pi/2$ (see Fig. 3.2).
- Calculated set: given the range of $0 < \vartheta_s \leq \pi/2$, we have a unique set of (3.27) (see Fig. 3.2 for $|\beta_{1,-1}(\vartheta_s)|$):

$$\mathcal{H}_{\text{mag}} = \{|\boldsymbol{\beta}(\vartheta_1)|, |\boldsymbol{\beta}(\vartheta_2)|, \cdots, |\boldsymbol{\beta}(\vartheta_{S1})|\} \tag{3.28}$$

  where $S1$ is the number of discrete samples of elevation. Also, the set of $\mathcal{H}_{\text{mag}}$ is calculated in advance without any prior recordings.

- Symmetric: the vector of (3.27) is symmetric to $\pi/2$ because $|P_{nm}(\cos\vartheta_s)| = |P_{nm}(\cos(\pi - \vartheta_s))|$ (see Fig. 3.2). Hence, we also have a unique mapping and set when $\pi/2 < \vartheta_s < \pi$.

Assuming the source's magnitude of relative harmonic coefficients are calculated, we show how to recover its elevation using the following two steps,

Step 1: Since the magnitude of the feature is symmetric to $\pi/2$, it cannot distinguish whether ϑ_s lies between $(0, \pi/2)$ or $(\pi/2, \pi)$. However, this can be known from the imaginary part of $\beta_{10}(k)$,

$$\text{Im}\{\beta_{10}(k)\} = \sqrt{3}\cos(\vartheta_s) \tag{3.29}$$

whose positive or negative characteristic only depends on the ϑ_s. Hence, we claim the estimated $\overline{\vartheta}_s$,

$$\begin{cases} 0 < \overline{\vartheta}_s < \pi/2, & \text{if } \text{Im}\{\overline{\beta}_{10}\} > 0 \\ \pi/2 \leq \overline{\vartheta}_s < \pi, & \text{if } \text{Im}\{\overline{\beta}_{10}\} \leq 0. \end{cases} \tag{3.30}$$

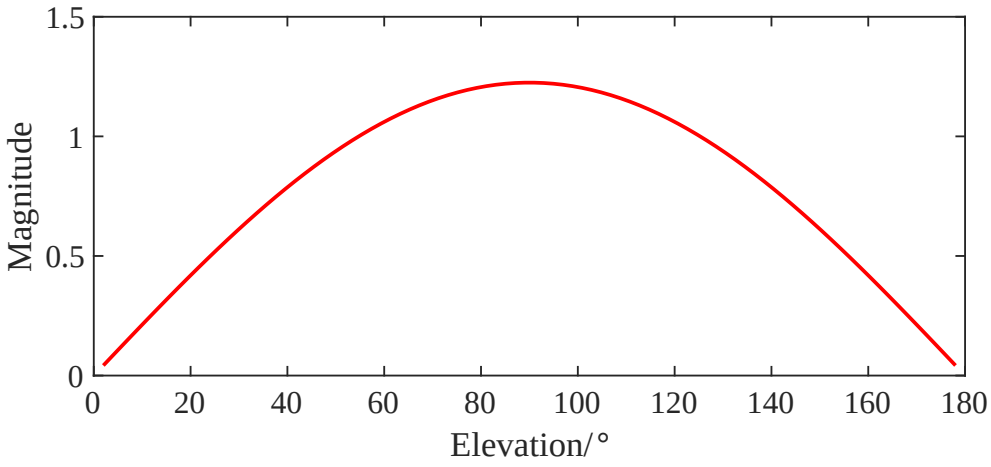


Figure 3.2: $|\beta_{1,-1}(\vartheta_s)|$ for the elevation ranging from 0 to π.

Step 2: Exploiting the unique mapping between $(0, \pi/2)$ or $(\pi/2, \pi)$, we then use a distance-based metric to compare it to the calculated set of $\mathcal{H}_{\text{mag}}$ to recover the source's elevation.

The Second Stage: Azimuth Estimation

Since the source elevation has already been estimated, we can recover the source's φ_s by searching over all possible azimuths as follows,

$$\arg\min_{(\varphi_s)} \sum_{n=0}^{N} \sum_{m=-n}^{n} \mid \overline{\beta}_{nm} - 2\sqrt{\pi} i^n Y_{nm}^*(\overline{\vartheta}_s, \varphi_s) \mid^2 \tag{3.31}$$

where $\overline{\beta}_{nm}$ is the recorded relative harmonic coefficients.

3.4.2 The Second Method

Different from the above method, the second approach first estimates the azimuth and then recovers the source elevation.

The First Stage: Azimuth Estimation

Let us define the ratio between the imaginary and real parts of the relative harmonic coefficients in (3.25),

$$\begin{aligned}\gamma_{nm}(\varphi_s) &= \frac{\mathrm{Im}\{\beta_{nm}(\vartheta_s,\varphi_s)\}}{\mathrm{Re}\{\beta_{nm}(\vartheta_s,\varphi_s)\}} \\ &= \begin{cases} -\tan(m\varphi_s), & \text{if } n = 0,2,4,\cdots \\ -\cot(m\varphi_s), & \text{if } n = 1,3,5,\cdots \end{cases}\end{aligned} \tag{3.32}$$

which only depends on the source's φ_s ($\mathrm{Re}\{\beta_{nm}(\vartheta_s,\varphi_s)\} \neq 0$). Combining the cases of (3.32) up to order N, we have a vector,

$$\boldsymbol{\gamma}(\varphi_s) = [\gamma_{00}(\varphi_s), \gamma_{1,-1}(\varphi_s), \cdots, \gamma_{NN}(\varphi_s)] \tag{3.33}$$

whose properties are as follows:

- Unique mapping: the (3.33) has a unique mapping to the φ_s when $0 < \varphi_s < \pi$.
- Calculated set: given the range when $0 < \varphi_s < \pi$, we have a unique set of (3.33):

$$\mathcal{H}_{\tan} = \{\boldsymbol{\gamma}(\varphi_1), \boldsymbol{\gamma}(\varphi_2), \cdots, \boldsymbol{\gamma}(\varphi_{S2})\} \tag{3.34}$$

  where $S2$ denotes the number of discrete samples.
- Periodic: the vector in (3.33) is periodic by π because of the tan/cot functions in (3.32). Therefore, we also have a unique mapping and set of (3.34) when $\pi < \vartheta_s < 2\pi$.

In the next, we explain how to estimate the source's azimuth given the source's $\boldsymbol{\gamma}(\varphi_s)$ from an unknown direction.

Step 1: The source's $\boldsymbol{\gamma}(\varphi_s)$ cannot distinguish whether the sound source lies in $(0,\pi)$ or $(\pi,2\pi)$ because of the periodic property. However, the real part of $\beta_{1,-1}(k)$ enables to address this issue,

$$\mathrm{Re}\{\beta_{1,-1}(k)\} = -\sqrt{\frac{3}{2}}\sin(\vartheta_s)\sin(\varphi_s) \tag{3.35}$$

whose positive or negative property only depends on the φ_s because $\sin(\vartheta_s) > 0$. Hence, we claim the estimated $\overline{\varphi}_s$,

$$\begin{cases} 0 < \overline{\varphi}_s \leq \pi, & \text{if } \mathrm{Re}\{\overline{\beta}_{1,-1}\} \leq 0 \\ \pi < \overline{\varphi}_s < 2\pi, & \text{if } \mathrm{Re}\{\overline{\beta}_{1,-1}\} > 0. \end{cases} \tag{3.36}$$

Step 2: Exploiting the unique mapping between $(0, \pi)$ or $(\pi, 2\pi)$, we also adopt a distance-based metric using the set of $\mathcal{H}_{\tan}$ to recover its φ_s.

The Second Stage: Elevation Estimation

With the source azimuth given by the first stage, we can recover the source ϑ_s by searching over all possible source elevations,

$$\arg\min_{(\vartheta_s)} \sum_{n=0}^{N} \sum_{m=-n}^{n} \mid \overline{\beta}_{nm} - 2\sqrt{\pi} i^n Y_{nm}^*(\vartheta_s, \overline{\varphi}_s) \mid^2 . \tag{3.37}$$

3.5 Application to Acoustic Enhancement

The above section addressed the estimation of relative harmonic coefficients as well as the source DOA in noisy environments. Here, we show that the proposed method is applicable for a spherical harmonics domain enhancement approach in noisy environments. According to the definition of relative harmonic coefficients, the spherical harmonic coefficients can be represented as a multiplication of relative harmonic coefficients and the received signal at the origin of the array (call as received signal). Hence, we can achieve the spherical harmonics coefficients estimations using the following three steps. Firstly, we estimate the relative harmonic coefficients using the above estimators. Secondly, we use a beamformer to estimate the received signal. Finally, we recover the original spherical harmonic coefficients by multiplying the estimated relative harmonic coefficients and received signal.

In this subsection, we apply the estimated source feature and source DOA for acoustic signal enhancement in spherical harmonics domain. Let us review the

noisy spherical harmonic coefficients in (3.12),

$$\overline{\alpha}_{nm}(k) = \beta_{nm}(k)\alpha_{00}(k) + \gamma_{nm}(k) \tag{3.38}$$

where

$$\beta_{nm}(k) = \frac{\alpha_{nm}(k)}{\alpha_{00}(k)} \tag{3.39}$$

denotes the relative harmonic coefficients defined in [1, 150]. This developed enhancement approach aims to recover the clean spherical harmonic coefficients of $\alpha_{nm}(k)$ from the noisy spherical harmonic coefficients of $\overline{\alpha}_{nm}(k)$. The acoustic model in (3.38) implies the accurate estimation of $\alpha_{nm}(k)$ can be divided into estimations of $\beta_{nm}(k)$ and $\alpha_{00}(k)$ in noisy environments, respectively. Since estimations of the relative harmonic coefficients of $\beta_{nm}(k)$ have already been addressed, we focus on the issue to estimate the clean $\alpha_{00}(k)$ using the steps explained in the following subsection.

3.5.1 Estimation of the Received Signal at the Origin

This subsection estimates the received signal of $\alpha_{00}(k)$ in (3.38) by steering a beamformer. Since the aperture of the recording area is much smaller when compared to its distance to the sound source, we use a method called as the maximum directivity beamformer toward to the far-field sound source [131, 152, 153],

$$S(k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \frac{i^{-n}}{(N+1)^2} Y_{nm}(\vartheta_{\text{est}}, \varphi_{\text{est}})\alpha_{nm}(k) \tag{3.40}$$

where $(\vartheta_{\text{est}}, \varphi_{\text{est}})$ denotes the estimated source's DOA using the localization approaches developed by this chapter. The spherical harmonic coefficient of $\alpha_{00}(k)$ is equivalent to the received signal estimated in (3.40) [25],

$$\alpha_{00}(k) = S(k). \tag{3.41}$$

3.5.2 Spherical Harmonic Coefficients Estimation

Multiplying the estimated relative harmonic coefficients $\beta_{nm}(k)$ and $\alpha_{00}(k)$ in (3.41), we recover the enhanced spherical harmonic coefficients as,

$$\alpha_{nm}(k) = \beta_{nm}(k)\alpha_{00}(k). \tag{3.42}$$

Note that, at the k-th frequency bin, the $\beta_{nm}(k)$ is fixed, while the $\alpha_{nm}(k)$ is updated by the dynamic $\alpha_{00}(k)$ over the time-varying source signal. When the $[\alpha_{00}(k), \cdots, \alpha_{NN}(k)]$ over the entire STFT bins are estimated, we can then reconstruct spectrogram of any individual microphone on the array, using the spherical harmonics representation in (3.2).

3.6 Simulations

This section uses extensive simulations to validate the proposed approaches. We conduct the evaluations in a reverberant room, whose size is $6 \times 4 \times 3$ m for the length, width and height, respectively. We record the incoming soundfield using an open-sphere spherical microphone array (with 32 channels and a radius of 4.2 cm). We use an available toolbox,[1] that implements the image source method [154], to generate the room impulse response (RIR) from the sound source to the microphone array. Speech signal randomly selected from the TIMIT database at the sampling frequency of 8 KHz is used as the input source signal. The original DOA estimation in (3.23) is taken as the baseline for comparisons. The source feature set of $\mathcal{H}$, $\mathcal{H}_{\text{mag}}$, and $\mathcal{H}_{\tan}$ are computed by sampling both the elevation and azimuth with a resolution of 2 degrees. The algorithms use the frequency bins ranging from 1600 Hz to 2400 Hz, recording the soundfield up to the 2nd order ($N = \lceil kr \rceil$) for the DOA estimations, so that the vector's dimension is 9. Lower frequency bins reduce the uniqueness of the feature set as the vector's dimension is reduced to 4, and higher frequency bins contain fewer speech components. The estimators given in (3.16) or (3.10) are used to estimate the source features in the noisy environments, which generally achieve equivalent accuracy of the estimations.

[1] https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator

3.6.1 DOA Estimations For a Static Sound Source

Performance of our system is measured using the mean absolute estimated error (MAEE/°) between the estimated and original DOA,

$$\text{MAEE} = \frac{1}{2}\left(|\vartheta_{\text{ori}} - \vartheta_{\text{est}}| + |\varphi_{\text{ori}} - \varphi_{\text{est}}|\right). \tag{3.43}$$

We first examine the performance of the methods using a static sound source somewhere in the room. Table 3.2 presents the localization errors using the proposed algorithms in diverse reverberation levels. The increased reverberation level implies more negative impacts caused by the coherent reverberations so that the localization accuracy degrades. Table 3.3 exhibits the localization errors under various noisy conditions, where the SNR level ranges from 5 dB to 25 dB. Since the feature estimator has already taken the noise into account, we recognize little degraded performance when the SNR level decreases. Examinations of the proposed methods in both Table 3.2 and 3.3 confirm that, although our proposed methods only use a 1-D searching, they still achieve competitive localization accuracy in comparison with the baseline which uses a 2-D searching.

This chapter emphasizes the speed of the proposed methods. The computational complexity of the decoupled DOA estimator is $O\left((N_\theta + N_\phi)(N+1)^2\right)$, where N_θ and N_ϕ denote the sampled directions over the elevation and azimuth, respectively. By contrast, the baseline DOA estimator has much larger computational cost, because it requires a two-dimensional search, i.e., $O\left(N_\theta N_\phi (N+1)^2\right)$. For validations, we measure the computational complexity of the algorithms by directly recording the time cost, using a Matlab implementation on a standard desktop (CPU Intel Core i7-4790 Quad 3.6 GHz, RAM 16 GB). Table 3.4 presents the speed of all the algorithms. We observe that the proposed methods only take less than 0.4 ms to search the source DOA over the 2-D direction space, achieving dramatically improved speed compared to the baseline approach.

3.6.2 DOA Tracking For a Moving Sound Source

Motivated by the reduced complexity, the remained content applies the proposed approaches into sound source tracking. The moving source's time-domain record-

Table 3.2: Source DOA estimation under various reverberation levels where the SNR is 25 dB.

MAEE/°	T_{60} (s)				
Methods	0.2	0.3	0.4	0.5	0.6
Baseline	1.01	1.33	1.74	1.93	2.31
Proposed1	1.00	1.38	1.76	1.91	2.34
Proposed2	1.05	1.90	2.48	3.28	4.05

Table 3.3: Source DOA estimation under various SNR levels where $T_{60} = 0.4$ s.

MAEE/°	SNR level (dB)				
Methods	5	10	15	20	25
Baseline	1.87	1.64	1.82	1.64	1.74
Proposed1	1.96	1.71	1.82	1.69	1.76
Proposed2	2.84	2.66	2.32	2.40	2.48

Table 3.4: Time cost by 20 executions, when searching the DOA.

Methods	Baseline	Proposed1	Proposed2
Time cost (ms)	309.5	7.9	8.1

ings are generated using an available toolbox in [95]. We use the same simulated room, whose six wall surface acoustic absorption coefficients are set at 0.9. The SNR level is 15 dB. We split the measured recordings into the frames lasting 0.5 s, and use the algorithms to estimate the source DOA over the instantaneous frames. For a smoother path, we synthesize the source trajectory using two successive estimations,

$$\overline{\Phi}(t) = w\Phi(t) + (1-w)\overline{\Phi}(t-1) \tag{3.44}$$

where t is the index of estimations, $\Phi(t)$ is the current estimated DOA of $(\vartheta_s^t, \varphi_s^t)$, and w denotes its weight (set at 0.7 in the simulation). Figure 3.3 and 3.4 exhibit the estimated source trajectory using the two proposed methods, respectively. Both the randomly generated trajectories by the moving source have been recovered accurately within a fast response time, which only takes about 0.1 s to process

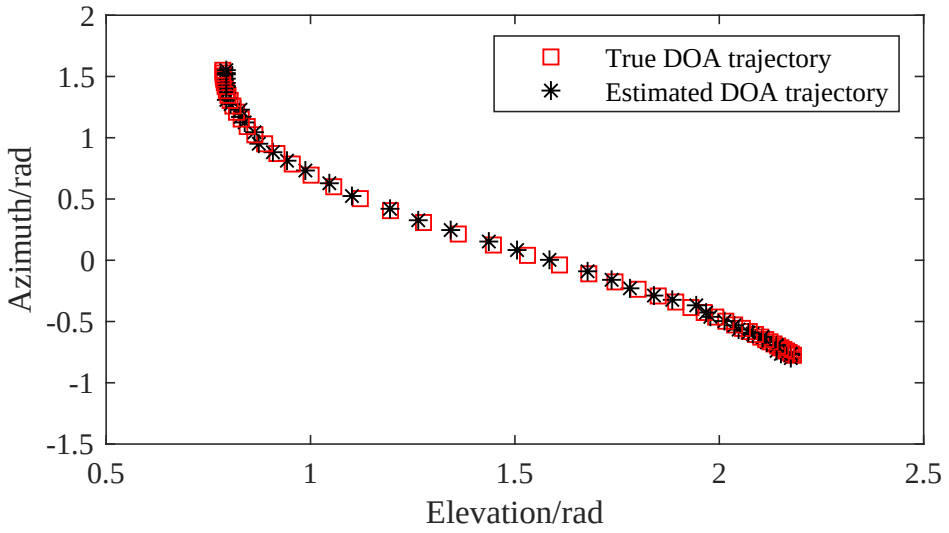


Figure 3.3: Original and estimated source trajectory using the first method.

each instantaneous frame.

3.6.3 Application in Acoustic Enhancement

Accuracy of the signal estimations is measured by the normalized mean squared error (NMSE/dB) of the original and estimated signal in noisy environments. We first define the metric for the relative harmonic coefficients,

$$\text{NMSE}_{\boldsymbol{\beta}} = 10\log_{10}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{||\boldsymbol{\beta}(k)-\overline{\boldsymbol{\beta}}(k)||_2^2}{||\boldsymbol{\beta}(k)||_2^2}\right) \tag{3.45}$$

where $||\cdot||_2$ denotes the ℓ_2 norm, $\boldsymbol{\beta}(k)$ and $\overline{\boldsymbol{\beta}}(k)$ denote the clean and estimated relative harmonic coefficients. Next, the metric for the received signal in the STFT domain is,

$$\text{NMSE}_{\boldsymbol{\alpha}_{00}} = 10\log_{10}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{||\boldsymbol{\alpha}_{00}(k)-\overline{\boldsymbol{\alpha}}_{00}(k)||_2^2}{||\boldsymbol{\alpha}_{00}(k)||_2^2}\right) \tag{3.46}$$

where $|\cdot|$ denotes the absolute operator, $\boldsymbol{\alpha}_{00}(k)$ and $\overline{\boldsymbol{\alpha}}_{00}(k)$ denote the vector of clean and estimated received signal over the time-varying bins, respectively. Finally,

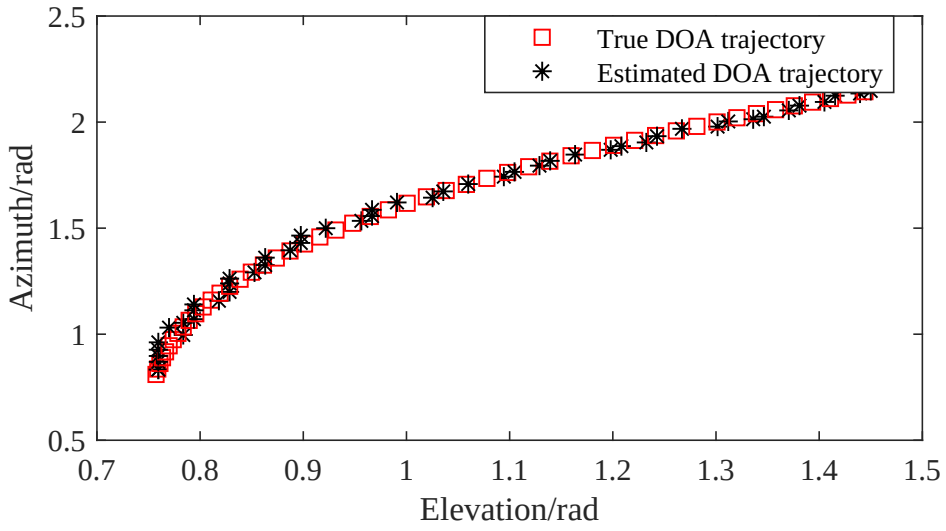


Figure 3.4: Original and estimated source trajectory using the second method.

the metric for spherical harmonics coefficients is,

$$\mathrm{NMSE}_{\boldsymbol{\alpha}} = 10\log_{10}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{||\boldsymbol{\alpha}(k)-\overline{\boldsymbol{\alpha}}(k)||_2^2}{||\boldsymbol{\alpha}(k)||_2^2}\right) \tag{3.47}$$

where $\boldsymbol{\alpha}(k)$ and $\overline{\boldsymbol{\alpha}}(k)$ denote the clean and estimated spherical harmonics coefficients, respectively.

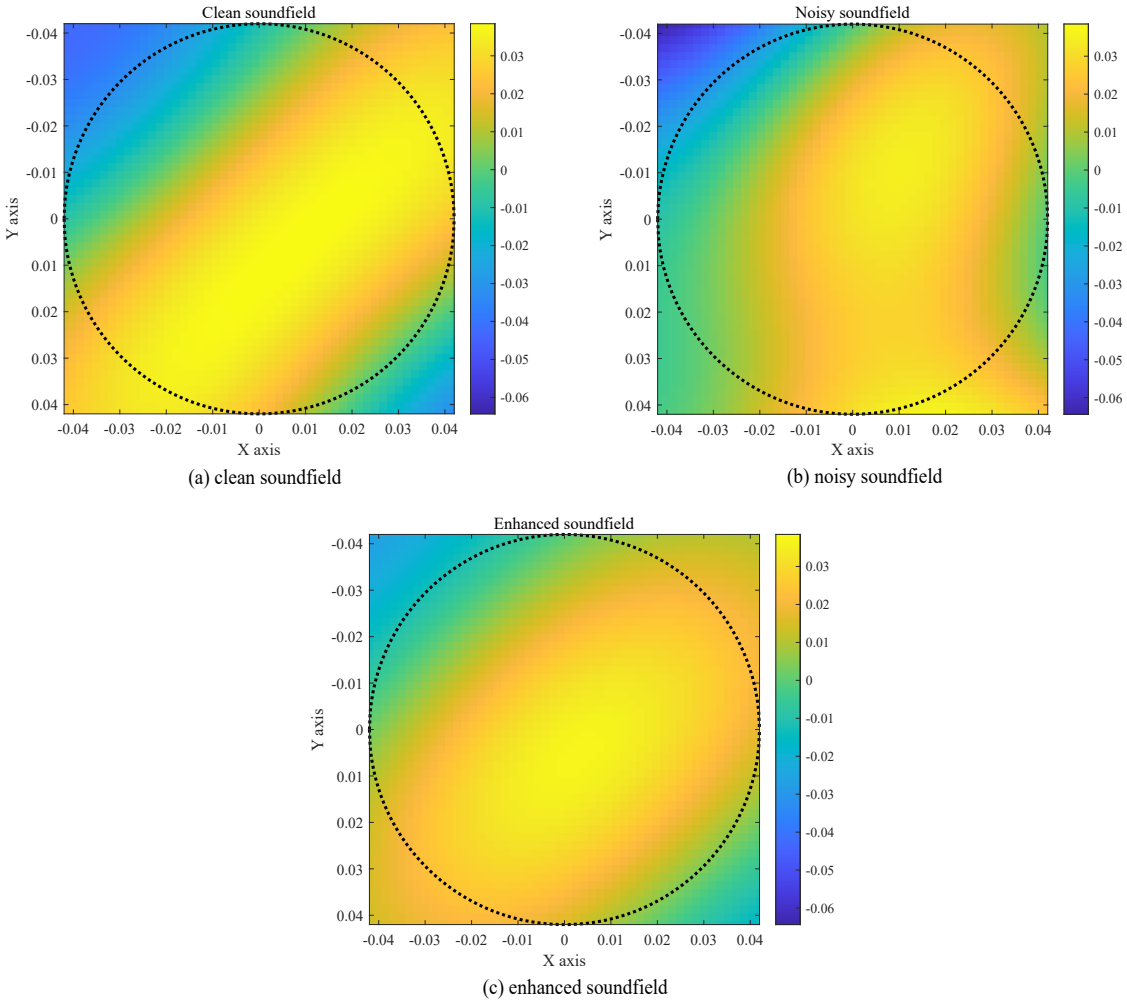
Table 3.5: Accuracy of estimations at various SNR levels, including relative harmonic coefficients, DOA and received signal.

SNR level (/dB)	25	20	15	10	5
$\mathrm{NMSE}_{\boldsymbol{\beta}}$ (dB)	-20.3	-15.0	-9.5	-5.5	-2.65
$\mathrm{NMSE}_{\boldsymbol{\alpha}_{00}}$ (/dB)	-15.1	-11.6	-7.9	-4.7	-2.2

To achieve consistent results, we use ten groups of sound sources, each propagating from a randomly unknown DOA. Thus, the values of both NMSE and MAEE, presented in the following, denote the mean value over all the cases. We evaluate the proposed method under various SNR conditions ranging from 5 dB to 25 dB. Table 3.5 depicts the accuracy of the estimations for relative harmonic

coefficients and received signal, respectively. We then multiply the estimated relative harmonic coefficients and received signal to obtain the spherical harmonic coefficients. Table 3.6 depicts the NMSE where the original values are taken as the baseline. As expected, we observe that a stronger noisy environment exerts a direct negative influence on the received signal. However, the proposed method improves the NMSE by around 3 dB.

Figure 3.5 exhibits the clean, noisy and enhanced soundfield over the micro-



(a) clean soundfield (b) noisy soundfield

(c) enhanced soundfield

Figure 3.5: (a) clean, (b) noisy and (c) enhanced soundfield over the microphone array when $z = 0$ (3 KHz, 5dB noise).

Table 3.6: Accuracy of the spherical harmonic coefficients estimations under various SNR levels.

$\text{NMSE}_{\boldsymbol{y}}$(/dB)	SNR levels (/dB)				
Method	25	20	15	10	5
Original	-11.9	-7.8	-4.8	-2.6	-1.2
Enhanced	-16.4	-11.8	-7.7	-4.4	-2.2

phone array whose coordinates are $z = 0$, due to the sound source propagating from $(\vartheta_s, \varphi_s) = (0.96, 5.17)$. We recognize that the enhanced soundfield generally gets rid of the distortions caused by the noise signal. Note that Fig. 3.5 shows the soundfield at a single STFT bin. By contrast, Figure 3.6 presents the noisy speech spectrogram in the entire STFT domain and time domain recordings using an ISTFT. Figure 3.7 presents the enhanced signal. Most of the noise is alleviated and the enhanced speech has satisfying intelligibility. While the above results are promising, a limitation of this approach lies in the beamformer in (3.40), which is designed for far-field propagation. Therefore, in near-field soundfield, the estimation performance of the signal in (3.41) will degrade unless an appropriate radial focused near-field beamformer is used.

3.7 Summary

This chapter presented a novel source feature called *relative harmonic coefficients* in the spherical harmonics domain. Its closed-form expression implies that the elevation and azimuth components in the relative harmonic coefficients are decoupled. Hence, two decoupled 2-D source DOA estimators, highlighted by a large reduction of computational complexity, were developed. Evaluations in both single source localization and tracking have confirmed the dramatic reduction in computational complexity while achieving competitive accuracy. This proposed approach is also applicable to an acoustic enhancement approach for cleaning noisy higher-order microphone recordings. It enables us to estimate the spherical harmonic coefficients up to the whole soundfield order. Extensive results using noisy speech measurements of a spherical microphone array confirmed the effective denoising performance.

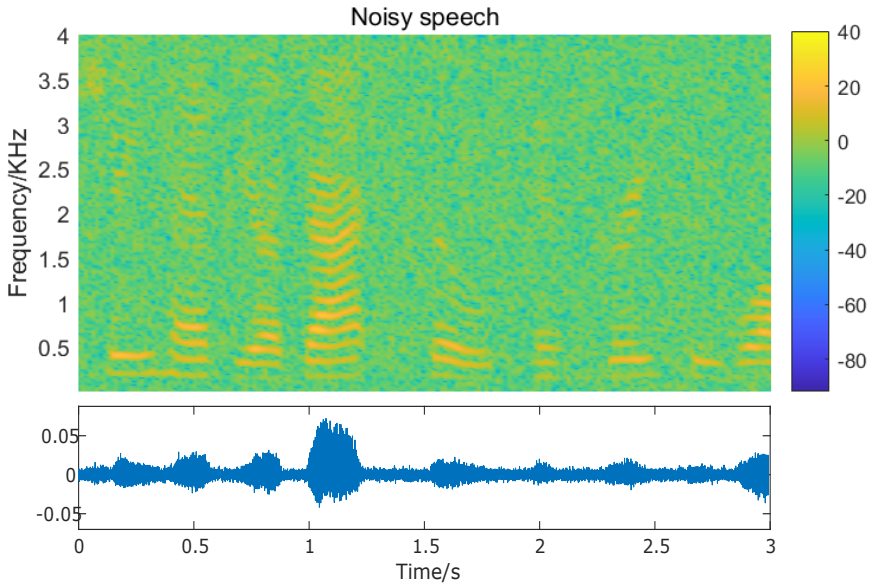


Figure 3.6: Noisy speech signal of the 1st microphone on the spherical microphone array (5dB noise).

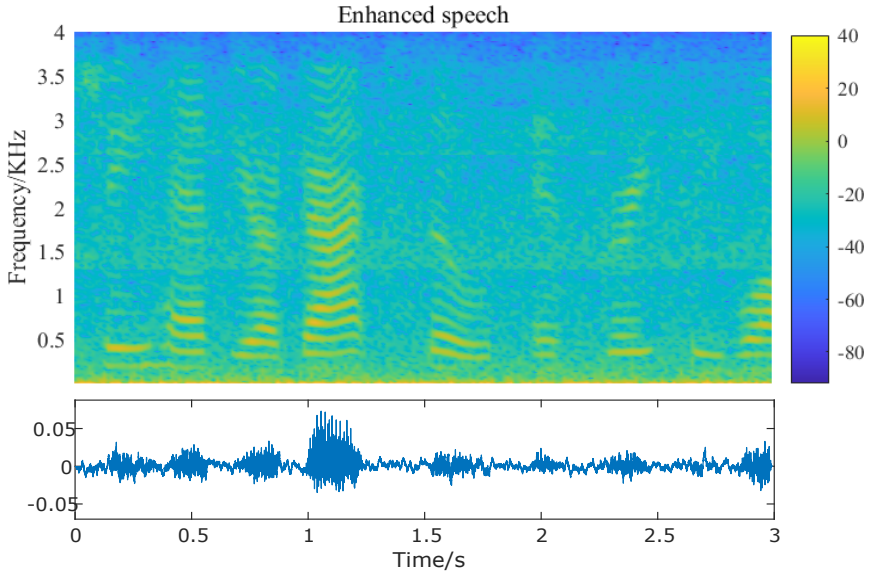


Figure 3.7: Enhanced speech signal of the 1st microphone on the spherical microphone array (5dB noise).

The proposed method is designed for far-field propagation. Hence, the performance of this method degrades severely in the environments where the far-field assumption hardly holds. Another major issue that remained to be solved is the degraded performance in strongly reverberant environments because the acoustic model has not yet taken acoustic reflections into account. To address the above issues, the next chapter of this thesis develops a learning-based localization approach that is suitable in a more complex/dynamic acoustic environment.

3.8 Related Publications

This chapter's work has ever been published in the following journal papers and conference proceedings.

- Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using realtive harmonic coefficients in modal domain", in 2019 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 348-352.
- Y. Hu, T. D. Abhayapala, P. N. Samarasinghe, and S. Gannot, "Decoupled direction-of-arrival estimations using relative harmonic coefficients", in 2020 *IEEE 28th European Signal Processing Conference (EUSIPCO)*, 246-250.
- Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Acoustic signal enhancement using relative harmonic coefficients: spherical harmonics domain approach", in 2020 *INTERSPEECH*, 5076-5080.

Chapter 4

Semi-Supervised Multiple Source Localization Using Relative Harmonic Coefficients Under Noisy and Reverberant Environments

Overview: *this chapter uses the relative harmonic coefficients to develop a semi-supervised algorithm that addresses the challenging multi-source localization problem in a noisy and reverberant environment. We investigate this source feature in reverberant environments by presenting (i) an illustration confirming its sole dependence on the source position in reverberant environments, (ii) a feature selector exploiting its inherent directivity over space. Source features at varied spherical harmonic modes, representing unique characterization of the soundfield, are merged/fused by the Multi-Mode Gaussian Process modeling. Based on the unifying model, we then formulate the mapping function revealing the underlying relationship between the source feature(s) and position(s) using a Bayesian inference approach. Another issue of the overlapped components is addressed by a pre-processing technique performing overlapped frame detection, which in turn reduces this challenging problem to a single source localization issue. It is highlighted that this data-driven*

method has a strong potential to be implemented in practice as only a limited number of labeled measurements is required. The proposed algorithm is evaluated using simulated recordings between multiple speakers in diverse environments, and the results confirm improved performance in comparison with the state-of-art methods. Additional assessments using real-life recordings further prove the effectiveness of the method, even in unfavorable circumstances with severe source overlapping.

4.1 Introduction

In Chapter 3, we developed two decoupled source localization approaches which were highlighted by low computational complexity. However, they are only limited to the single-source scenarios and their localization accuracy degrades severely in a complex acoustic environment, where the recordings are contaminated by the strong multi-path room reverberations as well as the noise with low signal-to-noise ratios. As we have discussed in Section 2.1, data-driven source localization algorithms have been widely used to address the degraded accuracy in the complex environments [22,23,45,57,81,58,80,97]. This chapter aims to address the multiple source localization in a noisy and reverberant environment by proposing a data-driven approach that uses the relative harmonic coefficients as its inputs.

The aforementioned deep learning-based algorithms, such as [23,59,80,97], accomplish source localization by classifying the desired source DOA into one of the candidate directions over the two-dimensional space. By contrast, this chapter intends to adopt a regression scheme, i.e., a Bayesian inference approach of Gaussian Process Regression (GPR) [155], because it suits more to localize the continuous variable of the source's Cartesian coordinates (i.e., x, y, z coordinates). Traditional GPR requires a single Gaussian Process modeling, while this chapter adopts the Multi Gaussian Process modeling [45] to the spherical harmonics domain (called as Multi-Mode Gaussian Process (MMGP)), to merge/fuse the relative harmonic coefficients over the varied spherical harmonic modes. Data-driven source localization is often criticized as a cumbersome task because it requires a large training set. To overcome the drawback, we are adopting the semi-supervised paradigm, previously used in [45,57,58], where only a small number of labeled samples is required. However, [45,57,58] only addressed the single-source scenario. Multiple

source localization becomes much more challenging because the overlapped components, especially significantly overlapped recordings, hinder an accurate localization of the sources. Recent studies [149, 90] addressed this issue using a pre-processing tool to detect and isolate the overlapped components. Motivated by this strategy, this chapter simplifies the challenging multi-source localization into a single source localization problem by developing a newly overlapped frame detector using the relative harmonic coefficients.

In comparison with Chapter 3, novel contributions by this chapter are briefly summarized as follows: (i) we study a semi-supervised multi-source localization approach, which only uses a small number of labeled training samples; (ii) we present a theoretical proof confirming that the relative harmonic coefficients only depend on its source position in reverberant environments; (iii) we develop a metric selecting the spherical harmonic modes that suit for source localization within a given area; and (iv) we provide a data-driven overlapped frame detection. The remaining part of the chapter is structured as follows. We first formulate the problem addressed in this chapter. Section 4.3 presents the source feature selector exploiting its inherent directivity. Section 4.4 derives the mapping function that merges/fuses the selected source features. Section 4.5 summarizes the block-diagram of the algorithm and explains the data-driven overlapped frame detection. Thereafter, extensive experimental results are reported in Section 4.6. Finally, conclusions are drawn and discussions are given in Section 4.7.

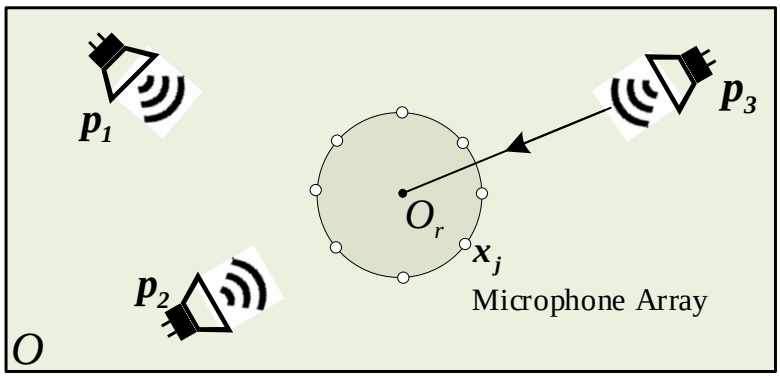


Figure 4.1: Multiple source localization using a higher-order microphone array in a noisy and reverberant environment (top view).

4.2 System Model

This section first briefly describes the challenging localization problem to be addressed. Then, we prove that the source feature of relative harmonic coefficients only depends on the source position in reverberant environments.

4.2.1 Problem Formulation

Let there be Q active sound sources inside the reverberant room (e.g., see Figure 4.1), whose Cartesian coordinates are $\boldsymbol{p}_q = [x_q, y_q, z_q]^T$ $(q = 1, \cdots, Q)$ with respect to the room origin of $O = [0, 0, 0]^T$. Consider a higher-order microphone array with M microphones that are located at $\boldsymbol{x}_j$ $(j = 1, \cdots, M)$ with respect to the array origin O_r. The sound pressure, measured by the j-th microphone of the array at the k-th frequency bin, is represented by:

$$\begin{aligned} \overline{P}(\boldsymbol{x}_j, k) &= P(\boldsymbol{x}_j, k) + V(\boldsymbol{x}_j, k), \quad j = 1, \cdots, M \\ &= \sum_{q=1}^{Q} S_q(k) A_q(\boldsymbol{x}_j, k) + V(\boldsymbol{x}_j, k) \end{aligned} \tag{4.1}$$

where $k = 2\pi f/c$ is the wavenumber, f is the frequency bin, c is the speed of sound, $S_q(k)$ is the q-th source signal, $A_q(\boldsymbol{x}_j, k)$ denotes the acoustic transfer function (ATF) from the q-th sound source to the j-th microphone, $P(\boldsymbol{x}_j, k)$ and $\overline{P}(\boldsymbol{x}_j, k)$ denote the clean and noisy sound pressure and $V(\boldsymbol{x}_j, k)$ represents the additive noise signal at the j-th microphone. Given the multi-source recordings of $\overline{P}(\boldsymbol{x}_j, k)$, we aim to accurately recover the positions of the sound sources, i.e., $\boldsymbol{p}_q$ where $q = 1, \cdots, Q$. In addition, we have $\mathcal{N}_D = \mathcal{N}_L + \mathcal{N}_U$ measurements in advance within a predefined source area of interest, consisting of $\mathcal{N}_L$ labeled samples whose known positions are $\mathbf{p} = \{\boldsymbol{p}_1, \cdots, \boldsymbol{p}_{\mathcal{N}_L}\}$, and $\mathcal{N}_U$ unlabeled samples randomly located at unknown positions. Note that the additive noise in (4.1) is assumed to be non-directional; otherwise, the directional noise could be treated as additional sources to be localized.

We intend to achieve our goal by solving two nontrivial issues: (i) developing an overlapped frame detector to discover the single-source components from the mixed

measurements so that the localization of multi-source is simplified into repetitive single source localizations; (ii) exploiting the source feature of relative harmonic coefficients to realize a data-driven single source localization in complex environments. Before showing the localization algorithm, we first show that the relative harmonic coefficients only depend on its source position in reverberant environments so that it can be used as the source feature to localize the sound source(s) in this environment.

4.2.2 Illustration of the Source Feature in Reverberant Environments

This subsection illustrates the composition of relative harmonic coefficients by deriving its theoretical expression in reverberant environments (both the arbitrary and rectangle reverberant rooms are used), which confirm to be only dependent on its source position.

Arbitrary reverberant soundfield

Assume a reverberant soundfield produced by the q-th sound source in the soundfield in Fig. 4.1. The spherical harmonic coefficient over the recording area is represented as,

$$\alpha_{nm}^{\text{rev}}(k) = \alpha_{nm}^{\text{dir}}(k) + \underbrace{\sum_{v=0}^{N}\sum_{u=-v}^{v} \widehat{\alpha}_{nm}^{vu}(k) S_q(k) ikj_v(kr_q) Y_{vu}^*(\theta_q, \phi_q)}_{\text{Reverberant-path}} \tag{4.2}$$

where $\widehat{\alpha}_{nm}^{vu}(k)$ is the coupling coefficient that is independent of the time-varying source signal [88]. Note that (4.2) considers an arbitrary acoustic environment so that the coupling coefficients have no explicit expression. Following the feature definition in (3.3), we have the corresponding relative harmonic coefficients,

$$\beta_{nm}^{\text{rev}}(k) = \frac{h_n(kr_q)Y_{nm}^*(\theta_q, \phi_q) + \sum_{v=0}^{N}\sum_{u=-v}^{v} \widehat{\alpha}_{nm}^{vu}(k) j_v(kr_q) Y_{vu}^*(\theta_q, \phi_q)}{h_0(kr_q)Y_{00}^*(\theta_q, \phi_q) + \sum_{v=0}^{N}\sum_{u=-v}^{v} \widehat{\alpha}_{00}^{vu}(k) j_v(kr_q) Y_{vu}^*(\theta_q, \phi_q)} \tag{4.3}$$

which only depends on the source position in a static acoustic environment where the settings of the environment and microphone array are assumed to remain fixed.

Rectangle reverberant soundfield

Above expression of (4.3) describes the relative harmonic coefficient given an arbitrary reverberant soundfield where the coupling coefficients have no explicit expression. Let us investigate it in a specific rectangle room characterized by the generalized image source method in the spherical harmonics domain [89].

Assuming a single sound source located at $\boldsymbol{p}_q = (x_q, y_q, z_q)$ in a reverberant room whose dimensions are (L_x, L_y, L_z) for length, width and height respectively, we claim that the corresponding relative harmonic coefficient over the recording area is represented as,

$$\begin{aligned}\beta_{nm}^{\text{rev}}(k) = \rho_o \sum_{\boldsymbol{p}=0}^{1} \sum_{\boldsymbol{r}=-\infty}^{\infty} \lambda_{x1}^{|r1-q|} \lambda_{x2}^{|r1|} \lambda_{y1}^{|r2-j|} \lambda_{y2}^{|r2|} \lambda_{z1}^{|r3-\ell|} \lambda_{z2}^{|r3|} \\ \times (-1)^{(j+\ell m)+\ell n} S_n^m(\boldsymbol{R}_p + \boldsymbol{R}_r)\end{aligned} \tag{4.4}$$

where $\boldsymbol{p} = (q, j, \ell)$ and $\boldsymbol{r} = (r_1, r_2, r_3)$ are triplet parameters controlling the indexing of the image sources in all dimensions, $\boldsymbol{R}_p = (x_r - x_q + 2qx_q, y_r - y_q + 2jy_q, z_r - z_q + 2\ell z_q)$, $\boldsymbol{R}_r = (2r_1L_x, 2r_2L_y, 2r_3L_z)$, (x_r, y_r, z_r) denotes the position of the origin of receiver area, and $\lambda_{x,i}, \lambda_{y,i}, \lambda_{z,i}$ with $i = 1, 2$, represent the wall reflection coefficients. In (4.4), the symbol of ρ_o denotes a fixed scalar that is determined by the setup of the room, and $S_n^m(\cdot)$ only depends on the positions of the sound source and microphone array. Detailed procedures to derive (4.4) are given in the Appendix A. In a static acoustic environment, (4.4) implies that the feature in a reverberant rectangle room also only depends on the source position.

4.3 Source Feature Selector

This section first shows that the proposed spherical harmonic domain feature has a unique directivity pattern over space. Thus, we then develop a quantitative metric to select a subset of the spherical harmonic modes that are suitable for source localization within a limited-size source area of interest.

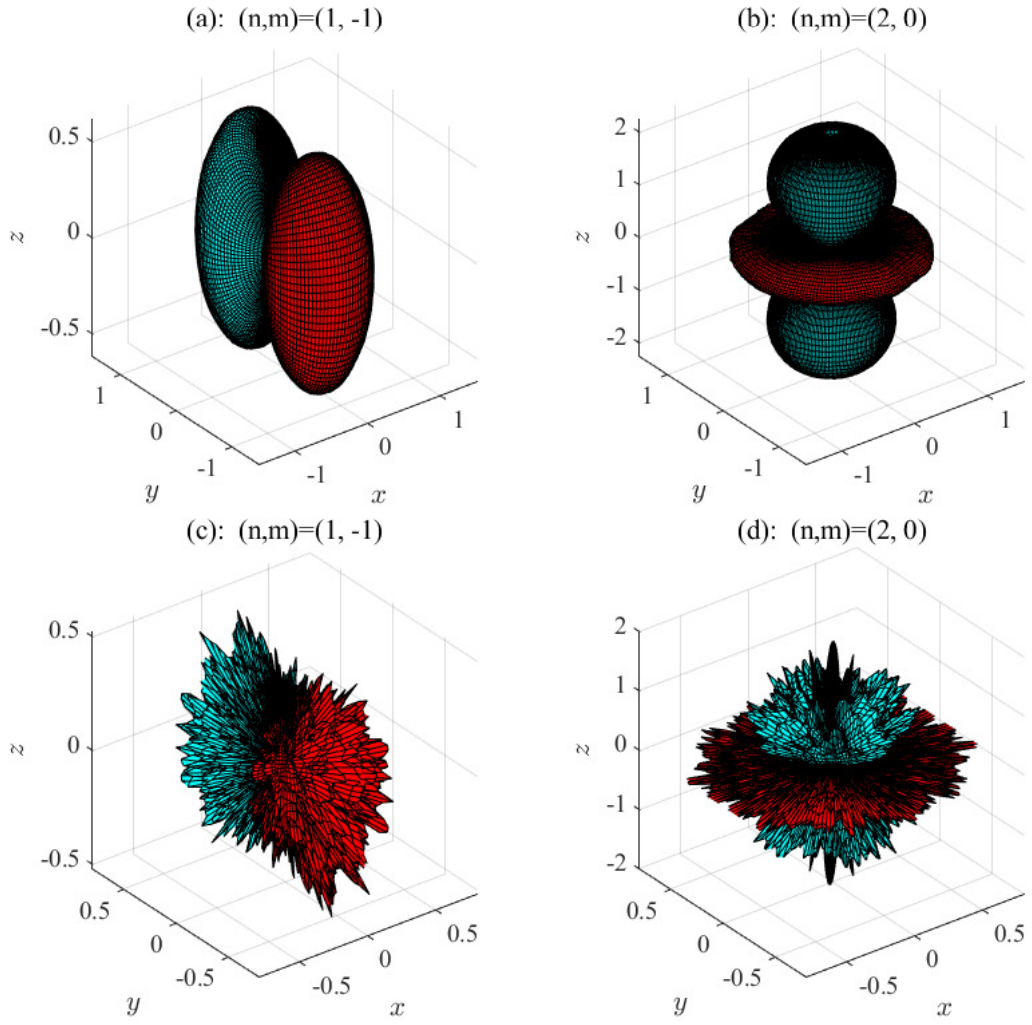


Figure 4.2: Real part of the source features at the spherical harmonic modes of $(1,-1),(2,0)$ respectively. The sub-figures (a)-(b) denote the source features using direct-path recordings where there exists almost no room reverberation. By contrast, sub-figures (c)-(d) denote the reverberant features whose $T_{60} = 500$ ms with a room reflection order of ten.

4.3.1 Directivity Pattern Analysis

The defined source feature of relative harmonic coefficients has a unique directivity pattern over the space due to its direct relation with the spherical harmonic function

(e.g., see both (4.3) and (4.4)). Figure 4.2 exhibits an example of the source features at the spherical harmonic modes of $(1, -1)$ and $(2, 0)$ respectively, each representing a distinct characterization/description of the soundfield. The red and cyan portions represent regions where the real parts of the features are positive and negative, respectively. The distance of the surface from the origin indicates the absolute value of the features in angular direction over space. We estimate the source features using the simulated recordings in a $6 \times 4 \times 3$ m room. We use a set of sound sources located on a spherical shell with respect to a spherical microphone array at the origin of the shell, i.e., $\mathbf{\Phi}^2 = \{(r, \vartheta_s, \varphi_s) : r = 1, 0 < \vartheta_s \leq \pi, 0 < \varphi_s \leq 2\pi\}$. Twenty frequency bins approximately ranging from 1500 Hz to 2500 Hz are used, which records the soundfield up to the 2nd order. Note that the presented figures denote the mean values over this wide frequency band. It is observed that the source features in a reverberant environment are less smoothly distributed over space, due to the random interfering signal of the room reverberations.

Generally, a unique directivity pattern assists in distinguishing the sound sources located at the area where the source features have a large difference between each other (i.e., active area). By contrast, the source features at any given harmonic mode also have an inactive area where they have little differences (i.e., the directivity pattern is not significant within some areas). The followings are some examples: (i) $(n, m) = (0, 0)$: the source features equal to 1 wherever the source locates; (ii) $(n, m) = (1, -1)$: the source features are close to zero for the sources located around the plate where $y = 0$; (iii) $(n, m) = (2, 0)$: when the sources are located on the horizontal plate where $z = 0$, their features are similar.

In practice, a data-driven source localization method generally implements within a limited region predefined in advance. Given the estimated source features up to the N-th order, we expect to select a subset of the spherical harmonic modes whose active area covers the source area for localization. As explained in the next subsection, we achieve the spherical harmonic modes selection by proposing a statistical metric based on the training set of source features.

4.3.2 Spherical Harmonic Modes Selector using the Training Set

Assume the coordinates of the predefined sound source area for localization are, e.g.,

$$\mathbf{\Phi} = \left\{(r, \vartheta_s, \varphi_s) : r \geqslant r_o, \vartheta_a < \vartheta_s \leq \vartheta_b, \varphi_c < \varphi_s \leq \varphi_d\right\} \tag{4.5}$$

in which $r_o, \vartheta_a, \vartheta_b, \varphi_c, \varphi_d$ are some constants. Consider $\mathcal{N}_D$ training samples distributed within this source area have been recorded and the corresponding training source features are estimated. We then construct a vector by collecting the relative harmonics coefficients at the mode of (n, m),

$$\left[\beta_{nm}^1(k), \beta_{nm}^2(k), \cdots, \beta_{nm}^{\mathcal{N}_D}(k)\right]^T. \tag{4.6}$$

As analyzed, the sound sources within the active area appear with varied values, i.e., the source features distribute more decentralized over the source area. For a quantitative measurement, we exploit the index of dispersion (i.e., Variance to Mean Ratio) [156] with respect to the vector of (4.6),

$$d_{nm}(k) = \left|\frac{\sigma_{nm}^2(k)}{\mu_{nm}(k)}\right| \tag{4.7}$$

where

$$\begin{aligned} \mu_{nm}(k) &= \frac{1}{\mathcal{N}_D}\sum_{\ell=1}^{\mathcal{N}_D}\beta_{nm}^{\ell}(k) \\ \sigma_{nm}^2(k) &= \frac{1}{\mathcal{N}_D - 1}\sum_{\ell=1}^{\mathcal{N}_D}|\beta_{nm}^{\ell}(k) - \mu_{nm}(k)|^2 \end{aligned} \tag{4.8}$$

denote the mean and variance of the elements in (4.6), respectively, and $\ell \leq \mathcal{N}_D$ denotes the index number. Note that above calculation only uses the source feature at the k-th frequency bin. At the case of a wide frequency band (e.g., F frequency

bins), we then compute the mean number as,

$$\overline{d}_{nm} = \frac{1}{F}\sum_{i=1}^{F} d_{nm}(k_i). \tag{4.9}$$

The measure of dispersion is successively applied for all the $(N+1)^2$ spherical harmonic modes to produce a vector as,

$$\left[\overline{d}_{00}, \overline{d}_{1,-1}, \cdots, \overline{d}_{NN}\right]^T. \tag{4.10}$$

Intuitively, we select the spherical harmonic modes exhibiting with a larger index of dispersion, i.e., the source features have more differences when the sources are located differently,

$$\overline{d}_{nm} > \zeta \tag{4.11}$$

where ζ is a positive threshold empirically specified as long as it performs with sufficient localization accuracy. For example, $\overline{d}_{00} = 0$, the source features at the spherical harmonic mode of $(0, 0)$ are discarded.

Up to now, we have estimated the training source features and selected a subset of the spherical harmonic modes with an improved validity to localize the sources within the defined area. In the next section, we show how to use the training samples to formulate a mapping function that recovers the unknown position of a given testing source.

4.4 Mapping Function Formulation for Data-driven Single Source Localization

This section develops a data-driven single source localization approach by formulating the mapping function that reveals the underlying relation between the source feature(s) and source position(s). We first use the Multi-Mode Gaussian Process (MMGP) to model the variable of source position, fusing/merging the features at the selected spherical harmonic modes. Then, we use the MMGP based Gaussian Process Regression (GPR) to recover the unknown source position. Note that

the proposed GPR based source localization approach localizes the source x, y, z-coordinate separately because the Gaussian Process modeling mainly applies into scalar variable [155]. Hence, the source position variable p used in this section denotes a scalar of p_x, p_y or p_z. Finally, we claim in advance that the underlying theory discussed in this section is a direct inspiration and adaptation of a recently proposed method in [45]. The original method exploits the ReTFs for the mapping function formulation while this section uses the RHCs defined in the spherical harmonic domain.

4.4.1 Multi-Mode Gaussian Process (MMGP)

Assume an arbitrary sound source whose feature matrix is $\boldsymbol{B} \in \mathbb{C}^{F \times V}$ where $V \leq (N+1)^2$ denotes the number of the selected spherical harmonic modes. Using a single feature vector at the v-th mode, we model the variable of its source position by a zero-mean Gaussian Process,

$$p^{(v)}(\boldsymbol{\beta}) \sim \mathcal{GP}(0, \mathcal{K}) \tag{4.12}$$

where p^v denotes the source position variable at the v-th mode, $\boldsymbol{\beta} \in \mathbb{C}^{F \times 1}$ denotes the feature vector containing all the F frequency bins at this mode, $\mathcal{K}$ denotes the kernel or covariance function that specifies the Gaussian Process. We adopt the manifold-based covariance function [45], where the relationship between two sources is not only a function of the current two samples but also exploits the information of the entire training set,

$$\mathrm{cov}(p_{n_i}^{(v)}, p_{n_j}^{(v)}) \equiv \sum_{n_\ell=1}^{\mathcal{N}_D} \mathcal{K}(\boldsymbol{\beta}_{n_i}, \boldsymbol{\beta}_{n_\ell}) \mathcal{K}(\boldsymbol{\beta}_{n_j}, \boldsymbol{\beta}_{n_\ell}) \tag{4.13}$$

where subscript of n_i and n_j denotes the index of two arbitrary sources, n_ℓ is the index of the training sources, and $\mathcal{K}(\cdot)$ is the kernel function between any pair of features. Theoretically, a series of kernel functions is applicable as long as its covariance matrix is positive semi-definite, and symmetric [155]. We use the

squared exponential (SE) covariance function as,

$$\mathcal{K}(\boldsymbol{\beta}_{n_i}, \boldsymbol{\beta}_{n_\ell}) = \exp\left(-\frac{\|\boldsymbol{\beta}_{n_i} - \boldsymbol{\beta}_{n_\ell}\|^2}{2\sigma_y^2}\right), 1 \leqslant n_i, n_\ell \leqslant \mathcal{N}_D \tag{4.14}$$

where $\|\cdot\|$ represents the Euclidean ℓ_2 norm, and σ_y denotes the characteristic length-scale hyperparameter that is initialized with a random value and then optimized using the marginal likelihood [155].

Note that the Gaussian Process modeling above only uses a single feature vector at the v-th spherical harmonic mode. By contrast, the MMGP fuses all the source features by modeling source position p as the mean of the Gaussian Processes between all the V spherical harmonic modes, i.e., the n_i-th source final position p_{n_i} equals to the average value of all the estimations,

$$p_{n_i} = \frac{1}{V}\left(p_{n_i}^{(1)} + p_{n_i}^{(2)} + \cdots + p_{n_i}^{(V)}\right). \tag{4.15}$$

We emphasize the difference between the Multi-Node Gaussian Process in [45] that fused recordings from the distributed microphone pairs and our proposed method in which we fuse the features of different spherical harmonic modes given by a higher-order microphone array[1]. Due to the assumption that the processes are jointly Gaussian, p also follows a zero-mean Gaussian Process, whose covariance between two arbitrary source positions is computed as,

$$\begin{aligned}\operatorname{cov}(p_{n_i}, p_{n_j}) &= \overline{\mathcal{K}}(\boldsymbol{B}_{n_i}, \boldsymbol{B}_{n_j}) \\ &= \frac{1}{V^2}\operatorname{cov}\left(\sum_{z=1}^{V} p_{n_i}^{(z)}, \sum_{w=1}^{V} p_{n_j}^{(w)}\right) \\ &= \frac{1}{V^2}\sum_{z,w=1}^{V} \operatorname{cov}(p_{n_i}^{(z)}, p_{n_j}^{(w)})\end{aligned} \tag{4.16}$$

in which $\overline{\mathcal{K}}(\cdot)$ denotes the kernel function of the MMGP, $\boldsymbol{B}_{n_i}$ and $\boldsymbol{B}_{n_j}$ are the feature matrix containing all the V modes, and z and w are the index of spherical harmonic mode. This paper defines the covariance of variables between two

[1]Please note that the Multi-Mode Gaussian Process refers to the proposed method by this paper while the Multi-Node Gaussian Process refers to the method in [45].

different modes as,

$$\text{cov}(p_{n_i}^{(z)}, p_{n_j}^{(w)}) \equiv \text{cov}(p_{n_i}^{(z)}, p_{n_j}^{(z)})\text{cov}(p_{n_i}^{(w)}, p_{n_j}^{(w)}) \tag{4.17}$$

where $\text{cov}(p_{n_i}^{(v)}, p_{n_j}^{(v)})$ denotes the covariance function in (4.13) using all the training samples at the v-th mode where $v = z, w$. Substituting (4.17) into (4.16), the final calculations of the covariance between the variables p_{n_i} and p_{n_j} are,

$$\begin{aligned}\text{cov}(p_{n_i}, p_{n_j}) &= \overline{\mathcal{K}}(\boldsymbol{B}_{n_i}, \boldsymbol{B}_{n_j}) \\ &= \frac{1}{V^2}\sum_{z,w=1}^{V} \text{cov}(p_{n_i}^{(z)}, p_{n_j}^{(z)})\text{cov}(p_{n_i}^{(w)}, p_{n_j}^{(w)})\end{aligned} \tag{4.18}$$

Note that the above calculations of the covariance between the positional variables only use the source features (i.e., source positional information is not required). Hence, both labeled and unlabeled training samples are exploited. In the next subsection, we show how to estimate the unknown testing source position using a GPR tool.

4.4.2 Estimate the Unknown Source Position Using GPR

Based on the MMGP, localization of a single sound source, located at an unknown source position, can be reviewed as a regression problem,

$$\begin{aligned}\overline{p}_{n_\ell} &= p_{n_\ell} + \varepsilon_{n_\ell} \\ &= f(\boldsymbol{B}_{n_\ell}) + \varepsilon_{n_\ell}, \quad n_\ell = 1, \cdots, \mathcal{N}_L\end{aligned} \tag{4.19}$$

where n_ℓ is the index of labeled training sources, $\overline{p}_{n_\ell}$ and p_{n_ℓ} denote the measured and desired source position, respectively, $f(\boldsymbol{B}_{n_\ell})$ is the mapping function between the n_ℓ-th source feature $\boldsymbol{B}_{n_\ell}$ and source position p_{n_ℓ} and $\varepsilon_l \sim \mathcal{N}(0, \sigma^2)$ denotes a zero-mean Gaussian noise (i.e., the calibration inaccuracies originating from inevitable errors such as imprecise positional measurements).

Given the feature matrix $\boldsymbol{B}^\star$ of a testing source, feature set of the labeled training samples, i.e. $\boldsymbol{B} = [\boldsymbol{B}_1, \cdots, \boldsymbol{B}_{N_L}]$, and their positional information $\mathbf{p}$, this chapter recovers the unknown position of the testing source using a standard

Bayesian approach,

$$\begin{aligned} Pr(f^{\star}|\boldsymbol{B}^{\star},\boldsymbol{B}) &= \int Pr(f^{\star},f|\boldsymbol{B}^{\star},\boldsymbol{B})df \\ &= \int Pr(f^{\star}|f,\boldsymbol{B}^{\star},\boldsymbol{B})Pr(f|\boldsymbol{B}^{\star},\boldsymbol{B})df \end{aligned} \tag{4.20}$$

in which symbols of f and $f^{\star}$ represent source position of $f(\boldsymbol{B})$ and $f(\boldsymbol{B}^{\star})$ respectively. From (4.20), we see probability formula for the testing source are determined by two conditional probability expressions, i.e., $Pr(f^{\star}|f,\boldsymbol{B}^{\star},\boldsymbol{B})$ and $Pr(f|\boldsymbol{B}^{\star},\boldsymbol{B})$. For the sake of clarity, we ignore the intermediate procedures and directly present the closed-form probability distribution of $Pr(f^{\star}|f,\boldsymbol{B}^{\star},\boldsymbol{B})$ and $Pr(f|\boldsymbol{B}^{\star},\boldsymbol{B})$, both of which follow a Gaussian distribution as,

$$Pr(f^{\star}|f,\boldsymbol{B}^{\star},\boldsymbol{B}) \sim \mathcal{N}\left(\mathbf{K}^{\star^T}\mathbf{K}^{-1}f, \mathbf{K}^{\star\star} - \mathbf{K}^{\star^T}\mathbf{K}^{-1}\mathbf{K}^{\star}\right) \tag{4.21}$$

$$Pr(f|\boldsymbol{B}^{\star},\boldsymbol{B}) \sim \mathcal{N}\left(\mathbf{K}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{p}, \sigma^2(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\right) \tag{4.22}$$

where notation of $\mathbf{I}$ is an Identity matrix, $\mathbf{K}^{*} \in \mathbb{R}^{N_T \times N_L}$ is the covariance matrix containing the covariance of two arbitrary positional variables between the training and testing sources, $\mathbf{K} \in \mathbb{R}^{N_L \times N_L}$ and $\mathbf{K}^{**} \in \mathbb{R}^{N_T \times N_T}$ represent the covariance matrix for the training and testing sources, respectively. Note that N_T above denotes the total number of testings based on the recordings from a single source. Multiplying Gaussian distributions of (4.21) and (4.22), probability distribution of the testing source position conditioned on the training set, i.e. $Pr(f^{\star}|\boldsymbol{B}^{\star},\boldsymbol{B})$, follows a Gaussian distribution as well,

$$\mathcal{N}\left(\mathbf{K}^{*}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{p}, \mathbf{K}^{**} - \mathbf{K}^{*}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{K}^{*^T}\right). \tag{4.23}$$

Hence, the unknown positions of the testing source $\mathbf{p}^{*} = [p_1^{*}, \cdots, p_{N_T}^{*}]^T$ is given by the mean value of the Gaussian distribution in (4.23) as the probability reaches its global maximum,

$$\mathbf{p}^{*} = \mathbf{K}^{*}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{p} \tag{4.24}$$

which can be interpreted as linear combination of the source positions in the labeled training set, i.e., $p^{*} = \mathbf{w}^T\mathbf{p}$ where $\mathbf{w}^T = \mathbf{K}^{*}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}$ are the linear weights.

Alternatively, the estimator of (4.24) can also be regarded as a linear combination $p^* = \mathbf{K}^*\mathbf{u}$, whose weights are $\mathbf{u} = (\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{p}$.

Some necessary comments are given with respect to the mapping function above:

- The mapping function is semi-supervised as it requires no positional information of the unlabeled samples. Although the unlabeled samples do not appear explicitly in (4.24), they play a part in the calculations of the covariance between positional variables. Usage of the unlabeled samples enables more precise measurement of the covariance for the MMGP modeling. Additionally, they exert a negligible influence on the practicality of the algorithm as we can easily obtain the unlabeled samples by randomly sampling the source area.

- There remain some unknown parameters in the estimated position of (4.24), i.e., parameters of the covariance function and noise covariance. This chapter uses the marginal likelihood [155] to specify the parameters. However, its non-convexity easily leads to local optimality with non-negligible errors from the global optimal results. To tackle this issue, we adopt the empirical method of cross-validation [155] to split the training set into two disjoint sets, one of which is used for training, and the other set, i.e., the validation or reference set, is used to monitor the performance.

- The number of parameters in our mapping function depends on the total number of spherical harmonic modes. It is more difficult to optimize a larger number of parameters simultaneously, as well as associate with increased algorithm complexity. From this point of view, the spherical harmonic modes selector not only increases the validity of the source features but also reduces the algorithm complexity.

4.5 Proposed Multiple Source Localization

4.5.1 Framework of the Algorithm

Multi-source localization in this chapter mainly considers the overlapped recordings as they are very common in practice, such as conversational recordings between

several speakers [149]. Figure 4.3 exhibits the overlapped recordings with a 40% overlapped ratio (i.e., the percentage of the overlapped periods among the recording). Figure 4.4 presents a compact block diagram of the proposed multi-source localization algorithm, which mainly consists of two disjoint stages, i.e., a training stage and a testing stage.

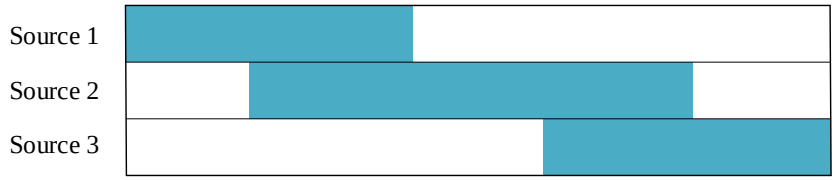


Figure 4.3: An example of overlapped multi-source recordings by 3 sound sources. The cyan color denotes the periods where a sound source is active.

Training stage: (i) Select $\mathcal{N}_L$ labeled and $\mathcal{N}_U$ unlabeled training samples within the defined reverberant sound source area of interest (e.g., Φ). (ii) Measure the recordings due to each training source separately using a higher-order microphone array and then collect the training feature set by estimating the features using the estimator given in (3.16). Note that, since the feature is independent of the source signal, we can use any given source signal (e.g., speech sentences or random signal) to drive the loudspeakers placed at different positions within the source area. (iii) Implement the defined metric of (4.11) to select a proper subset of spherical harmonic modes. (iv) Formulate the mapping function using the MMGP, optimize, and specify the parameters required by the test stage.

Testing stage:

(i) Record the overlapped recordings from multiple sources (e.g., $Q > 1$ sources) within the source area of Φ, divide them into source frames in the time domain (e.g., T frames in total and each lasting 0.5 s), and then obtain their source features using the feature estimator of (3.16). (ii) Use the overlapped frame detection, as explained in the next subsection, to detect and isolate the components overlapped by multiple sources. (iii) Only preserve the source features at the single source frames (e.g., N_T single source frames where $1 \leqslant N_T \leqslant T$), and estimate their positions using the mapping function of (4.24) obtained during the training stage.

(iv) collect all the estimated positions of the single-source frames and use a clustering tool (e.g., K-means [64]) for the final estimates. The final estimated positions correspond to the central location of each cluster.

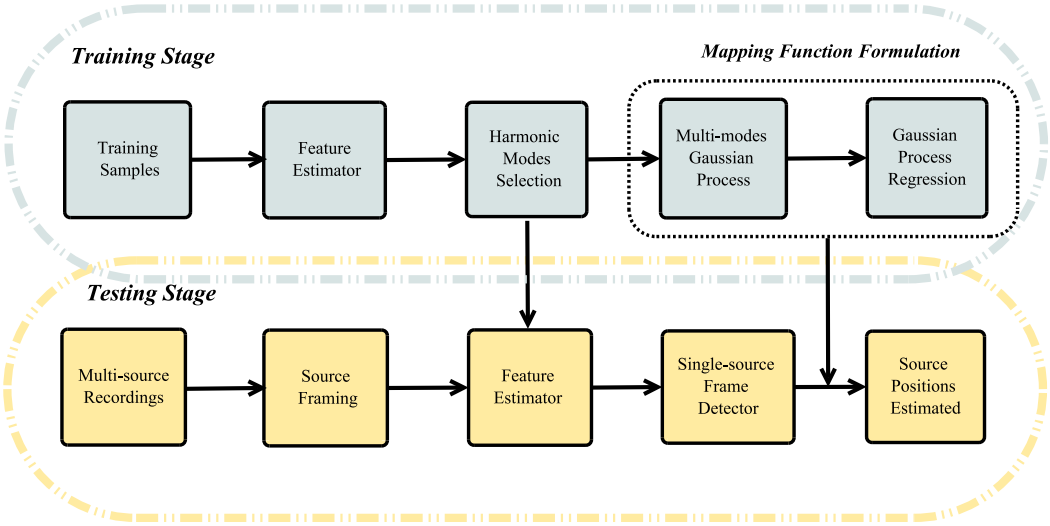


Figure 4.4: Block diagrams of the proposed multiple source localization approach, which mainly comprises of training and test stage respectively.

4.5.2 Overlapped Frame Detection Using the Training Set

This subsection explains the last step of the algorithm in Fig. 4.4, i.e., the overlapped frame detection that simplifies the multi-source localization into a single source localization.

Let us assume that the t-th frame originates from a single source located at $\boldsymbol{p}_q$. Due to the direct relation between source feature(s) and position(s), its feature $\boldsymbol{B}_t^\star$ has a strong similarity to features of the training samples located close to $\boldsymbol{p}_q$. By contrast, if the given source frame is overlapped by multiple sources, the similarity is much weaker since the feature now depends on the time-varying source signal. From this discussion, the single-source frames have a stronger similarity between a subset of the training features, while the overlapped frames, on the contrary,

have a weaker similarity. Hence, we can separate the overlapped and single source frames by introducing a proper metric measuring the similarity. For that, we use a distance function $\widetilde{k}(\cdot)$ to measure the similarity between the source features,

$$d(t, n_\ell) = \mathcal{T}(\boldsymbol{B}_t^\star, \boldsymbol{B}_{n_\ell}) \tag{4.25}$$

where $1 \leq t \leq T$ denotes the index of the segmented source frames in the time domain, $\boldsymbol{B}_{n_\ell}$ denotes the n_ℓ-th training feature matrix. Note that above-mentioned SE kernel function in (4.14), with unknown parameters, cannot be used as the distance function in (4.25). Several theoretical distance metrics can be used, such as the normalized Euclidean distance in (4.31) used in the experimental study. Intuitively, a smaller distance denotes the inputs have a stronger similarity. Then, we use a repetitive calculation over all the training samples to generate a vector,

$$\boldsymbol{d}(t) = \Big[d(t,1), d(t,2), \cdots, d(t, \mathcal{N}_D)\Big]^T \tag{4.26}$$

where both the labeled and unlabeled training samples are used because positional informational is not required. A small subset of elements in $\boldsymbol{d}(t)$ is used to compute the distance,

$$d(t) = \frac{1}{I} \sum_{i=1}^{I} \boldsymbol{d}_i^{\mathrm{s}}(t) \tag{4.27}$$

where $\boldsymbol{d}^{\mathrm{s}}(t)$ denotes the ascending sorted vector of $\boldsymbol{d}(t)$. This measure is successively applied for all T frames to produce,

$$\boldsymbol{d} = \Big[d(1), d(2), \cdots, d(T)\Big]^T. \tag{4.28}$$

Intuitively, given the vector of $\boldsymbol{d}$, we assume that the overlapped source frames, to be isolated from source localization, that satisfy the following inequality,

$$d(t) > \eta, \quad t = 1, \cdots, T \tag{4.29}$$

where η denotes a user defined threshold that is empirically specified. Note that the detection here directly uses the source features, not requiring the source position information, so that both the labeled and unlabeled training samples are exploited.

4.6 Experiments

4.6.1 Experimental Methodology

This section presents experimental results for multi-source localization in noisy and reverberant environments using both the simulated and real-life source recordings. The experiments are implemented following the procedures presented in Fig. 4.4. Note that the source localization approaches localize the source x, y, z-coordinate separately. For simplicity, the following localization scheme focuses on x-coordinate of the sources as localization of other coordinates follows a similar procedure. Performance of our localization system is evaluated using the mean absolute estimated error (MAEE/m),

$$\text{MAEE} = \frac{1}{Q}\sum_{q=1}^{Q} |x_{\text{ori}}(q) - x_{\text{est}}(q)| \tag{4.30}$$

where Q denotes the number of the sound sources presented in the environment, $x_{\text{ori}}(q)$ and $x_{\text{est}}(q)$ represents the original and estimated x-coordinate of the q-th sound source concerning the origin of the room (not the microphone array). Note that the distance function of $\widetilde{k}(\cdot)$ in (4.25), required by the overlapped frame detection, has not been specified yet. Here, we choose to use the normalized Euclidean distance function,

$$\mathcal{T}(\boldsymbol{B}_t^{\star}, \boldsymbol{B}_i) = \frac{||\boldsymbol{B}_t^{\star} - \boldsymbol{B}_1||_2}{||\boldsymbol{B}_t^{\star}||_2||\boldsymbol{B}_i||_2} \tag{4.31}$$

in which $||\cdot||_2$ represents a ℓ_2 norm of the input feature matrix. Note that other distance metrics can be equally used for (4.31).

The experiment adopts two additional source localization approaches for comparisons. (i) The distance function of (4.31), measuring the similarity between the source features of the testing and labeled training sources, is used. For this method, the estimated position equals the labeled training source which locates closest to the testing source. (ii) The other is the state-of-art Multi-Nodes Gaussian Process-based source localization approach using the source feature of ReTF. The original algorithm recently proposed in [45] aims at single source localization and uses ReTF between all pairs of microphones. For a fair comparison, we adjust

and estimate the ReTFs between all the microphones on the surface of the array and the one at the origin of the array, and then apply it to the multi-source localization assisted by the overlapped frame detector. Note that some structured spherical arrays, such as the rigid spherical arrays, only have microphones on the array surface. For such a case, we approximate the pressure at the array origin as the addition of the ones on the surface for the ReTF based localization method.

4.6.2 Simulation Setup

The size of the simulated reverberant room is $6 \times 4 \times 3$ m for the length, width and height, respectively. We set the left-front-bottom corner of the room as the reference origin for the source coordinates, i.e., $(0, 0, 0)$. We simulate an open-sphere spherical microphone array (32 channels and radius 4.2 cm) and place it at an unknown position in the room. The time-domain room impulse response from the sound sources to the microphone array is generated using an available toolbox (i.e., the same one used in Chapter 3) that implements the image source method [154]. Speech signal randomly selected from the TIMIT database at the sampling frequency of 8 KHz is used as the input source signal. We use a convolution operation between the simulated room impulse response and speech signals to generate the measured recordings. After that, Gaussian white noise is added into the time domain recordings. Then, the measured noisy recordings are segmented into 0.5s frames with a 50% overlapping. The segmented time-domain recordings are first transferred into the STFT domain and then decomposed into the spherical harmonics domain. Finally, the proposed estimator in (3.16) is used to compute the RHCs for all the segmented frames. Thirty frequency bins approximately ranging from 1500 Hz to 2500 Hz are exploited, which record the soundfield up to the 2nd order as $N = \lceil kr \rceil$ (i.e., 9 spherical harmonic modes). By contrast, lower frequency bins reduce the uniqueness of the RHC vector whose dimension is reduced to 4 (i.e., 4 spherical harmonic modes). The other drawback at low frequencies is the "Bessel zero problem", causing erroneous estimations of the desired spherical harmonics coefficients because the noise signal can be easily amplified [141]. Higher frequency bins contain less valid speech components.

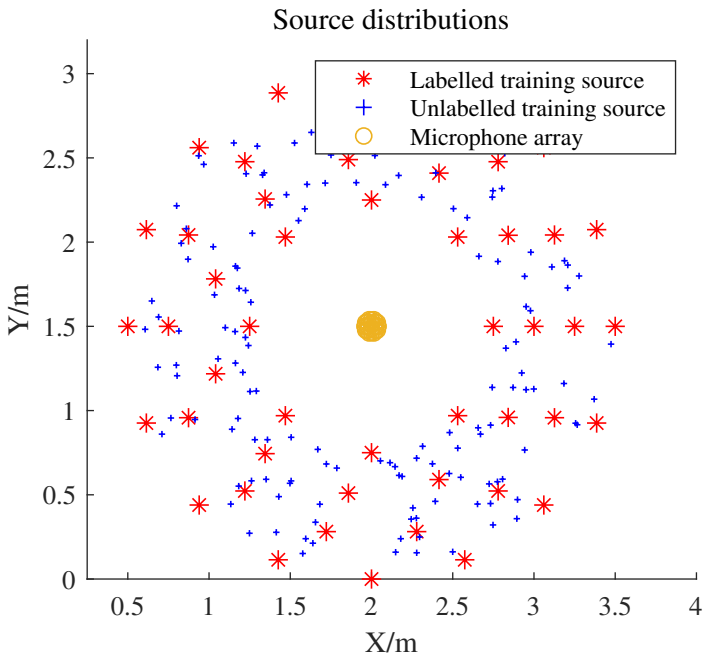


Figure 4.5: Top view of the simulated source distribution. The labeled and unlabeled samples are represented by the red and blue points respectively.

Table 4.1: MAEE of single source localization using different numbers of labeled training samples.

Number	20	33	49	66	86
MAEE/m	0.380	0.314	0.248	0.238	0.223

Sound Source Area for Localization

In the experiments, we apply the proposed method to the scenes of group conversations between multiple speakers. Consider a specific scenario to localize the speakers in a conference room. Hence, the sitting area around the conference table is taken as the source area for localization. Our first task is to select some labeled and unlabeled training samples over the defined source area. We address the problem using two separate steps as follows: (i) Labeled samples selection: Intuitively, the number of labeled training samples involves a trade-off: increasing the number generally leads to higher localization accuracy, while in return it increases the com-

plexity of the system. This algorithm's practicality is highlighted in this chapter. Hence, we select a relatively small number of labeled samples, while still achieving acceptable localization accuracy. Table 4.1 reports the accuracy of single source localization using an increasing number of labeled samples. From the results, we set the labeled training number to 49 because the accuracy starts to degrade severely when using a smaller number. (ii) Unlabeled samples selection: As explained, the unlabeled training samples are much easier to acquire, for simplicity, we directly select 250 unlabeled samples randomly distributed within the defined sound source area.

Table 4.2: Accuracy of overlapped frame detector under various reverberation levels, where the SNR level is 25 dB.

T_{60}/ms	300	400	500	600	700
Accuracy/%	75.0	73.3	71.7	68.3	65.0

Table 4.3: Accuracy of overlapped frame detector under various SNR levels, where the reverberation level is 700 ms.

SNR/dB	5	10	15	20	25
Accuracy/%	46.7	55.0	58.3	61.7	66.7

Figure 4.5 exhibits the sound source area filled by the selected training samples, which encircles the conference table whose radius is 0.75 m. The microphone array, placed at the center of the table, records the incoming soundfield in the reverberant room. In the training stage, we record the soundfield due to each training source separately and estimate the respective feature using the source feature estimator. After that, we implement the spherical harmonic modes selection using the metric in (4.11), and for this particular example, we preserve four spherical harmonic modes in total, whose indexes of (n, m) are $(1, -1), (1,\ 1), (2, -2), (2,\ 2)$, respectively.

4.6.3 Accuracy of Overlapped Frame Detection

The localization scheme proposed in this paper exploits a pre-processing step of the overlapped frame detector. Hence, the accuracy of the detection has a direct influ-

ence on the eventual localization performance. Before the source localization, let us evaluate the effectiveness of the detector. We measure conversational recordings due to three speakers within the defined source area. The recordings, lasting 30 seconds in total, are measured in a reverberant room where $T_{60} = 700$ ms, and are then contaminated by Gaussian white noise with an SNR of 25 dB. The overlapped ratio by the mixed recordings in the time domain is approximately 30%. Note that the overlapped frame detector in (4.27) has a parameter I. The exact number I depends on the total training samples used by the detector. Throughout the simulations, we set at I by around 2% of all the training samples. Hence, $I = 6$ when we use around 300 training samples in simulations. Figure 4.6 exhibits the conversational recordings with a 30% overlapped ratio. The 4-th sub-figure presents the calculated distance of the source frames to the training set. The 5-th sub-figure at the bottom displays the detected overlapped periods. This evaluation is performed in a reverberant room where $T_{60} = 700$ ms, and the recordings are contaminated by 25 dB noise. The results confirm the detector has successfully discovered most of the overlapped components. In the meanwhile, we notice that the detector occasionally detects the frames where the speech is weak or silent, i.e., absent or inactive speech. This is because the source feature is not accurately estimated there, thus has a larger distance to the training set. The capability to detect and remove the weak/inactive speech frames is beneficial for source localization because it ensures the selected frames contain valid speech signals.

We then examine the proposed detector using conversational recordings in diverse environments. We generate the multisource recordings in different acoustic environments, involving simultaneously three speaker positions, with a 30% overlap ratio. Table 4.2 and Table 4.3 report the performance of the detector at different reverberation and SNR levels, respectively. Note that, for each tested room reverberation time, we re-simulated all the training samples and re-calculate all the training feature set. For consistent results, we implement the evaluations up to five times. For each case, the three speakers originate from randomly selected source positions and use randomly selected speech sentences. Hence, each number in both Table 4.2 and Table 4.3 denote the mean detection accuracy of the five groups of evaluations. The results demonstrate that the accuracy gradually degrades in a more complex environment. Under most scenarios, it is capable to recognize more

than 50% of all the overlapped frames.

Finally, we confirm the direct influence of the overlapped frame detection on the localization accuracy. Five repetitive examinations are conducted in the $T_{60} = 700$ ms reverberant room where the SNR level is set at 25 dB. We still adopt three speakers whose overlapped ratio is 30%. We then segment the mixed recordings into 0.5 s frames and then apply the overlapped frame detection to recognize and isolate the overlapped frames. Finally, we apply the proposed semi-supervised localization method to estimate the unknown speakers' positions. The average MAEE over the five groups of evaluations using the overlapped frame detection is 0.205 m. by contrast, the MAEE without the overlapped frame detection is degraded to 0.255 m.

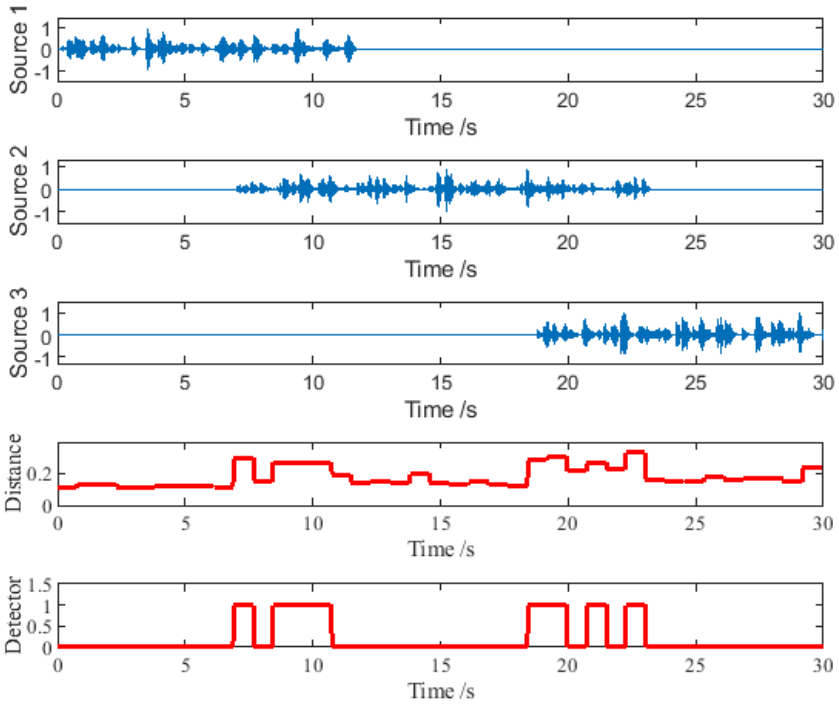


Figure 4.6: Conversation between three speakers (30s long), and the performance of the overlapped frame detector. The distance, calculated by (4.31), denotes the similarity between the features of the testing frame and training set. A larger distance implies this frame is more likely to be an overlapped one.

Table 4.4: MAEE of multiple source localization under various reverberation levels, where the SNR level is 15 dB.

MAEE/m	Reverberation levels (/ms)				
Methods	300	400	500	600	700
ReTF	0.183	0.214	0.253	0.240	0.265
Euclidean	0.301	0.288	0.259	0.296	0.298
All modes	0.207	0.237	0.229	0.259	0.285
Proposed	**0.179**	**0.166**	**0.186**	**0.194**	**0.228**

Table 4.5: MAEE of multiple source localization under various SNR levels, where the reverberation time is 700 ms.

MAEE/m	SNR levels (/dB)				
Methods	5	10	15	20	25
ReTF	0.333	0.301	0.279	0.273	0.244
Euclidean	0.250	0.289	0.315	0.327	0.311
All modes	0.336	0.282	0.289	0.267	0.260
Proposed	**0.246**	**0.221**	**0.232**	**0.192**	**0.204**

4.6.4 Performance of Multi-source Localization

Let us now evaluate the proposed localization method in comparison with the baseline methods. As introduced at the beginning of this section, one baseline is the ReTF based method using Multi-Node Gaussian Process modeling in [45]. The other baseline directly uses a distance metric in (4.31). Besides, we also examine the proposed method without the spherical harmonic modes selection to analyze the proposed feature selector's influence on the localization accuracy. Therefore, four localization approaches are implemented, whose abbreviations used below for convenience are denoted by 'ReTF', 'Euclidean', 'All modes', and 'Proposed', respectively. To increase the reliability of the results, under each acoustic environment (i.e., SNR and room reverberation time), ten successive examinations are implemented. And, each case uses three speakers with randomly selected source positions within the source area and randomly selected speech sentences. Hence, the values presented below denote the mean number over the ten successive evalu-

ations.

Diverse acoustic environments are simulated. We first analyze the impacts of reverberation on the localization algorithms. Table 4.4 displays the performance in different reverberation levels ranging from 300 ms to 700 ms. In each varied reverberation level, we re-recorded the training samples, optimized the parameters, and then applied the settings to the test stage. As expected, we observe that a higher reverberation level has negative impact on the localization accuracy. A stronger reverberation level implies an increased complexity of the acoustic path from the sound sources to the recording area, increasing the difficulty to accurately model the relation between the source features and source positions. We then evaluate the algorithms under various noisy conditions (SNR level ranging from 5 dB to 25 dB). Table 4.5 depicts the results. We recognize slightly degraded localization accuracy when the SNR level decreases. The strong robustness to noise is a result of the proposed biased feature estimator in Chapter 3, which has already alleviated some noise components. Since the estimator has not fully canceled the noise, the algorithms have non-negligible errors when the SNR level becomes very low. These results confirm the superiority of the proposed algorithm over the baseline methods. The improved localization accuracy when using selected harmonic modes, compared with that using all the spherical harmonic modes, validates the effectiveness of the spherical harmonic modes selection.

Table 4.6: Time cost by ten repetitive executions at the test stage.

Methods	Number of views	Time
ReTF	32	239.5s
All modes	9	69.3s
Proposed	4	30.4s

4.6.5 Algorithm Complexity Analysis

In addition to the localization accuracy, it is a necessity to evaluate the data-driven localization algorithm computational complexity. Several factors determine the proposed algorithm complexity, such as the number of labeled and unlabeled training samples, microphone channels in the array, and the soundfield order. Note

that both our proposed methods and the baseline using ReTF adopt multi Gaussian Process modeling so that they generally follow similar procedures. However, their numbers of views differ a lot, causing major consequences on the algorithm complexity. For validations, we evaluate the computational complexity of the algorithms by directly measuring their average time cost, using a Matlab implementation on a standard desktop (CPU Intel Core i7-4790 Quad 3.6 GHz, RAM 16 GB). Table 4.6 presents the speed of the algorithms as well as their numbers of view. The proposed method is much faster than the baseline as it only has 4 views in total. Intuitively, a smaller number of views implies that fewer parameters should be adjusted. A comparison between the method using either the selected number of modes or all the modes confirms the advantage of selecting modes on the computational complexity.

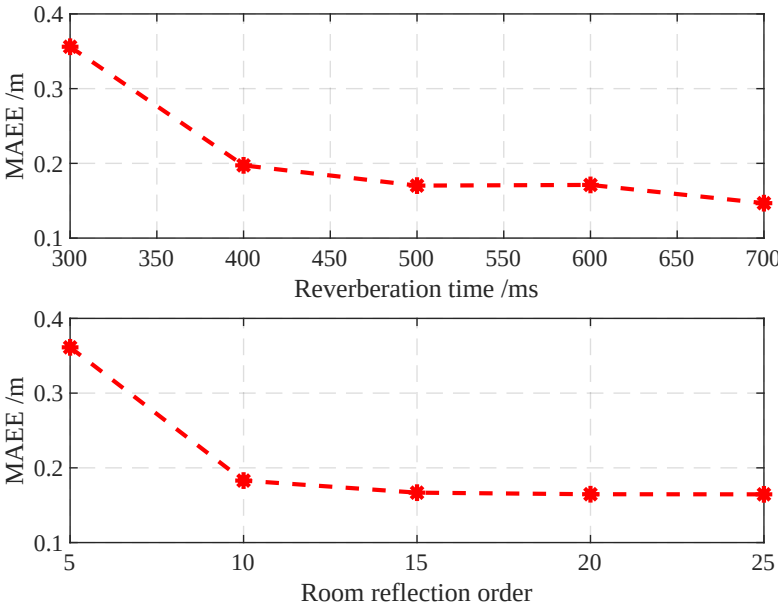


Figure 4.7: MAEE of multiple source localization when room reverberation level is changed during the test stage (SNR is 25 dB). The different room reflection orders are with $T_{60} = 700$ ms.

4.6.6 Robustness of the Algorithm

As assumed, the source feature solely depends on the source position in a static room environment. Hence, the aforementioned assessments assumed that the acoustic environment did not change between the train and test stages. However, this assumption hardly holds in practice. It occasionally happens that the setup of the room changes during the test stage. For example, the doors and windows may be opened or closed, or someone may walk around in the room. To meet practical requirements, our localization method should be robust to changes in the room characteristics. Hence, let us examine our method's robustness. Figure 4.7 reports the localization errors for room environments that are different between test and training stages. We simulate the changes in the test environment by using different room reverberation time as well as varied room reflection orders when $T_{60} = 700$ ms. In the training stage, we generate the training samples at the reverberation $T_{60} = 700$ ms, using a full reflection order. The examination results, presented in Figure 4.7, demonstrate slightly degraded accuracy when the test environment is not significantly different from that in the training stage. Hence, the localization method, learning the cues for localization in the training stage, is still applicable in the different/changed test environments. Additional evaluations at different reflection orders confirmed the improved localization accuracy at a higher reflection order. The reason is the testing source feature at a higher reflection order matches more to the training features that captured a full reflection pattern. However, Figure 4.7 implies the performance degrades more if the testing environment has more different characteristics in comparison with the training environment. And, it is recognized with dramatically reduced localization accuracy when the testing room environments change a lot (e.g., more than 0.35 m error when $T_{60} = 300$ ms or with room reflection order 5).

Table 4.7: Localization performance using different sound speeds in the test stage.

Speed (m/s)	336	339	343	346	350
Temperature (/°C)	8	14	20	25	30
MAEE/m	0.207	0.184	0.157	0.174	0.192

Additionally, the testing environment's temperature or air humidity also occasionally changes, which could be simulated by changing the speed of sound value by a few percent. Hence, we now change the speed of sound in the testing stage and examine the performance of the algorithm for both the training and testing stages the room reverberation time is $T_{60} = 700$ ms and the SNR level is 25 dB. Table 4.7 presents the proposed method's MAEE with various sound speeds ranging from 336 m/s to 350 m/s. Note that the reference temperature, in the training stage, is 20 °C and the corresponding speed is 343 m/s. We observe that with varying values of speed (caused by changes in room temperature), the localization accuracy sometimes degrades. However, with common indoor temperatures, the degradation is minimal.

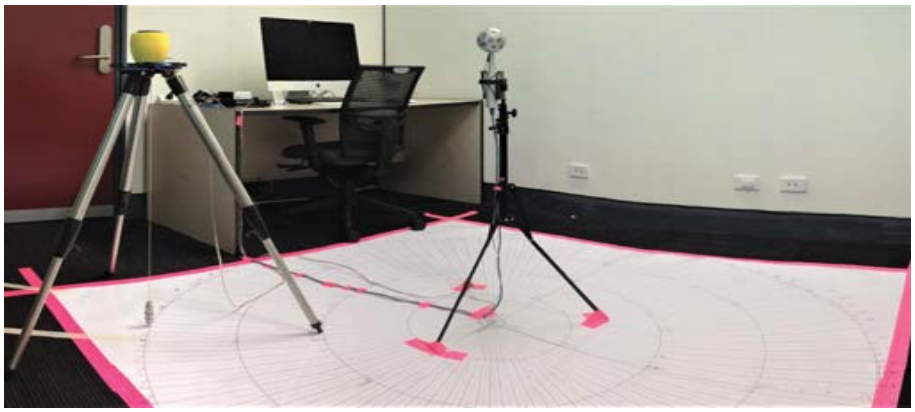
4.6.7 Real Recordings

This subsection validates the availability of the proposed algorithm under real-life scenarios, using practical recordings measured in the acoustic lab of Australian National University.

Experimental Setup

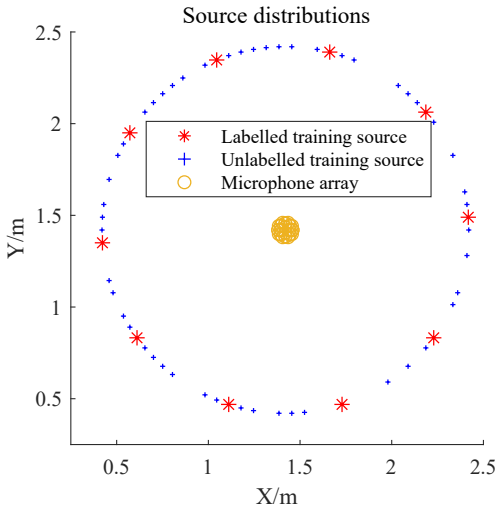
Figure 4.8 presents the setup for the practical measurements, a spherical microphone array called EigenMike and a circular source area, respectively. The EigenMike is a rigid 32-microphone array with a similar size as the above simulated open-sphere array. An advantage of using a rigid array is avoiding the division by very small values in (3.11) at low frequencies, alleviating the aforementioned "Bessel zero problem". The defined source area only comprises of 10 labeled training samples along with 80 unlabeled training samples. The EigenMike, placed at the center of the source area, measures the incoming soundfield. The experiment room dimensions are $[3.54, 4.06, 2.70]$ for the length, width and height, respectively, with the reverberation time around $T_{60} = 330$ ms. The same frequency band used by the simulated recordings, ranging from 1500 Hz to 2500Hz, is exploited for the real recordings.

Note that we obtain the real recordings using a convolution operation between the measured room impulse response (RIR) and the source signal. Hence, it is



(a)

(b)



(c)

Figure 4.8: (a): The setup for practical acoustic measurements used by our source localization approach in a reverberant room. (b): The commercial EigenMike and the mini-loudspeaker. (c): Top view of the defined source area in experiments, i.e., a 1m circle.

of great necessity to ensure high-quality RIR measurements. During practical recordings, the system time delay, caused by the hardware, for example, is unavoidable. It degrades the spatial measurement accuracy if the unknown delay is large. Here, we provide a calibration technique to measure the delay by attaching a mini-loudspeaker (Manufacturer: VISATON, External Diameter: 16mm) close to

the EigenMike (see Fig. 4.8 (b)). Specifically, when driving the desired loudspeaker, we simultaneously drive the mini-loudspeaker using a known labeled signal. Since the two speakers are driven synchronously, the delay can be detected by the location of the labeled signal within the measured recordings. Note that we just measure the system delay once as it generally keeps constant. When the delay is known, we then extract the source recordings right after the delay time where contains the valid source signal.

Validation of the Illustration in Section II C

Before presenting the localization accuracy, we first use real-life recordings to validate the illustration that the RHCs are independent of the particular source signal. We first compare the source features generated by the same sound source while using different source signals. For generality, ten pieces of random signal lasting around 0.5 seconds are used. For each signal, we calculate the mean values of the RHCs over a wide frequency band ranging from 1500 Hz to 2500 Hz. Figure 4.9 (a) depicts the real part of source features, using a sound source whose polar

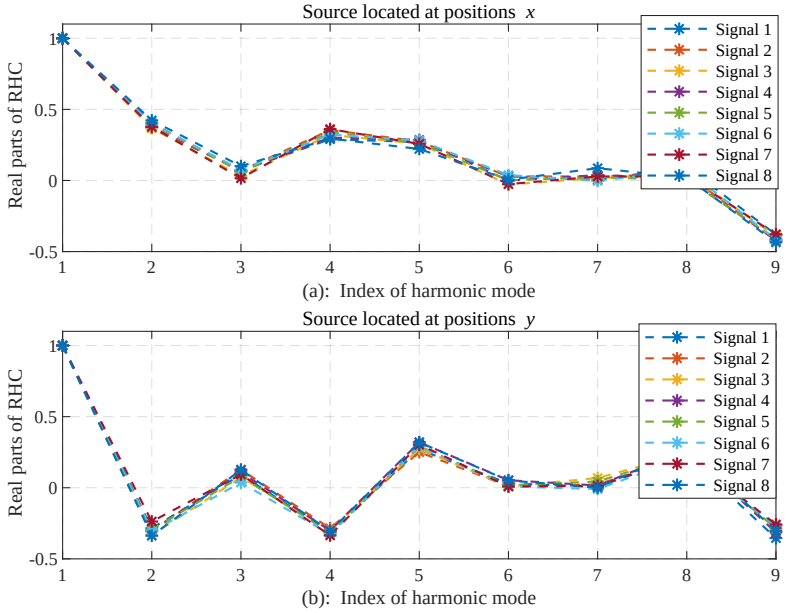


Figure 4.9: Real parts of the features for sources located at different positions. Note that, for convenience, the presented values denote the average over the wide frequency band.

coordinates are $(r_1, \theta_1, \phi_1) = (1, 1.57, 3.63)$ with respect to the EigenMike's origin. For this sound source, its source feature is repetitively estimated using ten random signals. The observed consistency of the features using different random signal confirms its independence from the specific signal. Note that the curves presented in Figure 4.9 (a) also contain a slight inconsistency. One possible reason is the feature estimator in (3.16) uses a short frame windowing, which cannot cover the full reverberated test signal and therefore causes slight inconsistency on the estimated RHC.

Then, we expect to see whether the source feature significantly changes if placing the sound source at a different source position. Figure 4.9 (b) depicts the real part of source features, due to the sound source located at a new position, i.e., $(r_2, \theta_2, \phi_2) = (1, 1.57, 0.56)$ with respect to the array origin. We use the same setting to estimate the source features as the case in Figure 4.9 (a). We observe much greater differences between the source features in sub-figure (a) and (b), representing the sources located at different positions have different source features. Above analysis confirms, in a real-life reverberant room, the defined feature is mostly source-independent and changes significantly when the source position changes.

Finally, we present a quantitative study on how the RHC changes when the source moves to different positions. We first pick one reference position located at $(1, 1.57, 3.21)$ within the source area in Figure 4.8. Then, we move the source to different positions with respect to the reference position and examine how the feature changes. For simplicity, the movement is carried along the azimuth axis only while elevation and distance are fixed. Note that we drive the source using a randomly generated signal and then use the proposed estimator to calculate the corresponding RHC. For quantitative evaluations, we use the normalized Euclidean distance function in (4.31) to measure the features' change. A larger distance value denotes the feature changes more significantly. Figure 4.10 denotes the changes of RHC against the increasing value of the source azimuth change. It is observed that the RHC changes proportionally to the deviation of source azimuth. Aforementioned analysis using real recordings verifies the arguments that the defining feature is source-independent and mainly depends on the source position. Thus, we conclude that the RHC contains relevant cues to localize the source position.

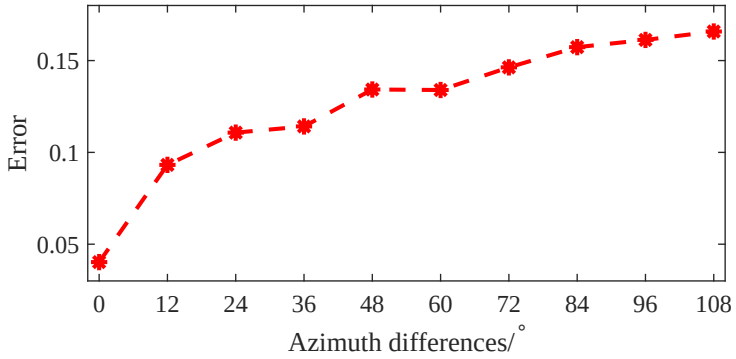


Figure 4.10: The changes of the source feature with an increasing change of the source azimuths.

Localization Using Conversational Recordings

We exactly follow the steps summarized in Fig. 4.4 to complete both the training and testing stage. We use ten measurement groups, each containing three sound sources at randomly selected positions within the circular area. Each source uses a unique speech sentence lasting around 20 s, and the mixed multi-source recordings measured by the array have an overlapped ratio of about 30%. Table 4.8 presents the performance using all the algorithms. Each number denotes the mean MAEE over the ten measurements. Improved localization accuracy over the baselines confirms the availability of the proposed multi-source localization approach under real-life scenarios.

Table 4.8: Average MAEE of multi-source localization using 10 groups.

Distance/m	ReTF	Euclidean	All modes	Proposed
MAEE/m	0.159	0.205	0.181	0.120

Localization Using Significantly Overlapped Recordings

The aforementioned examinations of the algorithms are limited to conversational recordings, whose overlapped ratios are generally mild (e.g., the overlapped ratio is 30% or less). In the remained content, we implement the proposed method in unfavorable circumstances where the recordings have a severe overlapped ratio (e.g., higher than 50%). Figure 4.11 demonstrates significantly overlapping recordings. Then, we use the proposed detector to recognize the overlapped frames. The 4-

Table 4.9: MAEE of multiple source localization using strong overlapped recordings.

MAEE/m	Overlapped ratio (%)				
Methods	50	60	70	80	90
ReTF	0.192	0.187	0.193	0.206	0.214
Euclidean	0.217	0.223	0.244	0.214	0.209
All modes	0.191	0.195	0.202	0.211	0.205
Proposed	**0.141**	**0.143**	**0.146**	**0.161**	**0.175**

th and 5-th sub-figure present the calculated distance to the training set and the detected overlapped periods, respectively. The results confirm that it successfully detects most of the overlapped components. We further evaluate the algorithm's localization accuracy using such severely overlapped recordings. We use ten measurement groups where each consists of three sound sources. The measured multi-source recordings have varied overlapped ratios ranging from 50% to 90%. Table 4.9 reports the localization accuracy using all the algorithms. The results show slightly degraded localization accuracy when the overlapped ratio gradually increases. The reason is the overlapped frame detection accurately isolates most invalid frames (even when the overlapped ratio is up to 90%), thus all the approaches are then capable of localizing the sources successfully. Being consistent with the above evaluations, the proposed algorithm outperforms the baselines by achieving improved localization accuracy.

4.7 Summary

This chapter used the relative harmonic coefficients to achieve a semi-supervised multi-source localization algorithm in a noisy and reverberant environment. Extensive simulations showed that the proposed algorithm achieved improved localization accuracy in comparison with the state-of-art approaches. Real-life evaluations confirmed the capability of this method even in unfavorable cases of severe source overlapping recordings. Several aspects of the proposed method are highlighted: (i) A further investigation of the relative harmonic coefficients: including a directiv-

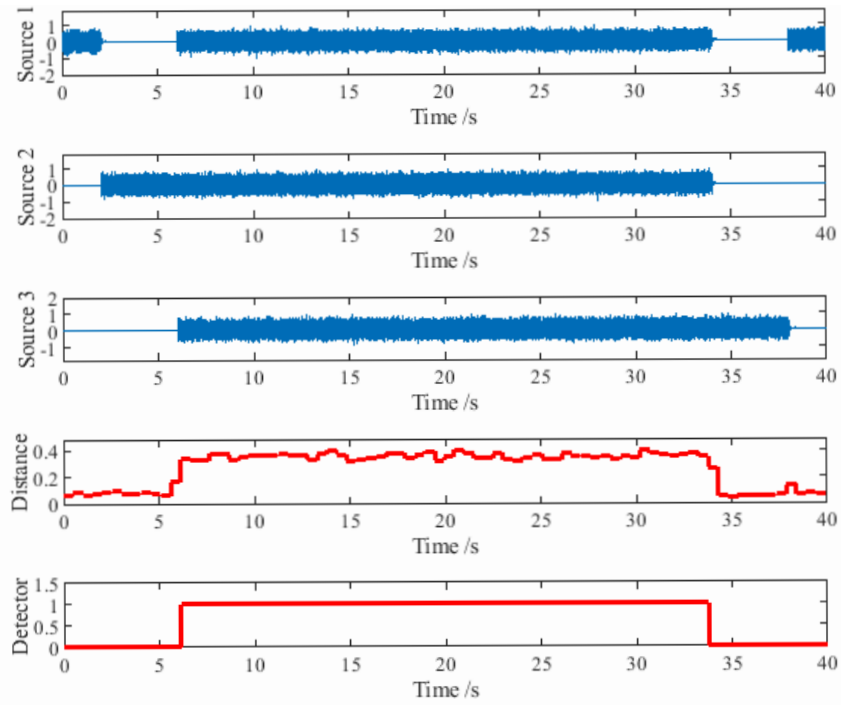


Figure 4.11: Overlapped frame detector for significantly overlapped recordings. Around 70% of the recordings, in the middle, are overlapped by the three sources sending out random source signal.

ity pattern analysis, a data-driven feature selector as well as an overlapped frame detector. (ii) The Multi-Mode Gaussian Process modeling (MMGP) efficiently fuses/merges the source features at the selected spherical harmonic modes, each representing a distinct/unique description of the soundfield. (iii) The unlabeled training samples not only enable a more precise measurement of the covariance for the MMGP modeling but also play an active role in the source feature selection and overlapped frame detection while exerting a negligible influence on the algorithm practicality.

While the proposed method performs better than the baseline methods, some inherent limitations of it include: (i) current biased feature estimator with relatively short window frames may not fully cover a strong reverberation, which causes some inconsistency between the testing and training features in strong reverberant environments; (ii) this paper mainly considers the overlapped recordings so that

is unusable for the simultaneous multi-source recordings, i.e., with an overlapped ratio of 100%. In the near future, we intend to propose a new feature estimator that better suits for strong noisy and reverberant environments and then achieve sufficient localization accuracy for simultaneous multiple source recordings in the complex environments.

4.8 Appendix A

Samarasinghe *et al.* [89] generalized the image source method into the spherical harmonics domain. Assume an outgoing sound source, soundfield over the recording area is,

$$P(\boldsymbol{x},k)=\sum_{n=0}^{N}\sum_{m=-n}^{n}\sum_{v=0}^{V}\sum_{u=-v}^{v}\widehat{\alpha}_{nm}^{vu}(k)\gamma_{vu}(k)\,j_n(kr)\,Y_{nm}(\theta,\phi) \tag{4.32}$$

in which $\gamma_{vu}(k)$ represents the spherical harmonic coefficients of the outgoing sound source, and $\widehat{\alpha}_{nm}^{vu}(k)$ is referred as the coupling coefficients relating the source and receiver region. In a rectangle room simulated by the image source method, the coupling coefficients are,

$$\begin{aligned}\widehat{\alpha}_{nm}^{vu}(k)=&\sum_{\boldsymbol{p}=0}^{1}\sum_{\boldsymbol{r}=-\infty}^{\infty}\lambda_{x1}^{|r1-q|}\lambda_{x2}^{|r1|}\lambda_{y1}^{|r2-j|}\lambda_{y2}^{|r2|}\lambda_{z1}^{|r3-\ell|}\lambda_{z2}^{|r3|}\\&\times(-1)^{(j+\ell u)+\ell v}S_{vn}^{((-1)^{q+j}u)m}(\boldsymbol{R}_p+\boldsymbol{R}_r).\end{aligned} \tag{4.33}$$

In the (4.32), the spherical harmonic coefficient follows as,

$$\alpha_{nm}(k)=\sum_{v=0}^{V}\sum_{u=-v}^{v}\widehat{\alpha}_{nm}^{vu}(k)\gamma_{vu}(k). \tag{4.34}$$

Its $\beta_{nm}(k)$, the ratio between $\alpha_{nm}(k)$ and $\alpha_{00}(k)$, is,

$$\beta_{nm}^{\text{rev}}(k)=\frac{\sum_{v=0}^{V}\sum_{u=-v}^{v}\widehat{\alpha}_{nm}^{vu}(k)\gamma_{vu}(k)}{\sum_{v=0}^{V}\sum_{u=-v}^{v}\widehat{\alpha}_{00}^{vu}(k)\gamma_{vu}(k)}. \tag{4.35}$$

Since the order of the omni-directional point source with respect to its location is zero, (4.35) can be simplified into,

$$\begin{aligned}\beta_{nm}^{\text{rev}}(k) &= \frac{\sum_{v=0}^{0}\sum_{u=-v}^{v}\widehat{\alpha}_{nm}^{vu}(k)\gamma_{vu}(k)}{\sum_{v=0}^{0}\sum_{u=-v}^{v}\widehat{\alpha}_{00}^{vu}(k)\gamma_{vu}(k)} \\ &= \frac{\widehat{\alpha}_{nm}^{00}(k)}{\widehat{\alpha}_{00}^{00}(k)} = \widehat{\alpha}_{nm}^{00}(k)\rho_o\end{aligned} \tag{4.36}$$

in which $\rho_o = 1/\widehat{\alpha}_{00}^{00}(k)$ is a constant. Substituting $v = 0$ and $u = 0$ into the coupling coefficient of (4.33), its relative harmonic coefficients are:

$$\begin{aligned}\beta_{nm}^{\text{rev}}(k) = \rho_o \sum_{\boldsymbol{p}=0}^{1}\sum_{\boldsymbol{r}=-\infty}^{\infty} &\lambda_{x1}^{|r1-q|}\lambda_{x2}^{|r1|}\lambda_{y1}^{|r2-j|}\lambda_{y2}^{|r2|}\lambda_{z1}^{|r3-\ell|}\lambda_{z2}^{|r3|} \\ &\times (-1)^{(j+\ell m)+\ell n} S_n^m(\boldsymbol{R}_p + \boldsymbol{R}_r)\end{aligned} \tag{4.37}$$

where the $S_n^m(\cdot)$, whose inputs are a combined coordinate of the sound source and microphone array, is written as,

$$S_n^m(\boldsymbol{x_o}) = 4\pi i^n \sum_{l=0}^{n} i^l (-1)^{-m} j_l(k|\boldsymbol{x_o}|) Y_{l(m)}^*(\theta_{x_o}, \phi_{x_o}) \sqrt{(2n+1)(2l+1)/4\pi} W_1 W_2, \tag{4.38}$$

where W_1 and W_2 denote the Wigner $3-j$ symbols. with

$$W_1 = \begin{pmatrix} 0 & n & l \\ 0 & 0 & 0 \end{pmatrix}, \text{and } W_2 = \begin{pmatrix} 0 & n & l \\ 0 & -m & m \end{pmatrix} \tag{4.39}$$

denote the Wigner $3-j$ symbols.

4.9 Related Publications

This chapter's work has ever been published/submitted in the following journal papers and conference proceedings.

- Y. Hu, P. N. Samarasinghe, S. Gannot, and T. D. Abhayapala, "Semi-supervised multiple source localization using relative harmonic coefficients under noisy and reverberant environments," *IEEE/ACM Transactions on*

Audio, Speech and Language Processing (TASLP), vol. 28, pp. 3108-3123, 2020.

- Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using realtive harmonic coefficients", in 2020 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571-575.

Chapter 5

Multiple Source Localization in Noisy Environments Using Relative Sound Pressure Based MUSIC

Overview: *This chapter addresses the multiple source localization for simultaneous multi-source recordings by developing a novel MUSIC algorithm. This developed MUSIC approach uses the estimations of the relative sound pressure for a higher-order microphone array, which is more suitable in noisy environments than the traditional MUSIC method. The proposed MUSIC approach is also decomposed into the spherical harmonics domain where a frequency smoothing technique is allowed to de-correlate the coherent source signals for improved localization accuracy. This algorithm is also capable of estimating the number of active sound sources, which is pre-requisite knowledge for the traditional MUSIC approach. Extensive experimental results using both simulated and real-life recordings confirm the advantages of the proposed algorithm over the traditional MUSIC method.*

5.1 Introduction

The multi-source localization method developed in Chapter 4 requires the availability of single-source components in the time domain, thus it is unusable for simultaneous multi-source recordings as they have no time-domain single-source components. By contrast, subspace methods use the simultaneous recordings directly [18, 21, 71]. The most popular subspace method shall be the multiple signal classification (MUSIC) method [18], attracting great attention due to its easy implementation with reasonable performance [72, 73, 74, 75, 76, 77, 78]. In the recent decade, the MUSIC based approaches using a higher-order microphone array are decomposed into the spherical harmonics domain (SHD) [16]. One of the main advantages of spatial decomposition is the decoupling of frequency-dependent and angular-dependent components. The spherical harmonics domain MUSIC (SHD-MUSIC) was first proposed in [101] by Abhayapala. Rafaely *et al.* in [17] improved the localization accuracy of SHD-MUSIC using frequency smoothing to de-correlate coherent source signals. Then, [157] studied the spherical harmonics domain root-MUSIC. The MUSIC approaches above only consider free-field propagation, thus the localization accuracy degrades in reverberate environments. To overcome this drawback, Birnie *et al.* recently proposed an improved SHD-MUSIC in [86] which uses a complete model of environmental reverberation so that it is more suitable for reverberate environments. However, this method requires the region-to-region coupling coefficients [89] to be known or measured in advance (refer to Chapter 4.2.2 for more details about the coupling coefficients). Nevertheless, the coupling coefficients depend on a set of parameters in the reverberate room (e.g., room sizes, wall reflection coefficients), which are hard to know in practice.

Another major limitation of MUSIC based methods is their vulnerability to noise [158]. This results in the commonly referred phenomena called the *subspace swap*, which is when the measured signal better approximates/represents the noise subspace rather than the signal subspace [73]. The noise signal exerts a more negative influence on the SHD-MUSIC approaches [17]. Apart from the drawbacks of subspace swap, they additionally suffer from the "Bessel zero problem", causing erroneous estimation of the desired spherical harmonic coefficients because the noise signal can be easily amplified in the spherical harmonics domain [141]. To alleviate

the problem, some specially structured microphone arrays have been designed, such as the dual spherical microphone array [159] and rigid spherical microphone array [160], while at the cost of more complicated microphone array requirements. Up to the author's best knowledge, the degraded localization accuracy of SHD-MUSIC methods in noisy environments remains to be improved.

The source feature relative transfer function (ReTF) relating a microphone pair has been widely used in recent source localization in noisy environments (discussed in Chapter 2.3). Inspired by the wide applications of ReTF, this paper defines the *relative sound pressure* as the ratio between the sound pressure on the surface of a higher-order microphone array and the pressure at the array origin. A robust method to estimate the quantity of relative sound pressure is provided. Then, the traditional MUSIC framework under far-field scenarios is re-defined using relative pressure estimations as the input. This new algorithm (abbreviated as RMUSIC) is shown to be capable of estimating multi-source DOAs with improved robustness to noise.

Since the relative sound pressure with respect to the origin can be interpreted as normalized pressure, the above framework is also capable of being represented in the spherical harmonics domain. Thus, a relative sound pressure based spherical harmonic domain MUSIC algorithm (SHD-RMUSIC) is also developed. This algorithm includes a frequency smoothing step for improved accuracy. We note that the "Bessel zero problem" encountered by open spherical arrays is naturally eased with the proposed SHD-RMUSIC due to its inherent robustness to noise. Lastly, both algorithms proposed (RMUSIC and SHD-RMUSIC) have the additional capability of estimating the active number of sound sources at least when realistic SNR levels are present. The remainder of the chapter is structured as follows. Firstly, we formulate the problem to be addressed and introduce the relative sound pressure. Then, we propose the novel RMUSIC exploiting relative sound pressure estimations. Thereafter, Section 5.4 investigates the spherical harmonics domain variation of RMUSIC, which is abbreviated by SHD-RMUSIC. Section 5.5 presents extensive experimental results. Finally, conclusions are drawn.

5.2 System Model

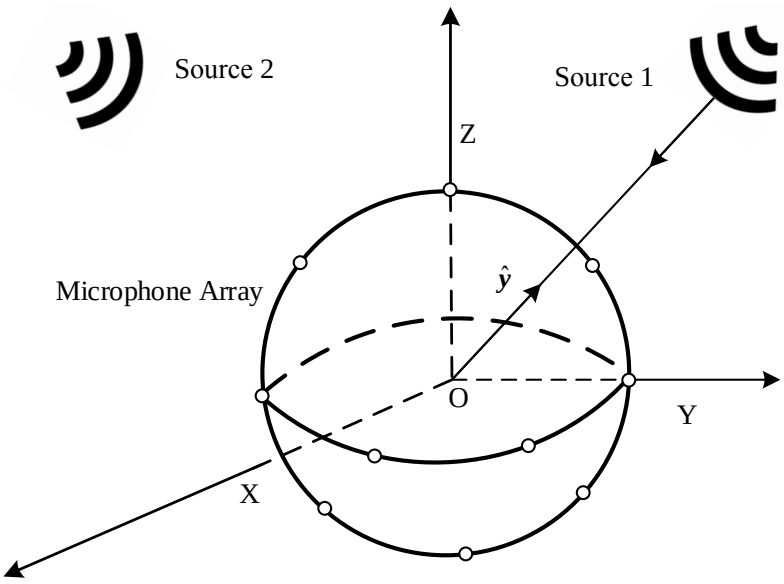


Figure 5.1: Multiple-source DOA estimation using a spherical microphone array.

5.2.1 Problem Formulation

Consider a higher-order microphone array, e.g., spherical microphone array, with M microphones, whose polar coordinates are $\boldsymbol{x}_j = (r, \theta_j, \phi_j)$, $j = 1, \cdots, M$, with respect to its local origin O (see Fig. 5.1). Assume there are L simultaneously active sound sources located at far-field of the array at angles $\Psi_q = (\theta_q, \phi_q)$, $q = 1, \ldots, Q$, with elevation θ_q and azimuth ϕ_q. Hence, the sound pressure, measured by the j-th microphone, in the frequency domain is written as,

$$\begin{aligned} \overline{P}(\boldsymbol{x}_j, k) &= P(\boldsymbol{x}_j, k) + n(\boldsymbol{x}_j, k) \\ &= \sum_{q=1}^{Q} S_q(k) e^{i\boldsymbol{k}_q^T \boldsymbol{x}_j} + n(\boldsymbol{x}_j, k) \end{aligned} \tag{5.1}$$

where $k = 2\pi f/c$ is the wave number, f is the frequency and c is the speed of sound, $P(\boldsymbol{x}_j, k)$ and $\overline{P}(\boldsymbol{x}_j, k)$ correspond to the noiseless and noisy sound pressure,

respectively, $S_q(k)$ denotes the q-th source signal as observed at the origin, $n(\boldsymbol{x}_j, k)$ denotes the additive noise signal at the j-th microphone, and the wavenumber vector is represented by $\boldsymbol{k}_q = (k\cos\phi_q \sin\theta_q, k\sin\phi_q \sin\theta_q, k\cos\theta_q)^T$. Note that the model in (5.1) assumes free field propagation. We then rewrite (5.1) in a vector form,

$$\mathbf{P}(k) = \mathbf{V}(k)\mathbf{s}(k) + \mathbf{n}(k) \tag{5.2}$$

where $\mathbf{P}(k)$ denotes the $M \times 1$ vector of the measured sound pressure at the microphones, $\mathbf{n}(k)$ denotes the $M \times 1$ noise vector, $\mathbf{s}(k)$ denotes the $L \times 1$ vector of the source signal,

$$\mathbf{s}(k) = [S_1(k), S_2(k), \cdots, S_q(k)]^T, \tag{5.3}$$

$\mathbf{V}(k)$ denotes the $M \times Q$ steering matrix,

$$\mathbf{V}(k) = [\mathbf{v}_1(k), \mathbf{v}_2(k), \ldots, \mathbf{v}_M(k)]^T \tag{5.4}$$

where $\mathbf{v}_j(k) = \left[e^{i\mathbf{k}_1^T \boldsymbol{x}_j}, e^{i\mathbf{k}_2^T \boldsymbol{x}_j}, \cdots, e^{i\mathbf{k}_L^T \boldsymbol{x}_j}\right]^T$ represents a steering vector of a microphone. Note that the additive noise in (5.1) is assumed to be non-directional (e.g., random white noise) otherwise, the directional noise can be treated as additional sources to be localized. This chapter intends to estimate the unknown DOAs of all the active sound sources, i.e., (θ_q, ϕ_q), $q = 1, \cdots, Q$, as well as counting the total sound source number using the noisy source recordings. Currently, the subspace method of MUSIC is one of the most frequently used approaches addressing this problem. However, the traditional MUSIC approach has two major drawbacks: (i) it is vulnerable to noise so that its localization accuracy degrades severely in noisy environments; (ii) it requires the source number to be known in advance, while this information is hardly known in practice. To overcome the above limitations, this chapter proposes an improved MUSIC approach using the normalized received signal called *relative sound pressure*, as is introduced in the next subsection.

5.2.2 Relative Sound Pressure (RSP) Definition

This subsection introduces the relative sound pressure of a higher-order microphone array. Let us consider the j-th microphone on the surface of the spherical

microphone array in Fig. 5.1. Its relative sound pressure with respect to the sound pressure at the origin of the array $\boldsymbol{x}_o = (0, 0, 0)$ is defined by,

$$Q(\boldsymbol{x}_j, k) = \frac{P(\boldsymbol{x}_j, k)}{P(\boldsymbol{x}_o, k)}, \qquad j = 1, \cdots, M. \tag{5.5}$$

The above definition requires the availability of the recordings at the array origin. However, some array structures, such as the rigid spherical arrays [160], only have microphones on the array surface. For such a case, we approximate the pressure at the array origin as the addition of the ones on the surface of the array, i.e.,

$$P(\boldsymbol{x}_o, k) \approx \frac{1}{M} \sum_{j=1}^{M} P(\boldsymbol{x}_j, k). \tag{5.6}$$

Note that at the case of a single source, i.e., $L = 1$, the relative sound pressure of the two microphones is equivalent to its relative transfer function (ReTF) [30],

$$Q(\boldsymbol{x}_j, k) = \frac{P(\boldsymbol{x}_j, k)}{P(\boldsymbol{x}_o, k)} = \frac{S(k)A(\boldsymbol{x}_j, k)}{S(k)A(\boldsymbol{x}_o, k)} = \frac{A(\boldsymbol{x}_j, k)}{A(\boldsymbol{x}_o, k)} \tag{5.7}$$

in which $S(k)$ is the source signal, $A(\boldsymbol{x}_j, k)$ and $A(\boldsymbol{x}_o, k)$ represent the acoustic transfer function from the sound source to the microphones, respectively, and $A(\boldsymbol{x}_j, k)/A(\boldsymbol{x}_o, k)$ denotes the ReTF between the two microphones. However, the RSP is no longer identified with the ReTF in the events when there are multiple sound sources.

5.2.3 Estimation of Relative Sound Pressure

Computing the relative sound pressure using a ratio between two microphone pressure contains non-negligible errors in noisy environments, especially when the pressure at the denominator in (5.5) is weak. This subsection overcomes this issue by presenting an alternative estimation of the relative sound pressure, where both the noiseless and noisy environments are taken into account.

Noiseless Environment

The original definition in (5.5) is equivalent to,

$$Q(\boldsymbol{x}_j, k) = \frac{P(\boldsymbol{x}_j, k)P^*(\boldsymbol{x}_o, k)}{|P(\boldsymbol{x}_o, k)|^2}. \tag{5.8}$$

Assuming the source signal is stationary or less dynamic over a short time period, we then represent (5.8) using,

$$Q(\boldsymbol{x}_j, k) = \frac{S_{p_j p_o}(k)}{S_{p_o p_o}(k)} \tag{5.9}$$

where

$$S_{p_o p_o}(k) = \mathbb{E}\left\{P(\boldsymbol{x}_o, k)P^*(\boldsymbol{x}_o, k)\right\} \tag{5.10}$$

denotes the power spectral density (PSD) of $P\boldsymbol{x}_o, k)$, $\mathbb{E}\{\cdot\}$ denotes the statistical expectation operator, and

$$S_{p_j p_o}(k) = \mathbb{E}\left\{P(\boldsymbol{x}_j, k)P^*(\boldsymbol{x}_o, k)\right\} \tag{5.11}$$

denotes cross PSD (CPSD) between $P(\boldsymbol{x}_j, k)$ and $P(\boldsymbol{x}_o, k)$.

Noisy Environment

Substituting the noisy sound pressure of (5.1) to (5.9), the noisy relative sound pressure follows,

$$\overline{Q}(\boldsymbol{x}_j, k) = \frac{S_{\overline{p}_j \overline{p}_o}(k)}{S_{\overline{p}_o \overline{p}_o}(k)} \tag{5.12}$$

where $S_{\overline{p}_j \overline{p}_o}(k)$ and $S_{\overline{p}_o \overline{p}_o}(k)$ represent the noisy CPSD and PSD, respectively. The noisy CPSD and PSD can be further simplified by substituting the noise and signal

components,

$$\begin{aligned} S_{\overline{p}_j\overline{p}_o}(k) &= S_{p_jp_o}(k) \\ S_{\overline{p}_o\overline{p}_o}(k) &= S_{p_op_o}(k) + S_{n_on_o}(k) \end{aligned} \tag{5.13}$$

where

$$S_{n_on_o}(k) = \mathbb{E}\left\{ n(\boldsymbol{x}_o, k) n^*(\boldsymbol{x}_o, k) \right\} \tag{5.14}$$

represents the PSD of the noise at the reference microphone. Note that (5.13) is due to the assumption that the source signal and incoherent noise signal are uncorrelated so that their CPSD between the microphone pair is zero. Substituting (5.13) to (5.12), we have the noisy relative sound pressure,

$$\overline{Q}(\boldsymbol{x}_j, k) = \frac{S_{p_jp_o}(k)}{S_{p_op_o}(k) + S_{n_on_o}(k)}. \tag{5.15}$$

Dividing (5.15) by (5.9), we derive the following relation between the noisy and noiseless relative sound pressure,

$$\overline{Q}(\boldsymbol{x}_j, k) = Q(\boldsymbol{x}_j, k)\rho(k) \tag{5.16}$$

where

$$\rho(k) = \frac{T(\boldsymbol{x}_o, k)}{T(\boldsymbol{x}_o, k) + 1} \tag{5.17}$$

only depends on the signal to noise ratio (SNR) at the origin of the array, i.e., $T(\boldsymbol{x}_o, k) = S_{p_op_o}(k)/S_{n_on_o}(k)$, and the dependency of $\rho(k)$ on $\boldsymbol{x}_o$ is omitted for convenience. Similar to the ReTF, the relative sound pressure, represented using the power spectral density between two microphones, is also robust to the noise. Section 5.4 presents proof confirming the relative sound pressure is less sensitive to noise than the sound pressure.

5.3 RMUSIC: Relative Sound Pressure Based MUSIC

This section outlines an approach to estimate multi-source DOAs using relative sound pressure based on the standard MUSIC algorithm framework.

5.3.1 Far-field Relative Sound Pressure

Substituting the sound pressure using plane waves modeling to (5.5), we have a linear representation of the relative sound pressure in a noiseless environment,

$$\begin{aligned} Q(\boldsymbol{x}_j, k) &= \frac{\sum_{q=1}^{Q} S_q(k) e^{i\boldsymbol{k}_q^T \boldsymbol{x}_j}}{\sum_{q=1}^{Q} S_q(k) e^{i\boldsymbol{k}_q^T \boldsymbol{x}_o}} = \frac{\sum_{q=1}^{Q} S_q(k) e^{i\boldsymbol{k}_q^T \boldsymbol{x}_j}}{\sum_{q=1}^{Q} S_q(k)} \\ &= \sum_{q=1}^{Q} \overline{s}_q(k) e^{i\boldsymbol{k}_q^T \boldsymbol{x}_j}, \qquad j = 1, \cdots, M \end{aligned} \tag{5.18}$$

where

$$\overline{s}_q(k) = \frac{S_q(k)}{\sum_{q=1}^{Q} S_q(k)} \tag{5.19}$$

denotes its relative component of the q-th source signal among all the sources. We rewrite (5.18) in the vector form,

$$Q(\boldsymbol{x}_j, k) = \mathbf{v}_j^T(k) \overline{\mathbf{s}}(k) \tag{5.20}$$

where $\overline{\mathbf{s}}(k)$ is a $L \times 1$ vector as,

$$\overline{\mathbf{s}}(k) \;=\; [\overline{s}_1(k), \overline{s}_2(k), \cdots, \overline{s}_q(k)]^T \tag{5.21}$$

and $\mathbf{v}_q(k)$ is the steering vector. By substituting (5.20) to (5.16), the noisy relative sound pressure is represented as,

$$\overline{Q}(\boldsymbol{x}_j, k) = \mathbf{v}_j^T(k) \overline{\mathbf{s}}(k) \rho(k). \tag{5.22}$$

In consideration of all the M microphones, we have a matrix form of (5.22),

$$\overline{\mathbf{Q}}(k) = \mathbf{V}(k)\overline{\mathbf{s}}(k)\rho(k) \tag{5.23}$$

where $\overline{\mathbf{Q}}(k)$ denotes the vector of noisy relative sound pressure for all the microphone channels, $\mathbf{V}(k)$ represents the steering matrix of (5.4), and $\rho(k)$ is the scalar of (5.17).

5.3.2 MUSIC Using Relative Sound Pressure

This subsection shows how to modify the MUSIC approach for multi-source localization, using the relative sound pressure in (5.23). Let us calculate the $M \times M$ correlation matrix of the noisy relative sound pressure,

$$\begin{aligned} \mathbf{S}_{\overline{\mathbf{Q}}}(k) &= \mathbb{E}\left\{\overline{\mathbf{Q}}(k)\overline{\mathbf{Q}}^{H}(k)\right\} \\ &= \mathbf{V}(k)\mathbf{R}_{\mathbf{S}}(k)\mathbf{V}^{H}(k) \end{aligned} \tag{5.24}$$

where

$$\mathbf{R}_{\mathbf{S}}(k) = \mathbb{E}\left\{\overline{\mathbf{s}}(k)\rho(k)\overline{\mathbf{s}}^{H}(k)\rho^{*}(k)\right\} \tag{5.25}$$

is a full rank matrix. In practice, we obtain the eigenvectors using a singular value decomposition of the covariance matrix,

$$\mathbf{S}_{\overline{\mathbf{Q}}}(k) = \begin{bmatrix} \overline{\mathbf{U}}_s & \overline{\mathbf{U}}_n \end{bmatrix} \begin{bmatrix} \overline{\mathbf{\Sigma}}_s & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \overline{\mathbf{U}}_s^{H} \\ \overline{\mathbf{U}}_n^{H} \end{bmatrix} \tag{5.26}$$

where the frequency dependency is also omitted for convenience. Note that, different from the traditional MUSIC method [18], the covariance matrix in (5.24) does not have an additive item corresponding to the noise's covariance matrix. Since the $\mathbf{R}_{\mathbf{S}}(k)$ is a full rank $L \times L$ matrix, the $M \times M$ matrix of $\mathbf{S}_{\overline{\mathbf{Q}}}(k)$ has L non-zero eigenvalues (i.e., $L \times L$ diagonal matrix of eigenvalues $\overline{\mathbf{\Sigma}}_s(k) = \mathrm{diag}(\sigma_1(k), \cdots, \sigma_L(k))$) and $M - L$ zero eigenvalues. The matrixes of $\overline{\mathbf{U}}_s(k)$ and $\overline{\mathbf{U}}_n(k)$ in (5.26) denote the subspaces corresponding to the $\overline{\mathbf{\Sigma}}_s(k)$ and zero eigenvalues, respectively. Above analysis shows that the proposed RMUSIC has a larger variation between its eigenvalues corresponding to subspaces of $\overline{\mathbf{U}}_s(k)$ and $\overline{\mathbf{U}}_n(k)$, respectively. The

greater difference between the sorted eigenvalues enables us to estimate the number of sound sources more easily, thus not requiring this prior knowledge anymore.

Above analysis only uses the recordings at the k-th frequency bin. **Algorithm 1** presents the steps of the algorithm considering a wide frequency band. We provide more detailed explanations on key steps of the algorithm below:

Calculate the relative sound pressure

Given the multi-channel measurements, the relative sound pressure of the microphone array is calculated using (5.12) at each individual frequency bin.

Estimate the number of sound sources

We first compute the covariance matrix of the relative sound pressure in (5.24) and then apply the singular value decomposition in (5.26) to compute the vector of eigenvalues, i.e., $\boldsymbol{\sigma}(k) = [\sigma_1(k), \cdots, \sigma_M(k)]^T$. For a wide frequency band, it is intuitive to compute the average eigenvalues as follows,

$$\overline{\boldsymbol{\sigma}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\sigma}(k) \tag{5.27}$$

where $\overline{\boldsymbol{\sigma}}$ denotes the average vector of eigenvalues over the K frequency bins of interest. Generally, an average vector reduces the computational complexity as it avoids repetitive processing at each frequency bin respectively. As analyzed above, the number of sound sources is counted based on the significant difference between eigenvalues in $\overline{\boldsymbol{\sigma}}$ as the eigenvalues associated with subspace of $\overline{\mathbf{U}}_n(k)$ are closer to zero (see an example in Figure 5.2).

Compute the pseudo-spectrum over space

For each frequency bin, we first calculate the subspace $\overline{\mathbf{U}}_n(k)$ and then use (5.28) to compute the pseudo-spectrum over the two-dimensional directional space. We search the steering vector $\mathbf{a}(k, y_s)$ in (5.28) over the sampled two-dimensional space, i.e., $\boldsymbol{\Phi} = \left\{(\vartheta_\ell, \varphi_\ell) : 0 < \vartheta_\ell \leq \pi, 0 < \varphi_\ell \leq 2\pi\right\}$. A higher space resolution to sample the space increases the localization accuracy while at the cost of a higher compu-

tational expense. Finally, we exploit the (5.29) for the average pseudo-spectrum over the wide frequency band of interest.

Achieve DOA estimations

Given the pseudo-spectrum over space, the multi-source DOA estimations is completed by searching the L significant peaks within the spectrum.

Algorithm 1: RMUSIC.

Input: Time-domain recordings.

Output: DOA estimations.

1. Transfer the recordings into STFT domain.
2. Calculate the relative sound pressure.
3. Estimate the number of sources via the average eigenvalues in (5.27).
4. **For** $k = 1, 2, \cdots$, do until finished:
   1). Calculate the covariance matrix $\mathbf{S_p}(k)$.
   2). Calculate the subspace $\overline{\mathbf{U}}_n^H(k)$ via SVD.
   3). Calculate the pseudo-spectrum over space,

$$M(k, y_s) = \frac{1}{||\overline{\mathbf{U}}_n^H(k)\,\mathbf{a}(k, y_s)||^2}. \tag{5.28}$$

5. Average the spectrum over a wide band,

$$\widetilde{M}(y_s) = \frac{1}{K}\sum_{k=1}^{K} M(k, y_s). \tag{5.29}$$

6. Search the L peaks of the spectrum by $\widetilde{M}$.

5.4 SHD-RMUSIC: Spherical Harmonics Domain RMUSIC

This section decomposes the proposed RMUSIC approach into the spherical harmonics domain, i.e., SHD-RMUSIC, allowing a frequency smoothing to de-correlate

the coherent source signal for improved localization accuracy.

5.4.1 Decompose Relative Sound Pressure into Spherical Harmonics Domain

The measured relative sound pressure over the microphone array, i.e., $\overline{Q}(\boldsymbol{x}_j, k)$, $j = 1, \cdots, M$, can be decomposed into the spherical harmonics domain using a set of orthogonal spatial functions [16],

$$\overline{Q}(\boldsymbol{x}_j, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \overline{\beta}_{nm}(k)\, j_n(kr)\, Y_{nm}(\theta_j, \phi_j) \tag{5.30}$$

where $n(\geqslant 0)$ and m are integers, $\overline{\beta}_{nm}(k)$ is the spherical harmonic coefficient of the relative sound pressure, $j_n(\cdot)$ is the spherical Bessel function, $N = \lceil kr \rceil$ is the truncated order of the soundfield [103] and $Y_{nm}(\theta, \phi)$ denotes the spherical harmonics function. The spherical harmonic coefficients, i.e., $\overline{\beta}_{nm}(k)$, characterizing/describing the soundfield in spherical harmonics domain, can be measured using this spherical microphone array (M discrete microphones),

$$\overline{\beta}_{nm}(k) = \frac{1}{j_n(kr)} \sum_{j=1}^{M} a_j \overline{Q}(\boldsymbol{x}_j, k)\, Y^*_{nm}(\theta_j, \phi_j). \tag{5.31}$$

Note that both (5.30) and (5.31) use the measurements of an open-sphere spherical microphone array. More details about the background knowledge of spherical harmonics decomposition are referred to Chapter 2.

Traditional spherical harmonics decomposition of the noisy sound pressure, used by the SHD-MUSIC approach [17], suffers from the "Bessel zero problem". This is due to the spherical Bessel function $j_n(kr)$, fed with a small input, approaches zero crossings [141]. As a result, the noise component in the measured spherical harmonic coefficients is greatly amplified. By contrast, it becomes a less serious issue by the spherical harmonics decomposition in (5.31) because the relative sound pressure is less sensitive to noise.

Algorithm 2: SHD-RMUSIC.

Input: Time-domain recordings.

Output: DOA estimates.

1. Transfer the recordings into STFT domain.
2. Calculate the relative sound pressure.
3. Calculate its spherical harmonics coefficients.
4. **For** $k = 1, 2, \cdots$, do until finished:
   Calculate the covariance matrix $\mathbf{S}_{\overline{\boldsymbol{\beta}}}(k)$.
5. Smoothed covariance matrix,

$$\widetilde{\mathbf{S}}_{\mathbf{p}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_{\overline{\boldsymbol{\beta}}}(k). \tag{5.32}$$

6. Estimate the number of sources via eigenvalues.
7. Calculate the subspace $\mathbf{U}_n$.
8. Calculate the pseudo-spectrum $\widetilde{M}\,(y_s)$.
9. Search the L peaks of the pseudo-spectrum.

5.4.2 SHD-RMUSIC with Frequency Smoothing

The plane waves modeling of the steering vector in (5.22), due to the q-th sound source, can also be decomposed into the spherical harmonics domain [120, 121],

$$e^{i\mathbf{k}_q^T \boldsymbol{x}_j} = \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n Y_{nm}^*(\Psi_q)\, j_n(kr)\, Y_{nm}(\theta_j, \phi_j). \tag{5.33}$$

Substituting (5.30) and (5.33) to (5.22), we derive the expression of the spherical harmonics coefficients of the noisy relative sound pressure,

$$\overline{\beta}_{nm}(k) = \mathbf{y}_{nm}(k)\overline{\mathbf{s}}(k)\rho(k) \tag{5.34}$$

where $\bar{\mathbf{s}}(k)$ is the vector of (5.21), and $\mathbf{y}_{nm}(k)$ is the steering vector at order n and degree m associating with all the sources,

$$\mathbf{y}_{nm}(k) = 4\pi \left[i^n Y_{nm}^*(\Psi_1), i^n Y_{nm}^*(\Psi_2), \cdots, i^n Y_{nm}^*(\Psi_Q) \right]. \tag{5.35}$$

Note that (5.34) only includes a single spherical harmonic mode. Combining all the cases up to the N-th order, we rewrite (5.34) in a matrix form,

$$\overline{\boldsymbol{\beta}}(k) = \mathbf{Y}(k)\bar{\mathbf{s}}(k)\rho(k) \tag{5.36}$$

where $\mathbf{Y}(k)$ denotes the $(N+1)^2 \times L$ steering matrix in the spherical harmonics domain,

$$\mathbf{Y}(k) = \left[\mathbf{y}_{00}(k), \mathbf{y}_{1,-1}(k), \ldots, \mathbf{y}_{NN}(k) \right]^T. \tag{5.37}$$

The correlation matrix of the noisy spherical harmonic coefficients over the time-varying source signal is,

$$\begin{aligned} \mathbf{S}_{\overline{\boldsymbol{\beta}}}(k) &= \mathbb{E}\left\{ \overline{\boldsymbol{\beta}}(k) \overline{\boldsymbol{\beta}}^H(k) \right\} \\ &= \mathbf{Y}(k) \mathbf{R}_{\mathbf{S}}(k) \mathbf{Y}^H(k) \end{aligned} \tag{5.38}$$

where the covariance matrix $\mathbf{R}_{\mathbf{S}}(k)$ and steering matrix $\mathbf{Y}(k)$ contain the frequency and angular components, respectively. The MUSIC approach assumes $\mathbf{R}_{\mathbf{S}}(k)$ to be a full rank matrix. However, this assumption hardly conforms to reality because the speech recordings from multiple speakers, especially in a reverberant enclosure, maybe coherent, i.e.,

$$rank \ \ \mathbf{R}_{\mathbf{S}}(k) < Q. \tag{5.39}$$

It is a common advantage that the frequency-dependent and angular-dependent components are decoupled in the spherical harmonics domain (see (5.38)). Hence, we de-correlate the coherent source signal by implementing the frequency smoothing that computes the smoothed covariance matrix as the average of covariance matrices at different frequency sectors [17],

$$\widetilde{\mathbf{S}}_{\mathbf{p}} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_{\overline{\boldsymbol{\beta}}}(k) = \mathbf{Y}(k) \widetilde{\mathbf{R}}_{\mathbf{S}}(k) \mathbf{Y}^H(k) \tag{5.40}$$

where

$$\widetilde{\mathbf{R}}_{\mathbf{S}}(k) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{R}_{\mathbf{S}}(k) \tag{5.41}$$

where K frequency bins are exploited. Finally, the smoothed covariance matrix is decomposed using a singular value decomposition and the pseudo-spectrum is calculated to complete the multi-source DOA estimations. **Algorithm 2** presents the general steps of SHD-RMUSIC method, which are derived similar to those in **Algorithm 1**.

5.5 Experimental Validation

This section evaluates the proposed MUSIC methods in diverse environments, using simulated recordings as well as real-life recordings measured in an acoustic lab at the Australian National University. The following experiments are implemented by using the detailed steps presented in both the **Algorithms 1** and **2**.

5.5.1 Simulation Setting

We generate the simulated recordings inside a $6 \times 4 \times 3$ m room. Several sound sources are simultaneously active within the enclosure and use the speech sentences randomly selected from the TIMIT database (lasting 4 seconds and re-sampled to 8 KHz). Then, we measure the incoming soundfield in the room using a simulated open-sphere spherical microphone array (with 32 microphones and a radius of 4.2 cm). We use an open spherical array for convenience, yet the developed theory is directly extendable for rigid arrays when the scattering is incorporated in the theory. Thereafter, we model the room characteristics from the sources to the microphone array, using an available room impulse response (RIR) generator (i.e., the same one used in Chapter 3). The time-domain recordings measured by the array are contaminated by the randomly generated noise signal at all the 32 microphones. We then use the STFT to transfer the recordings into the frequency domain. Thirty frequency bins ranging from 2700 Hz to 3600 Hz, which exactly measure the soundfield up to the 3rd order as $N = \lceil kr \rceil$ (i.e., 16 spherical harmonic modes in total), are used by the localization algorithms. When calculating

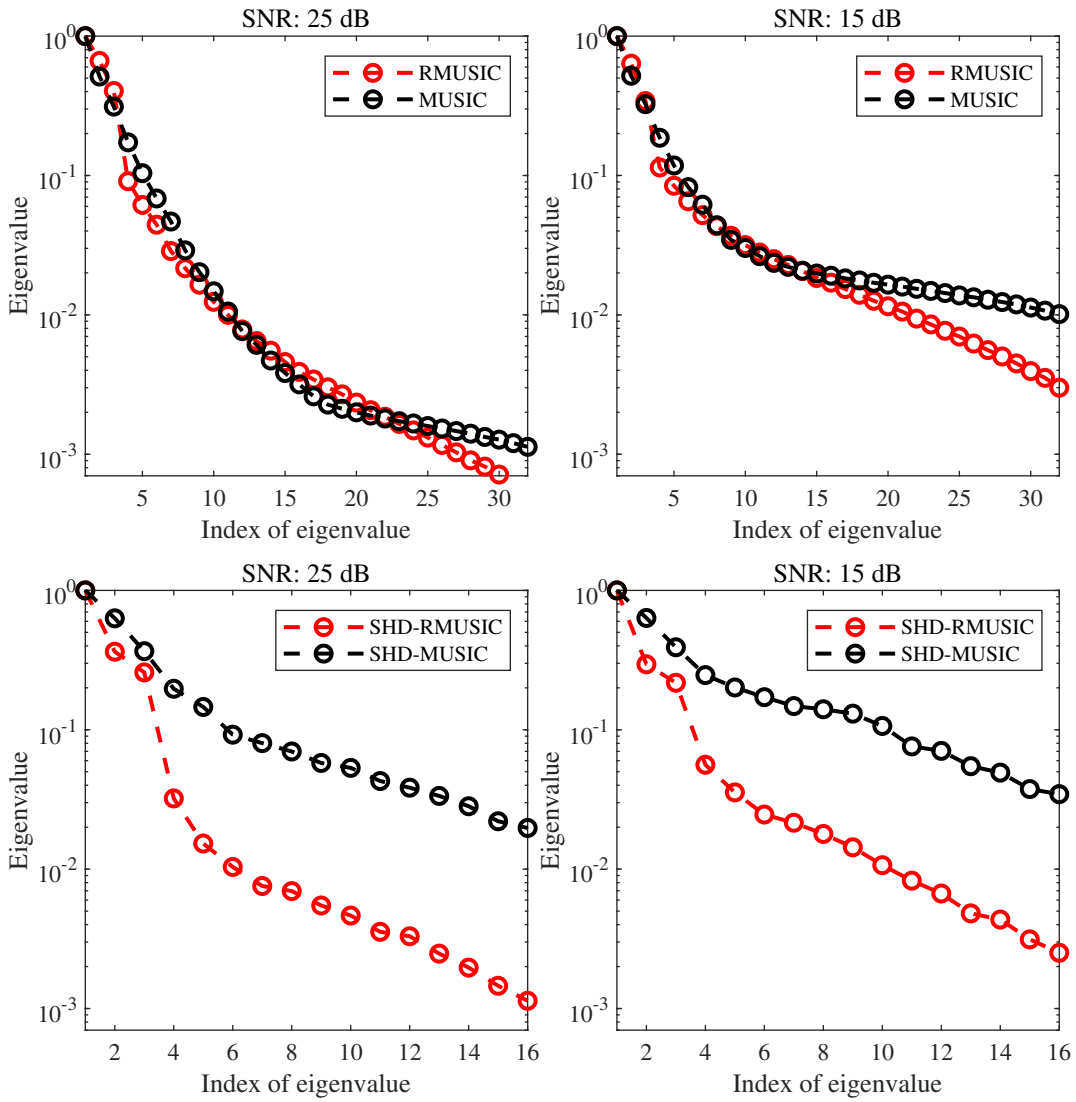


Figure 5.2: Normalized eigenvalues obtained via a singular value decomposition of the source signal's covariance matrix (room reverberations $T_{60} = 0.3$ s).

the relative sound pressure, we assume the speech signal is stationary over about 0.1 s, and utilize the Welch method to estimate the PSD and CPSD, with 0.016 s windows and 50% overlap.

5.5.2 Baseline Methods and Evaluation Metrics

We evaluate the proposed methods in comparison with three additional multi-source localization approaches, (i) the traditional sound pressure based MUSIC, (ii) SHD-MUSIC developed by Rafaely *et al.* [17], and (iii) a recently proposed approach we developed in [1] using the source feature of relative harmonic coefficients. For notational convenience, we abbreviate the four types of the investigated MUSIC approaches to 'RMUSIC', 'SHD-RMUSIC', 'MUSIC', 'SHD-MUSIC', respectively. By contrast, the other approach in [1] is significantly different from the MUSIC based methods as it exploits a pre-processing technique to detect the components where only a single source is active. All the localization algorithms here require to sample the two-dimensional space, i.e., $\mathbf{\Phi} = \Big\{(\vartheta_q, \varphi_q) : 0 < \vartheta_q \leq \pi, 0 < \varphi_q \leq 2\pi\Big\}$. A higher spatial resolution increases the accuracy of the DOA estimations while at the cost of a higher computational expense. Here, both the elevation and azimuth grid are uniformly divided into 90 samples so that there are 8100 samples in total.

The following experiments implement the algorithms up to $\mathcal{M}_{\text{tot}} > 1$ (i.e., the total number of tests) times to achieve more reliable results. Each testing uses the sound sources located at randomly selected DOAs. To fairly evaluate the algorithms, we use two qualitative metrics to measure the performance: (i) the success-ratio (SR):

$$\text{SR} = \frac{\mathcal{M}_{\text{suc}}}{\mathcal{M}_{\text{tot}}} \times 100\% \tag{5.42}$$

where $\mathcal{M}_{\text{suc}}$ denotes the number of cases that successfully detect all the L sound sources. A larger success-ratio indicates the algorithm has a stronger ability to localize all the sources in the environment. (ii) the mean absolute estimated error (MAEE/°) between their estimated and original source DOAs:

$$\text{MAEE} = \frac{1}{2LM_{\text{suc}}} \left(\sum_{m=1}^{M_{\text{suc}}} \sum_{q=1}^{Q} |\theta_{\text{ori}}^{m}(q) - \theta_{\text{est}}^{m}(q)| + |\phi_{\text{ori}}^{m}(q) - \phi_{\text{est}}^{m}(q)| \right) \tag{5.43}$$

which measures the average numerical accuracy over the $\mathcal{M}_{\text{suc}}$ successful testings.

5.5.3 Robustness Analysis

Before examining the localization accuracy, we show that the relative sound pressure is less sensitive to the noise, in comparison with the sound pressure. We first present a theoretical proof and then evaluate it using simulated measurements.

Theoretical proof

To evaluate the robustness, we use a distortion ratio between the clean and noisy signal of the j-th microphone,

$$\Gamma_{Q(\boldsymbol{x}_j,k)} = \left| \frac{\overline{Q}(\boldsymbol{x}_j,k) - Q(\boldsymbol{x}_j,k)}{Q(\boldsymbol{x}_j,k)} \right| \tag{5.44}$$

where $Q(\boldsymbol{x}_j,k)$ and $\overline{Q}(\boldsymbol{x}_j,k)$ denote the clean and noisy relative sound pressure, and $|\cdot|$ denotes the absolute operator. Intuitively, a larger value means more distortions. Substitute (5.9) and (5.15) into (5.44), its distortion ratio is simplified as,

$$\Gamma_{Q(\boldsymbol{x}_j,k)} = \left| \frac{S_{v_o v_o}(k)}{S_{p_o p_o}(k) + S_{v_o v_o}(k)} \right| = \frac{1}{T_{\boldsymbol{x}_o}(k) + 1} \tag{5.45}$$

where $T_{\boldsymbol{x}_o}(k)$ is the SNR at the array origin. The distortion ratio of the sound pressure at the j-th microphone is,

$$\Gamma_{P(\boldsymbol{x}_j,k)} = \left| \frac{\overline{P}(\boldsymbol{x}_j,k) - P(\boldsymbol{x}_j,k)}{P(\boldsymbol{x}_j,k)} \right| = \left| \frac{n(\boldsymbol{x}_j,k)}{P(\boldsymbol{x}_j,k)} \right|. \tag{5.46}$$

We discuss the robustness under the following scenarios,

- *Noise signal is not stronger than the source signal, i.e., SNR* $\geqslant 0$ *dB*: the following inequality holds,

$$\Gamma_{P(\boldsymbol{x}_j,k)} \geqslant \left| \frac{n(\boldsymbol{x}_j,k)}{P(\boldsymbol{x}_j,k)} \right|^2 > \frac{|n(\boldsymbol{x}_j,k)|^2}{|P(\boldsymbol{x}_j,k)|^2 + |n(\boldsymbol{x}_j,k)|^2}. \tag{5.47}$$

We divide the numerator and denominator by $|n(\boldsymbol{x}_j, k)|^2$,

$$\Gamma_{P(\boldsymbol{x}_j,k)} > \frac{1}{T_{\boldsymbol{x}_j}(k)+1} \tag{5.48}$$

in which $T_{\boldsymbol{x}_j}(k)$ denotes the SNR at the j-th microphone. It generally holds that the SNR within the array stays at similar levels, i.e., $T_{\boldsymbol{x}_j}(k) \approx T_{\boldsymbol{x}_o}(k)$. Thus, we derive,

$$\Gamma_{P(\boldsymbol{x}_j,k)} > \Gamma_{Q(\boldsymbol{x}_j,k)}. \tag{5.49}$$

- *Noise signal is stronger than the source signal, i.e., SNR* < 0 *dB*: the above inequality in (5.47) does not hold. However, this scenario is beyond the concern of this chapter as the localization algorithm to be developed is an unsupervised approach, which is unusable in unfavorable circumstances with such severe noise components.

Verification using recordings

Table 5.1: Distortions of relative sound pressure and pressure at varying SNR levels using the metric of (5.50)

Error/dB	SNR level (dB)				
Types	5	10	15	20	25
Error_{Γ_Q}	-0.61	-1.59	-3.03	-5.34	-7.85
Error_{Γ_P}	10.05	7.58	5.06	2.57	0.11

With the measured recordings at hand, we calculate the above distortions in (5.49) over the STFT domain as,

$$\begin{aligned} \text{Error}_{\Gamma_P} &= 10\log_{10}\left(\frac{1}{TFM}\sum_{t=1}^{T}\sum_{k=1}^{F}\sum_{j=1}^{M}\Gamma_{P_t(\boldsymbol{x}_j,k)}\right) \\ \text{Error}_{\Gamma_Q} &= 10\log_{10}\left(\frac{1}{TFM}\sum_{t=1}^{T}\sum_{k=1}^{F}\sum_{j=1}^{M}\Gamma_{Q_t(\boldsymbol{x}_j,k)}\right) \end{aligned} \tag{5.50}$$

where T, F, and M denote the total number of time, frequency bins, and microphones on the array, respectively, and t, k and j denote the corresponding index number. Table 5.1 presents the errors of the relative sound pressure and direct sound pressure, where various SNR levels are considered. Note that each value displayed in the table denotes the mean number over 5 tests. As expected, the measurements of both the relative sound pressure and pressure have increased distortions when the SNR level gradually decreases. However, we see that the distortion of the relative sound pressure is about 8 dB smaller than the sound pressure, indicating improved robustness to the noise.

5.5.4 Source Number Estimation

The significant difference between the sorted eigenvalues in (5.26) enables to count the sound source number. Before showing the localization accuracy, let us first count the unknown source number from the multi-source recordings. Figure 5.2 presents the normalized eigenvalues using all the four types of MUSIC based methods. Note that each value denotes the mean number over the wide frequency band. The environment originally contains three sound sources whose elevation and azimuth are $(149°, 259°)$, $(30°, 68°)$, $(95°, 101°)$, respectively. The dimensions of eigenvalues for both the RMUSIC and MUSIC are 32, i.e., the total number of microphones. By contrast, the dimensions for SHD-RMUSIC and SHD-MUSIC equal to the total number of spherical harmonics modes (i.e., 16). We easily observe that the eigenvalues of RMUSIC have a relatively large difference between the 3rd and 4th one, indicating the source number is 3. However, the gap between the 3rd and 4th eigenvalues by the sound pressure is less obvious. This phenomenon is more significant in the spherical harmonics domain, where the proposed SHD-RMUSIC has greater variations between the 3rd and 4th eigenvalues. We point out the proposed methods enable the source number counting in a typical environment. However, we cannot ensure an accurate counting at lower SNR conditions (e.g., 5 dB), thus the prior knowledge of the sound source number, at extremely low SNR levels, is still required.

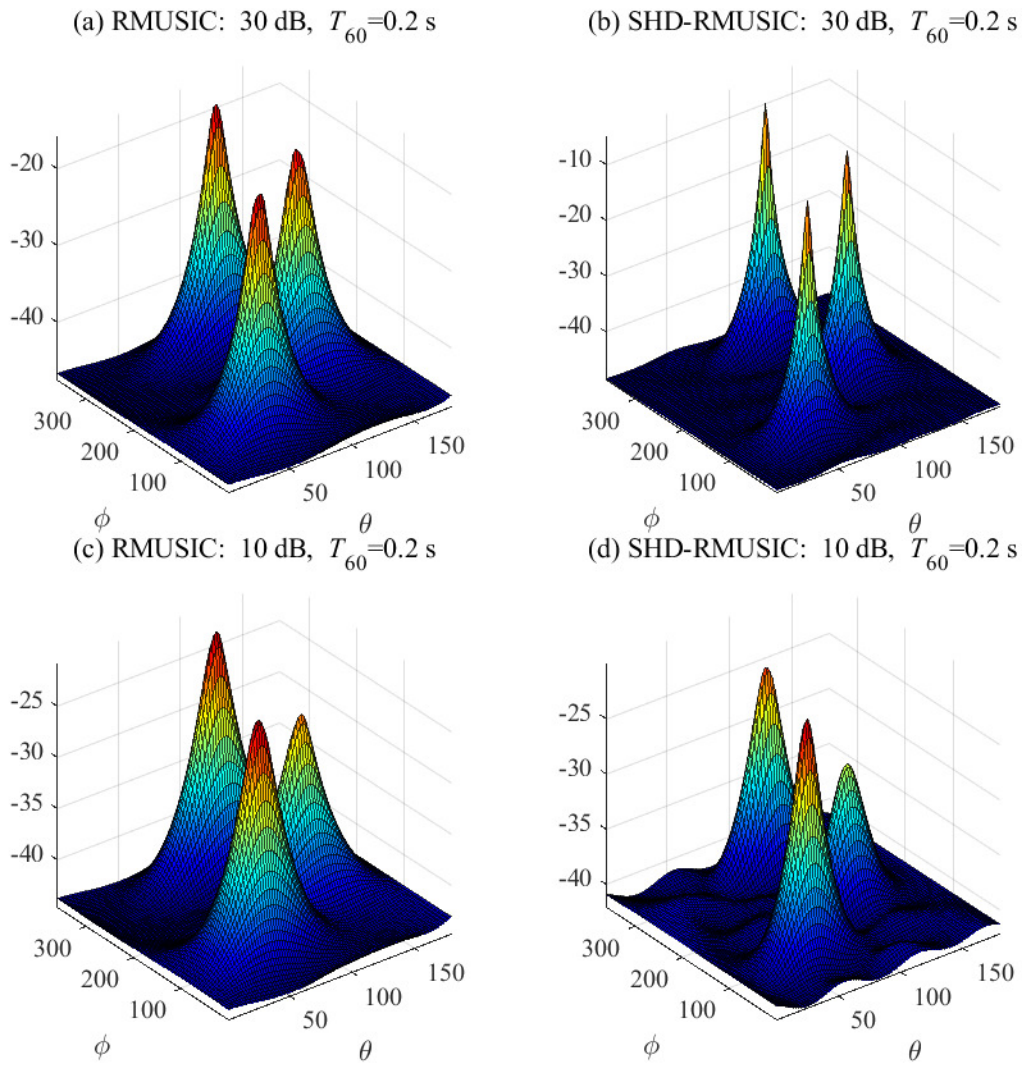


Figure 5.3: Pseudo-spectrum of three simultaneous sound sources using the proposed methods when $T_{60} = 0.2$ s.

5.5.5 DOA Estimations Under Various Scenarios.

This subsection implements the proposed DOA estimations under different scenarios, and then analyzes the performance.

- *Scenario 1* - different room reverberation and SNR levels in Fig. 5.3 and Fig. 5.4: We simulate three sound sources whose elevation and azimuth are $(107°, 303°)$

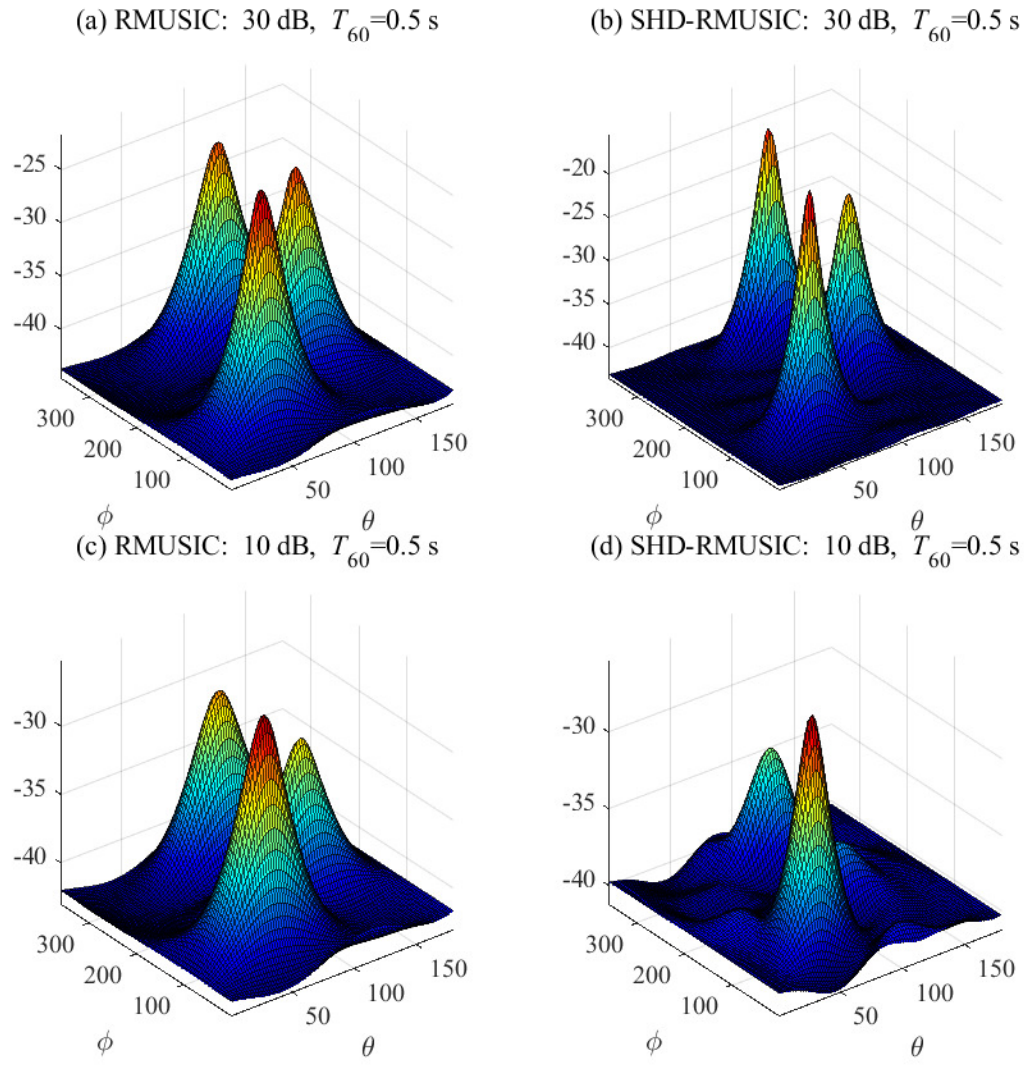


Figure 5.4: Pseudo-spectrum of three simultaneous sound sources using the proposed methods when $T_{60} = 0.5$ s.

and $(60°, 93°)$, $(123°, 173°)$. We measure the multi-source recordings considering different acoustic environments and then exhibit the performance by directly plotting their pseudo-spectrum. Under all the scenarios, the method using RMUSIC performs with three obvious peaks, i.e., the detected source DOAs. By contrast, the proposed SHD-RMUSIC has sharper peaks in a typical environment. However, in a complex environment, i.e., the SNR is 10 dB and $T_{60} = 0.5$ s (see

Fig. 5.4 (d)), the proposed SHD-RMUSIC fails to localize the three sources. The reason is as follows: the SHD-RMUSIC is more sensitive to the noise than the RMUSIC, due to the aforementioned "Bessel zero problem".

- *Scenario 2* - adjacent sound sources: In practical conditions, the sound sources may propagate from adjacent directions. Here, we simulate three adjacent sound sources whose elevation and azimuth are $(60°, 101°)$, $(88°, 145°)$ and $(91°, 117°)$, respectively, i.e., the sources elevation and azimuth only differ by $10°$ to $15°$ between the adjacent sources. The room reverberation level is set at $T_{60} = 0.2$ s and the SNR level is 30 dB. Figure 5.5 plots the pseudo-spectrum of both the proposed approaches. Under this challenging scenario, the RMUSIC fails to distinguish the three sources. By contrast, the method using SHD-RMUSIC

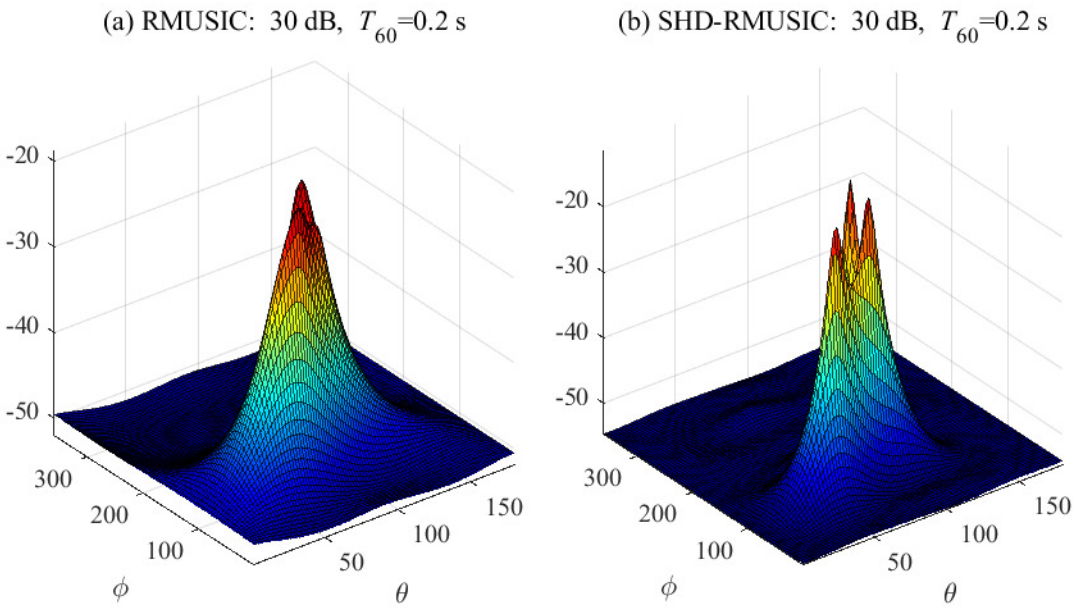


Figure 5.5: Pseudo-spectrum of three adjacent sound sources.

is still capable to detect three peaks. The sharper peaks are claimed to be an advantage by the SHD-RMUSIC because of the frequency smoothing technique to de-correlate the coherent source signal [17].

- *Scenario 3* - impact of source to microphone distance: The proposed methods assume the soundfield follows a far-field scenario. Hence, it is a necessity to investigate whether the multiple sources to microphone distance has an impact on the proposed methods. For a wide range of distances, we increase the dimension of the room to $10 \times 8 \times 6$ m for the length, width, and height, respectively. We simulate three sound sources whose elevation and azimuth are $(40°, 24°)$, $(16°, 250°)$, $(145°, 218°)$, respectively. The microphone array is still placed in the middle of the room, however, we set the sources at varied positions, with distances ranging from 0.5 m to 2.5 m to the microphone array. Note that the evaluations are implemented in a typical environment where the $T_{60} = 0.3$ s and the SNR level is 30 dB. For the varied distances, the MAEE using both the methods remains at about 1 degree, confirming the far-field assumption generally holds.

5.5.6 Comparison with Traditional MUSIC Methods

Let us evaluate the proposed MUSIC approaches in comparison with the traditional ones. Note that the traditional SHD-MUSIC in [17] uses a frequency smoothing technique as well. In the sequel, we implement the four MUSIC approaches using fifty repetitive measurements where the multiple sound sources propagate from randomly selected DOAs.

Although the proposed algorithms assume a free field propagation, we expect them to be robust enough in the typical reverberate environments. Table 5.2 displays their performance using both the metrics of SR in (5.42) and MAEE in (5.43). Different reverberant environments, whose T_{60} ranges from 0.2 s to 0.6 s, are examined. A higher room reverberation level indicates a larger multi-path distortion caused by the reflections from the multiple sources so that the source recordings are less incoherent. Since the proposed algorithms assume a far-field scenario where only the direct-path recordings are considered, we observe the localization accuracy degrades at a higher reverberation level. After that,

Table 5.2: Multi-source localization error under various reverberation levels where the SNR is 25 dB.

SR/MAEE°	T_{60} (s)		
Methods	0.2	0.4	0.6
MUSIC	0.90/1.23	0.82/2.06	0.72/3.17
RMUSIC	0.96/1.08	0.84/1.68	0.74/2.44
SHD-MUSIC	0.96/1.28	0.94/2.04	0.92/2.48
SHD-RMUSIC	0.98/1.14	0.96/1.41	0.92/2.06

Table 5.3: Multi-source localization error under various SNR levels where $T_{60} = 0.3$ s.

SR/MAEE°	SNR level (dB)		
Methods	10	20	30
MUSIC	0.84/1.69	0.84/1.58	0.84/1.57
RMUSIC	0.84/1.81	0.88/1.55	0.88/1.50
SHD-MUSIC	0.78/3.49	0.94/1.65	0.94/1.59
SHD-RMUSIC	0.90/3.74	0.98/1.72	0.98/1.45

we evaluate the algorithms under various noisy conditions. Table 5.3 presents their performance where the SNR level ranges from 10 dB to 30 dB. We recognize degraded localization accuracy when the SNR level gradually decreases. In most cases, we observe improved robustness of the RMUSIC as well as SHD-RMUSIC, in comparison with the MUSIC and SHD-MUSIC. To conclude, the evaluations under diverse environments, confirm the superiority of the proposed approaches over the traditional ones. Especially, the proposed SHD-RMUSIC outperforms the others under almost all the scenarios, achieving a success-ratio higher than 90% with an MAEE less than 4 degrees.

The proposed methods achieve improved localization accuracy at the cost of larger computational complexity. The increased expense is due to the calculations of the relative sound pressure. For validations, we measured their complexity by directly recording the time cost over ten repetitive cases, using a Matlab implementation on a standard desktop (CPU Intel Core i7-4790 Quad 3.6 GHz, RAM 16 GB). Based on our measurements lasting 4 s long, the time cost by the MU-

SIC and SHD-MUSIC approach is 2.8 s and 5.6 s, respectively. By contrast, the average time cost by the RMUSIC and SHD-RMUSIC approach is 4.7 s and 7.7 s, respectively. Both SHD-MUSIC and SHD-RMUSIC approaches require more time than MUSIC and RMUSIC methods because the process of transforming microphone signals into the spherical harmonic domain is costly. Although the proposed MUSIC methods require an additional time cost, it causes little influence on the algorithms because the MUSIC approaches are currently mainly used under off-line processing scenarios.

5.5.7 Comparison with the Multi-source Localization Technique Given in [1]

Table 5.4: Multi-source localization error using different source numbers.

SR/MAEE/°	Number of sources		
Methods	2	3	4
Baseline	0.90/4.22	0.88/3.33	0.82/2.80
RMUSIC	0.96/1.56	0.88/1.50	0.64/1.69
SHD-RMUSIC	1.00/1.25	0.96/1.36	0.88/2.17

The proposed MUSIC approaches use simultaneous multi-source recordings. This subsection compares the proposed methods with the other type of multi-source localization techniques, requiring detection of single-source components [1, 24]. A recently proposed method in [1] is taken as the baseline. This approach generally consists of two stages. In the first stage, it implements the pre-processing step to detect the single-source STFT bins. After that, it implements a single source localization given the detected single-source STFT bins. Aforementioned evaluations in subsection 5.5.5 and 5.5.6 only use source recordings generated by three sound sources. Here, different source numbers are considered. Table 5.4 presents the MAEE of all the algorithms where the $T_{60} = 0.3$ s and SNR is 20 dB. The results displayed are also computed using fifty repetitive measurements. As expected, the success ratio gradually decreases when the number of sound sources gradually increases. Particularly, the proposed RMUSIC degrades severely when

the source number is 4. This is because it hardly distinguishes the adjacent sources when there exists a larger number of sources in the environment. The baseline approach also degrades given a larger number of sound sources because there remain fewer single source frames/bins available for single source localization. From the results, we recognize that the proposed algorithms, especially the SHD-RMUSIC, outperform the baseline approach under most of the scenarios.

5.5.8 Verification Using Real Recordings

In this subsection, we use real-life recordings to validate the effectiveness of the proposed algorithms under practical scenarios. The experiments are implemented in the acoustic lab of Australian National University, whose room dimensions are $[6.7, 3.6, 2.8]$ for the length, width, and height, respectively. We place a commercial EigenMike at the corner of the room, where the average reverberation time is around $T_{60} = 250$ ms. The algorithms use the same parameters setting like that in the above simulations to process the real-life recordings. Before examining the localization accuracy, we first implement the source number counting by calculating the eigenvalues using both the proposed algorithms. Both 3 and 4 source scenarios are taken into account. Figure 5.6 presents the examples of the sorted eigenvalues due to three and four sound sources, respectively, where the source number can be easily counted by the sorted eigenvalues. We then evaluate their localization accuracy. For both the 3 and 4 source scenarios, ten cases of randomly selected real loudspeakers are used for the practical measurements. Figure 5.7 presents their MAEE calculated by (5.43). The average MAEE over the ten cases is about 5 degrees, indicating the proposed methods succeed in estimating the source's DOAs. Note that the accuracy using real-life recordings have larger errors than that using simulated recordings. This is due to the unavoidable fact that the practical measurements contain some non-negligible errors, including the measuring errors, and deviations in the loudspeaker positions.

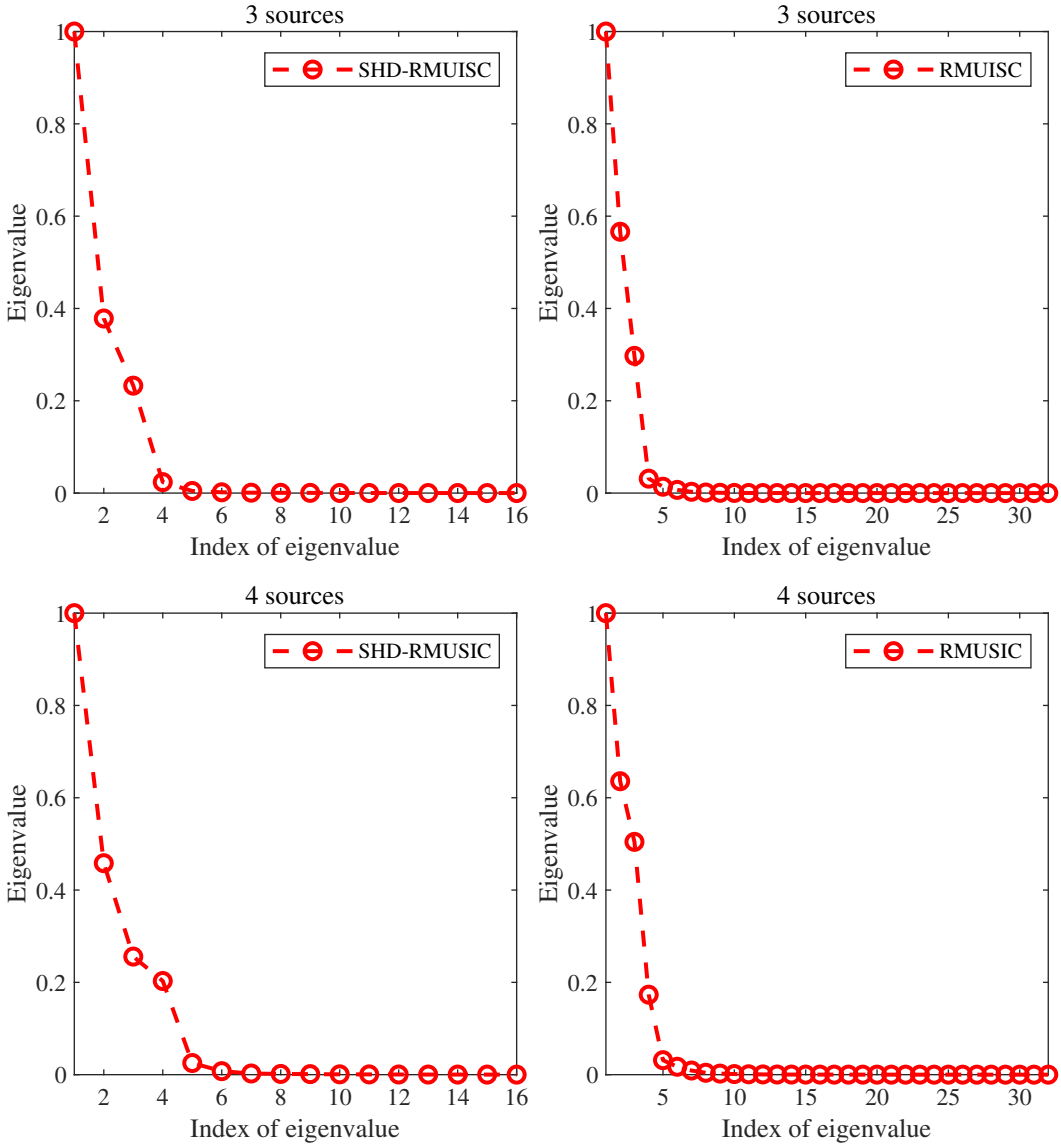


Figure 5.6: Normalized eigenvalues obtained using the real recordings. The two sub-figures on the left and right side correspond to 3 and 4 sources, respectively.

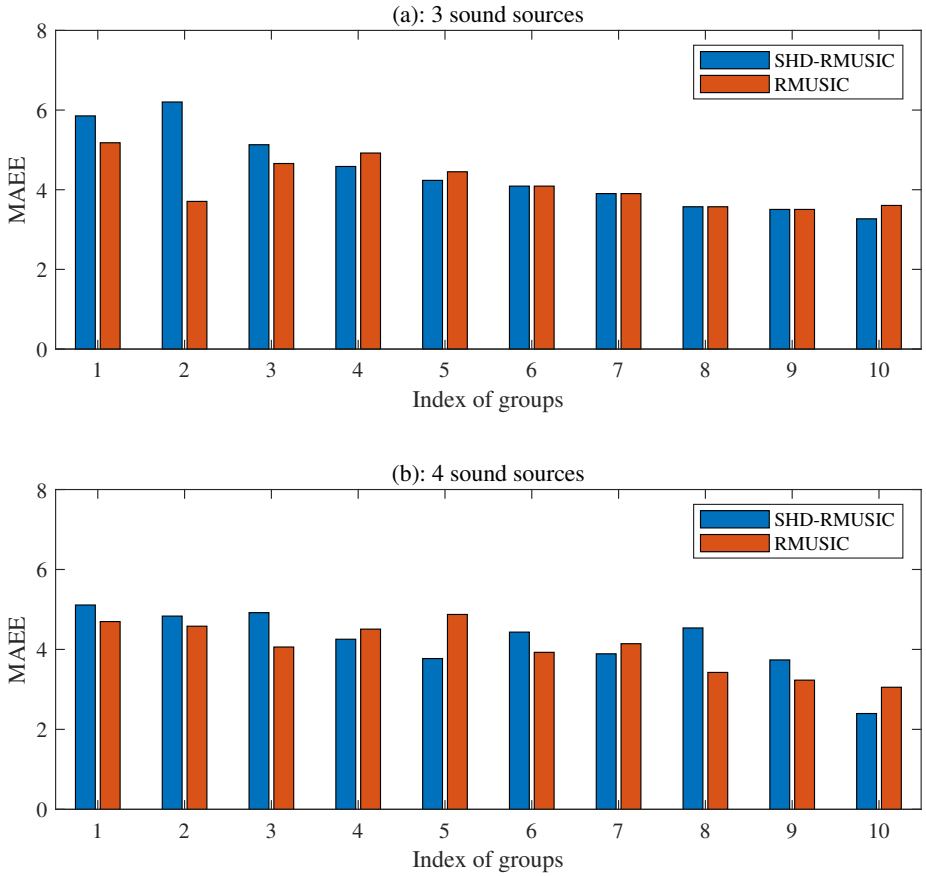


Figure 5.7: MAEE of the proposed approaches using real-life recordings.

5.6 Summary

This chapter addressed the problem of multi-source DOA estimations using simultaneous recordings by developing a relative sound pressure based MUSIC. We also decomposed the proposed algorithm into the spherical harmonics domain where a frequency smoothing technique is used to de-correlate the coherent source signals for improved accuracy. Extensive evaluations in diverse environments, using both simulated and real-life recordings, confirmed improved accuracy in comparison with the traditional approaches, at the cost of slight computational expense. Although some progress has been achieved, some inherent drawbacks are summarized: (i)

Lower SNR level: current approaches still cannot achieve satisfying performance at lower SNR conditions (i.e., 5 dB and lower); (ii) Room reverberations: the developed approaches have not taken the acoustic reflections into account, thus are unusable in strong reverberant environments. A potential solution is to develop a relative sound pressure based source feature and then feed it as the inputs of learning-based multi-source DOA estimators for improved performance under severely noisy and reverberant environments. In addition, another future direction is to develop a method that automatically detects the source number based on the sorted eigenvalues.

5.7 Related Publications

This chapter's work has ever been published/submitted in the following journal papers and conference proceedings.

- Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe, "Multiple source direction of arrival estimations using relative sound pressure based MUSIC", *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 29, pp. 253-264, 2021.

Chapter 6

Modeling Characteristics of Real Loudspeakers Using Various Acoustic Models

Overview: *The accuracy of the sound source localization algorithms, used in practice, are strongly influenced by the inherent characteristics of the commercial loudspeakers. This chapter analyzes such characteristics of loudspeakers by deriving equivalent theoretical acoustic models. Several acoustic models are investigated, including plane waves decomposition, point source decomposition, and mixed source decomposition. Each proposed model employs three effective sparse decomposition algorithms for optimized solutions, including iteratively reweighted least squares (IRLS), matching pursuit (MP), and least absolute shrinkage and selection operator (LASSO). A successful model shall enable the prediction of the soundfield outside the original recording region. Therefore, we validate the effectiveness of the models by comparing the simulated soundfield with secondary measurements obtained beyond the original area. Experimental results have confirmed that both the plane wave and mixed source model achieve promising performance concerning the proposed metrics.*

6.1 Introduction

In the above chapters, several sound source localization algorithms have been proposed. Although the techniques generally achieve improved performance in comparison with the baseline approaches, there still exist some factors with negative impacts on the localization accuracy. One issue that needs attention shall be the assumption of omnidirectional behavior of the sound sources, which is actually not the case with commercial loudspeakers. Up to the best knowledge of the author, there are very few techniques that take such considerations into source localization techniques, attempting to be technically accurate and precise. Note that, apart from source localization, this is a common issue suffered by other spatial acoustic processing techniques. For example, spatial soundfield reproduction uses an array of loudspeakers to create an immersive soundfield over a predefined spatial region so that listeners within the area can experience a virtual but realistic replication of the original soundfield [2, 142, 161]. Current soundfield reproduction systems, such as [106, 162, 163], also assume the omnidirectional behavior of sound sources. However, the inconsistency with the true characteristics of commercial loudspeakers leads to less immersive performance in practice.

This chapter conducts original research to model the characteristics of real loudspeakers using a set of equivalent source models over a wide frequency band. Apart from the common plane wave model, we explore an equivalent point source decomposition model with various radii and a mixed acoustic model that combines both the plane wave and point source decompositions. Since loudspeakers are modeled individually, the resulting incident field at the listening area is inherently sparse, especially in terms of incident direction. Therefore, the aforementioned equivalent source models can be further optimized by exploiting the feature of spatial sparsity. Thus, each acoustic model analyzes a range of sparsity exploitation algorithms [164], including iteratively re-weighted least squares (IRLS), matching pursuit (MP), and least absolute shrinkage and selection operator (LASSO). A successful theoretical model of the loudspeaker shall enable the ability to predict its incident field outside of the original measurements. Hence, we validate the effectiveness of proposed acoustic models by comparing the simulated and measured sound pressure over an extended area. Extensive experiments using real-life

recordings are conducted to validate the effectiveness of the proposed models.

6.2 Problem Formulation

In this section, we describe the soundfield produced by a loudspeaker, observed in a listening area, and then formulate the problem to be addressed. Typically, any arbitrary soundfield at a point $\boldsymbol{x} = (r, \theta, \phi)$ within a spherical listening region of radius R can be decomposed into modal domain [16] by,

$$P(\boldsymbol{x}, k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \alpha_{nm}(k)\, j_n(kr)\, Y_{nm}(\theta, \phi) \tag{6.1}$$

where $N = \lceil kR \rceil$ indicates the order of soundfield [165], $j_n(\cdot)$ stands for spherical Bessel functions, $Y_{nm}(\theta, \phi)$ is the spherical harmonics function with order n and degree m and $\alpha_{nm}(k)$ represents the spherical harmonic coefficients.

Traditionally, we assume that the loudspeakers act as an omni-directional point sources [165, 166] so that $\alpha_{nm}(k)$ due to a loudspeaker located at (r_s, θ_s, ϕ_s) can be calculated as,

$$\alpha_{nm}(k) = 4\pi i k h_n(kr_s) Y_{nm}(\theta_s, \phi_s) \tag{6.2}$$

where $h_n(\cdot)$ denotes the spherical Hankel function. However, the resulting harmonic coefficients in (6.2) are not accurate for commercial loudspeakers used in the real scenario as they are non-ideal speakers. Thus, the problem addressed is to propose and compare various acoustic models to model the soundfield coefficients $\alpha_{nm}(k)$ in (6.2) due to a real commercial loudspeaker by a limited number of measurements and predict the reproduced sound field over an extended area.

6.3 Acoustic Source Models

This section discusses the proposed equivalent source models of plane wave, point source and the mixed source model in details, respectively.

6.3.1 Plane Wave Modeling

Suppose we can represent an equivalent soundfield due to a loudspeaker by a finite number of planes waves arriving from an equiangular grid over all 3D directions,

$$P(\boldsymbol{x}, k) = \int s(\widehat{\boldsymbol{y}}, k) e^{ik\widehat{\boldsymbol{y}}\cdot\boldsymbol{x}} d\widehat{\boldsymbol{y}} \tag{6.3}$$

where $s(\widehat{\boldsymbol{y}}, k)$ is the complex weight of the plane wave arriving from the direction $\widehat{\boldsymbol{y}}$. Instead of considering infinite plane waves arriving from all directions, in practice, we only consider a finite number of planes waves arriving from an equiangular grid over 3D directions. Then, we can approximate (6.9) by

$$P(\boldsymbol{x}, k) \approx \sum_{l=1}^{L} s(\widehat{\boldsymbol{y}}_l, k) e^{ik\widehat{\boldsymbol{y}}_l\cdot\boldsymbol{x}} \tag{6.4}$$

where L denotes the source number. Using Gegenbauer expansion [167], we can write its decomposition in modal domain,

$$e^{ik\widehat{\boldsymbol{y}}\cdot\boldsymbol{x}} = \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi i^n Y_{nm}^*(\widehat{\boldsymbol{y}}) j_n(kr)\, Y_{nm}(\theta, \phi). \tag{6.5}$$

By substituting (6.5) into (6.4) and equaling to (6.1), the equivalent harmonic coefficients $\alpha_{nm}(k)$ can be expressed as,

$$\alpha_{nm}(k) = \sum_{l=1}^{L} 4\pi i^n Y_{nm}^*(\widehat{\boldsymbol{y}}_l) s(\widehat{\boldsymbol{y}}_l, k). \tag{6.6}$$

which relates the spherical harmonic coefficients $\alpha_{nm}(k)$ that characterise the loudspeaker to an equivalent set of plane wave weights $s(\widehat{\boldsymbol{y}}_l, k)$, $l = 1, \cdots, L$. Finally, we can write (6.6) in matrix form as,

$$\boldsymbol{\alpha} = \boldsymbol{H}_{pw} \boldsymbol{s}_{pw} \tag{6.7}$$

where $\boldsymbol{\alpha} = [\alpha_{00}, \cdots, \alpha_{NN}]^T$, $\boldsymbol{s}_{pw} = [s(\widehat{\boldsymbol{y}}_1, k), \cdots, s(\widehat{\boldsymbol{y}}_L, k)]^T$ and

$$\boldsymbol{H}_{pw} = 4\pi i^n \begin{bmatrix} Y_{00}^*(\widehat{\boldsymbol{y}}_l) & \cdots & Y_{00}^*(\widehat{\boldsymbol{y}}_L) \\ \cdots & \cdots & \cdots \\ Y_{NN}^*(\widehat{\boldsymbol{y}}_l) & \cdots & Y_{NN}^*(\widehat{\boldsymbol{y}}_L) \end{bmatrix}. \tag{6.8}$$

6.3.2 Point Source Modeling

Suppose we can represent an equivalent soundfield due to a loudspeaker by an infinite number of point sources that lie on the surface of a sphere with radius of r_p,

$$P(\boldsymbol{x}, k) = \int s(\widehat{\boldsymbol{y}}_m, k) e^{ik||\widehat{\boldsymbol{y}}_m - \boldsymbol{x}||_2} / ||\widehat{\boldsymbol{y}}_m - \boldsymbol{x}||_2 d\widehat{\boldsymbol{y}} \tag{6.9}$$

where $s(\widehat{\boldsymbol{y}}_m, k)$ is the complex weight of the point source arriving from the direction $\widehat{\boldsymbol{y}}_m$. Similarly, we propose to employ M discrete point sources to realize the modeling by approximating the recorded sound pressure,

$$P(\boldsymbol{x}, k) \approx \sum_{m=1}^{M} s(\widehat{\boldsymbol{y}}_m, k) e^{ik||\widehat{\boldsymbol{y}}_m - \boldsymbol{x}||_2} / ||\widehat{\boldsymbol{y}}_m - \boldsymbol{x}||_2. \tag{6.10}$$

Various setting of radius r_p makes a difference for the performance and its impact will be investigated in experiments. With Gegenbauer expansion [167], it can be decomposed as,

$$\frac{e^{ik||\widehat{\boldsymbol{y}} - \boldsymbol{x}||_2}}{||\widehat{\boldsymbol{y}} - \boldsymbol{x}||_2} = \sum_{n=0}^{N} \sum_{m=-n}^{n} 4\pi ikh_n(kr_p) Y_{nm}^*(\widehat{\boldsymbol{y}}) j_n(kr)\, Y_{nm}(\theta, \phi). \tag{6.11}$$

By substituting (6.11) into (6.10) and equaling to (6.1), we obtain,

$$\alpha_{nm}(k) = \sum_{m=1}^{M} 4\pi ikh_n(kr_p) Y_{nm}^*(\widehat{\boldsymbol{y}}_m) s(\widehat{\boldsymbol{y}}_m, k). \tag{6.12}$$

Represent (6.12) in matrix form as,

$$\boldsymbol{\alpha} = \boldsymbol{H}_{ps} \boldsymbol{s}_{ps} \tag{6.13}$$

where $\boldsymbol{\alpha} = [\alpha_{00}, \cdots, \alpha_{NN}]^T$, $\boldsymbol{s}_{ps} = [s(\widehat{\boldsymbol{y}}_1, k), \cdots, s(\widehat{\boldsymbol{y}}_M, k)]^T$ and

$$\boldsymbol{H}_{ps} = 4\pi i k h_n(kr_s) \begin{bmatrix} Y_{00}^*(\widehat{\boldsymbol{y}}_1) & \cdots & Y_{00}^*(\widehat{\boldsymbol{y}}_M) \\ \cdots & \cdots & \cdots \\ Y_{NN}^*(\widehat{\boldsymbol{y}}_1) & \cdots & Y_{NN}^*(\widehat{\boldsymbol{y}}_M) \end{bmatrix}. \tag{6.14}$$

6.3.3 Mixed Source Modeling

This subsection proposes the mixed source model that exploits and combines both the plane wave and point source efficiently and models the recorded harmonic coefficients $\alpha_{nm}(k)$ in a joint way.

$$\begin{aligned} \alpha_{nm}(k) &= \sum_{l=1}^{L} 4\pi i^n Y_{nm}^*(\widehat{\boldsymbol{y}}_l) s(\widehat{\boldsymbol{y}}_l, k) \\ &\quad + \sum_{m=1}^{M} 4\pi i k h_n(kr_s) Y_{nm}^*(\widehat{\boldsymbol{y}}_m) s(\widehat{\boldsymbol{y}}_m, k). \end{aligned} \tag{6.15}$$

Combination of (6.7) and (6.13) leads to matrix from of (6.15),

$$\boldsymbol{\alpha} = \boldsymbol{H}_{ms} \boldsymbol{s}_{ms} \tag{6.16}$$

where $\boldsymbol{\alpha} = [\alpha_{00}, \cdots, \alpha_{NN}]^T$, $\boldsymbol{H}_{ms} = [\boldsymbol{H}_{pw}\ \boldsymbol{H}_{ps}]$ and $\boldsymbol{s}_{ms} = [\boldsymbol{s}_{pw}\ \boldsymbol{s}_{ps}]^T$.

To combine the two parts fairly, the grid of directions for the plane wave and point source shall be set in the same manner so that $L = M$. Moreover, for the sake of fair possibility to be selected, the amplitude of the plane wave and point source from the same direction $\widehat{\boldsymbol{y}}$ ought to be equal. Therefore, the radius r_s of point source within the mixed model is set for each frequency that follows $|kh_n(kr_s)| = 1$.

6.4 Sparse Decomposition

This section attempts to seek optimized or desired solutions for the acoustic models formulated in Section 6.3. Given (6.7), (6.13) and (6.16), traditional least square

methods can provide accurate solutions while it has a tendency to spread the components of $\boldsymbol{\alpha}$ among a large number of source candidates in $\boldsymbol{H}$. Since we mainly consider the sound field generated by a single loudspeaker, we are to exploit the spatial sparsity feature by adding a sparse constraint using ℓ_p norm to the vector of driving signal $\boldsymbol{s}$ shown in (6.17) as it manages to accomplish the modeling using only a small number of source candidates.

$$\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_p^p, \text{ s.t. } \boldsymbol{H}\boldsymbol{s} = \boldsymbol{\alpha}. \tag{6.17}$$

This chapter investigates three effective sparse algorithms, e.g. IRLS, MP, and LASSO, to solve the sparse problem in (6.17) and introduce each of them in brief. Note that all the subscripts in this section are abandoned for the sake of generality.

IRLS; this algorithm [168, 169] replaces the ℓ_p norm objective function in (6.17) by a form of weighted ℓ_2 norm.

$$\min_{\boldsymbol{s}} \sum_{i=1}^{M} \boldsymbol{w}_i \boldsymbol{s}_i^2, \text{ s.t. } \boldsymbol{H}\boldsymbol{s} = \boldsymbol{\alpha} \tag{6.18}$$

where $\boldsymbol{w}_i = |\boldsymbol{s}_i^{(m-1)}|^{p-2}$ and the driving signal for the next iterate $\boldsymbol{s}^{(m)}$ can be given explicitly,

$$\boldsymbol{s}^{(m)} = Q_m \boldsymbol{H}^T (\boldsymbol{H} Q_m \boldsymbol{H}^T)^{-1} \boldsymbol{\alpha} \tag{6.19}$$

where Q_m is diagonal matrix with entries $1/\boldsymbol{w}_i = |\boldsymbol{s}_i^{(m-1)}|^{2-p}$.

MP; with a proper initialization, the MP conducts in iterative and greedy procedures to select the m-th column that has maximally inner product with current residual $R^m\boldsymbol{\alpha}$.

$$\boldsymbol{h}^{(m)} = arg \max_{\boldsymbol{h} \in \boldsymbol{H}} |\boldsymbol{h}^T (R^m \boldsymbol{\alpha})| \tag{6.20}$$

whose corresponding driving signal is calculated by,

$$\mathbf{s}^{(m)} = |(\boldsymbol{h}^{(m)})^T R^m \boldsymbol{\alpha}|. \tag{6.21}$$

Above procedures come to stop after a sufficient number of iterations or when the residual $R^{(m)}\boldsymbol{\alpha}$ is close to zero [170].

LASSO; it reformulates (6.17) by combining the objective function and sparse

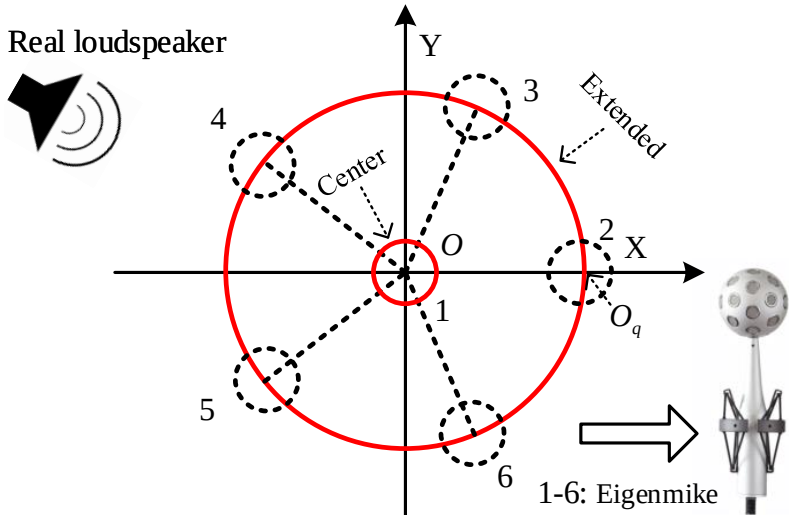


Figure 6.1: Vertical view of the system setup for experiments.

constraint into a united expression.

$$\boldsymbol{s} = arg\min_{\boldsymbol{s}} \ ||\boldsymbol{\alpha} - \boldsymbol{H}\boldsymbol{s}||_2^2 + \lambda||\boldsymbol{s}||_1. \tag{6.22}$$

The parameter λ controls the extent of sparsity for vector $\mathbf{s}$. An optimal variable selection of LASSO for (6.22) can be realized by the coordinate descent algorithms [171].

6.5 Validation of the Proposed Models

Intuitively, a successful theoretical model of the loudspeaker shall enable the ability to predict its incident field outside of the original measurements. Therefore, for validations, we use the proposed models to simulate the soundfield over an extended area and compare that to the real recordings.

In [140, 172, 173, 174], the authors presented a simplified array geometry for three-dimensional soundfield capture, which employs a horizontal planar array of first-order microphones. The reader is encouraged to refer to [140] for a detailed description of the theory behind the design. To analyze the accuracy over the extended area, the [120] used a larger array of higher-order microphone microphones

to fully capture a soundfield over the while spatial area. Then, the prediction errors over the extended area are computed using the recorded and predicated spherical harmonics coefficients. However, this validation method has an inherent drawback as it is only suitable for a low/narrow frequency band (around 300 Hz in [120]).

For evaluations over a wide frequency band, we design a practical measuring setup using direct sound pressure measurements recorded by Eigenmike, which is demonstrated in Fig. 6.1 and note that the speaker symbolizes a real commercial loudspeaker. The small red circle of radius 4.2 cm at the center of Fig. 6.1, represents the Eigenmike recording used for modeling while the remained 5 Eigenmikes placed along the boundary of a larger red circle with a radius of 0.3 m are employed for validations.

Assuming stationary conditions, we propose to emulate only one single Eigenmike moving along a horizontal circle to record the sound field for validations. The strategy to move the Eigenmike separately for each recording provides two main benefits: (i) the recording for the spatial soundfield can be accomplished conveniently by one single Eigenmike alone that reduces hardware costs; (ii) it avoids perturbations of the scattering effect when to set various Eigenmike nearby in the sound field. Details of the setup will be presented in experiments. For each model, the sound pressure over the 5 Eigenmikes (160 channels in total) beyond the center can be approximated or simulated by the selected source candidates and their driving signal. To analyze the accuracy, the predication errors over the extended area in terms of soundfield pressure are computed as (6.23) shows below,

$$\text{Error}_P(k) = 10\log_{10}\left(\frac{\sum_{q=1}^{Q}|P^{\text{rec}}(\theta_q,\phi_q,k)-P^{\text{pre}}(\theta_q,\phi_q,k)|^2}{\sum_{q=1}^{Q}|P^{\text{rec}}(\theta_q,\phi_q,k)|^2}\right) \tag{6.23}$$

in which the $P^{\text{rec}}(\cdot)$ and $P^{\text{pre}}(\cdot)$ represents the recorded and predicted sound pressure for each channel, respectively.

However, a single metric of numerical sound pressure errors cannot fully evaluate or reflect the performance of the proposed sparsity exploited models. It is conceivable that a successful acoustic model exploiting sparsity shall enable the ability to select active candidates that enclose the direction of the real loudspeaker.

Therefore, apart from numerical error, spatial distributions of the active or selected source candidates with respect to the direction of the real loudspeaker are investigated and taken into account for thorough evaluations. And such characteristics can be exhibited easily by plotting the distributions of the selected candidates as Fig. 6.6 shows.

6.6 Experiments

This section presents an experimental set up where a commercial loudspeaker broadband response is recorded and modeled using the proposed equivalent source models. The performance of the proposed models is analyzed using secondary measurements over the extended area.

6.6.1 Experimental Setup

The experiment setup mainly consists of four stages. Firstly, we use an Eigenmike to record the soundfield due to a loudspeaker of interest. The chosen loudspeaker is a single unit from the loudspeaker array that is originally built for sound field reproduction (i.e., see Fig. 6.2). The selected loudspeaker is located at $(r, \theta, \phi) = (1, 0.55, 0.62)$ with respect to the center of the listening area. The loudspeaker's incident sound field (only the direct path) is recorded at the origin using an Eigenmike as shown in Fig. 6.1. Secondly, the Eigenmike recordings are used to derive a set of theoretical source models using the proposed sparsity exploited methods. Thirdly, for validation of the proposed models outside of the original recording area, secondary soundfield recordings are obtained using a moving Eigenmike (see Fig. 6.1). Finally, we analyze the accuracy of the proposed methods using the metric given in (6.23). Furthermore, we also study the concept of sparsity exploited equivalent source models by analyzing the spatial distribution of the proposed source decompositions.

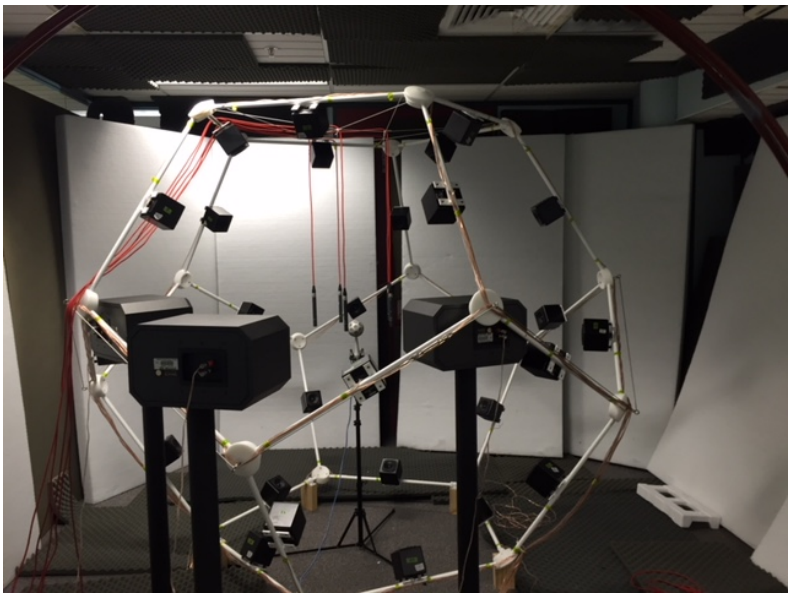


Figure 6.2: A 30-units loudspeaker array and the EigenMike at the center.

6.6.2 Accuracy of the Proposed Models Using the Numerical Metric

Here, we use the metric of (6.23) to study the accuracy of the proposed models outside of the region where original recordings were taken. Note that we mainly consider frequencies below 1 kHz, because when soundfield reproduction accuracy is important, low frequencies are those most affected by loudspeaker directivity.

Figure 6.3 (a), (b) and (c) report the accuracy in terms of sound pressure errors for the plane wave, point source, and mixed source models using the three sparse algorithms. Under low-frequency conditions, the soundfield over the whole area in Fig. 6.1 may share certain similarities that make it hard to distinguish the effectiveness of proposed models. Therefore, we calculate the differences of the sound pressure over the extended area with respect to the soundfield at the center and take it as a baseline.

Generally, each acoustic model using any sparse algorithm achieves satisfying performance, especially when the frequencies are below 500Hz. The performance decreases when the frequency increases in that the desired field is under-sampled

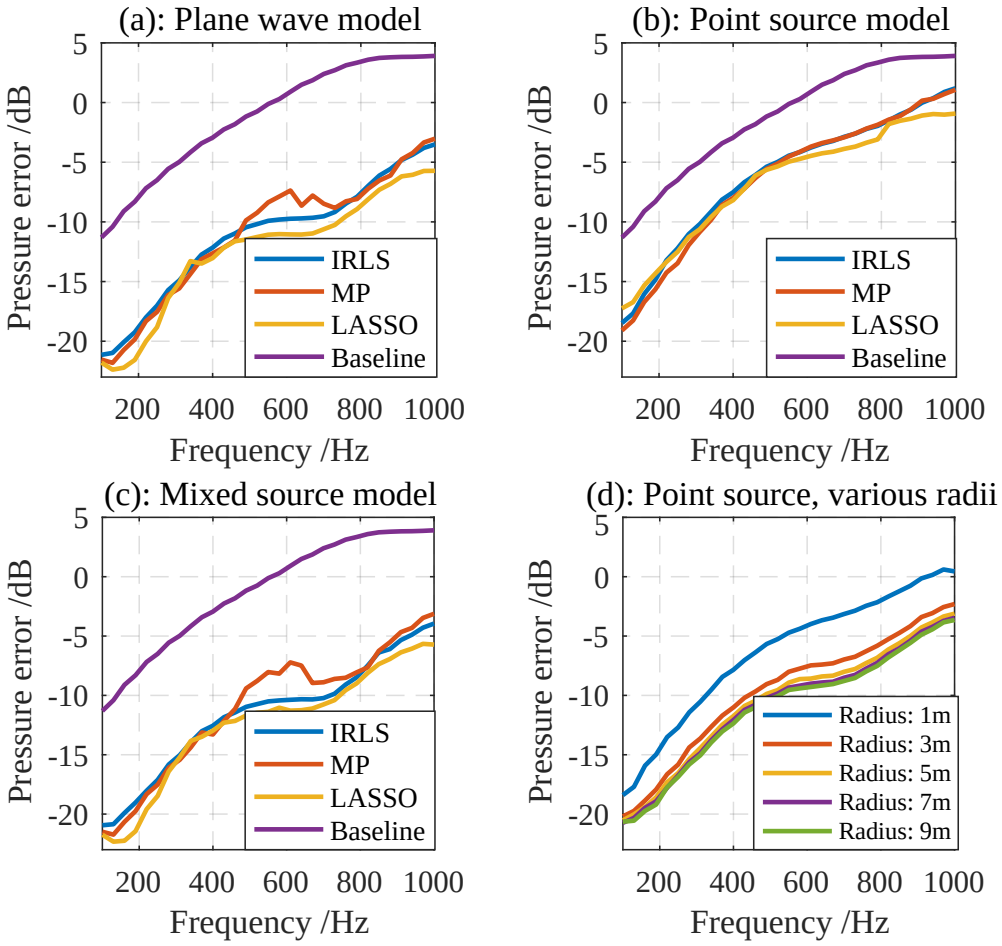


Figure 6.3: Sound pressure errors for all acoustic models

at higher frequencies. Among the sparse algorithms, the LASSO achieves the least numerical errors while it has a slow convergence rate and suffers from an expensive cost to determine the sparse parameter λ in (6.22). The IRLS shows to be with more errors at low frequencies while outperforms the MP at higher frequency bins. For the acoustic models, the mixed acoustic model achieves competitive performance with a slight improvement when compared to the plane wave model. However, the point source model, with the same radius as that of the mixed model, appears to be with more errors than the other two source models.

Figure 6.3 (d) presents the mean sound pressure errors using all the three sparse algorithms for the point source model when they are placed at various radii. It

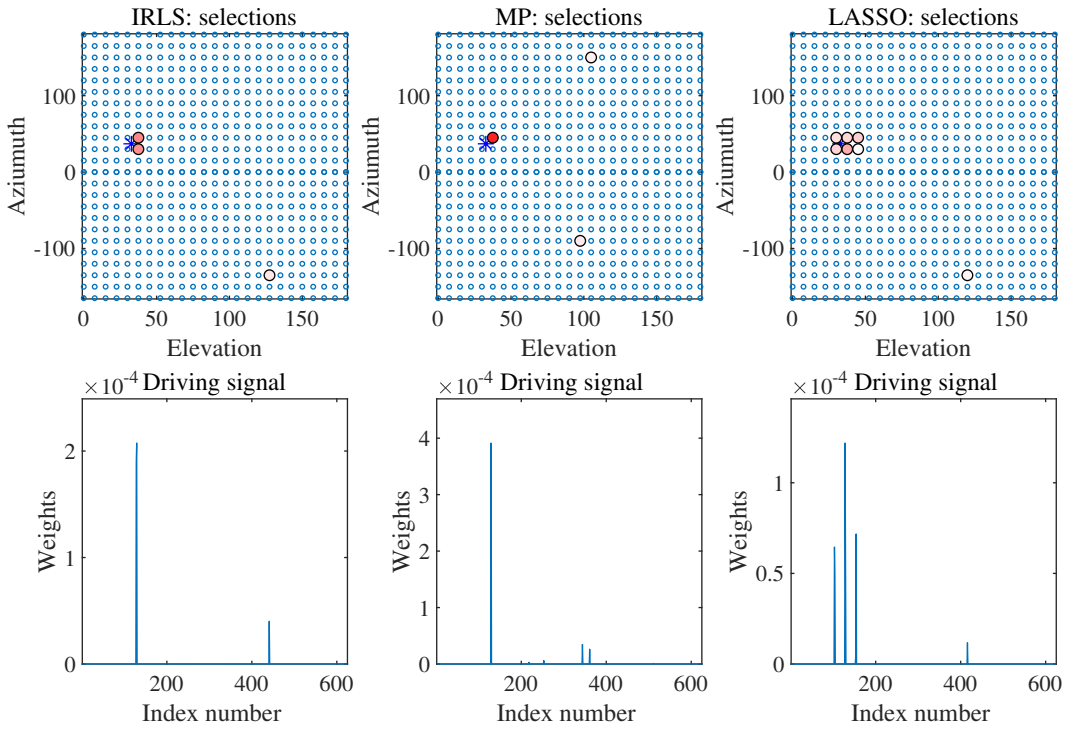


Figure 6.4: Sparse distributions of selected candidates along with corresponding magnitude of driving signal for plane wave model at 600Hz.

demonstrates that the larger radius that lies further away from the sensors acts more like plane waves and leads to reduced modeling errors.

6.6.3 Analysis of the Spatial Distribution of Sparsity Exploited Source Models

The Fig. 6.4, Fig. 6.5 and Fig. 6.6 present the spatial distributions of selected source candidates for all proposed models using all sparse algorithms when $f = 600$ Hz. Note that such distributions share similar profile over other frequency bins considered. For each sub-figure, the small circle lined in gird stands for the location of a candidate (625 in total) and the notation of $*$ in blue indicates the direction of the real speaker. The magnitude of the driving signal represents the degree of activity or importance for that candidate. The four double-columns marked by

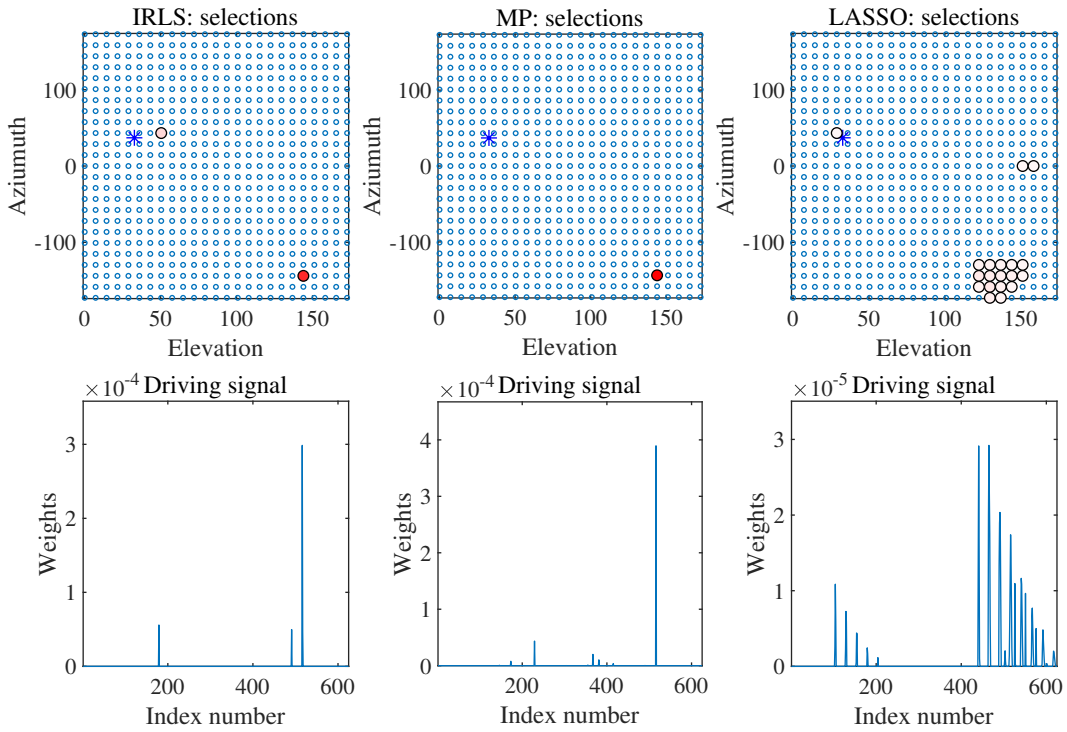


Figure 6.5: Sparse distributions of selected candidates along with corresponding magnitude of driving signal for point source model at 600Hz.

brackets correspond to distributions of plane wave model, point source model, the parts of plane wave, and points source within the mixed source model, respectively. Three lines of figures from top to down employ IRLS, MP, and LASSO, respectively.

Results verify that most of the active candidates selected by the plane wave model lie around the direction of the real speaker while the point source model fails. The mixed source model exhibits promising distributions in that the plane wave parts play the leading role and a small number of components originate from the point source as well, which shall be kinds of signal components sharing similar characteristics with the point source. Considering the sparse solutions, the MP turns out to be the most sparse one due to the exponential decay of the residual error [170]. The IRLS, with less sound pressure errors than MP, provides satisfying sparse source distributions as well. Though the LASSO produces the least numer-

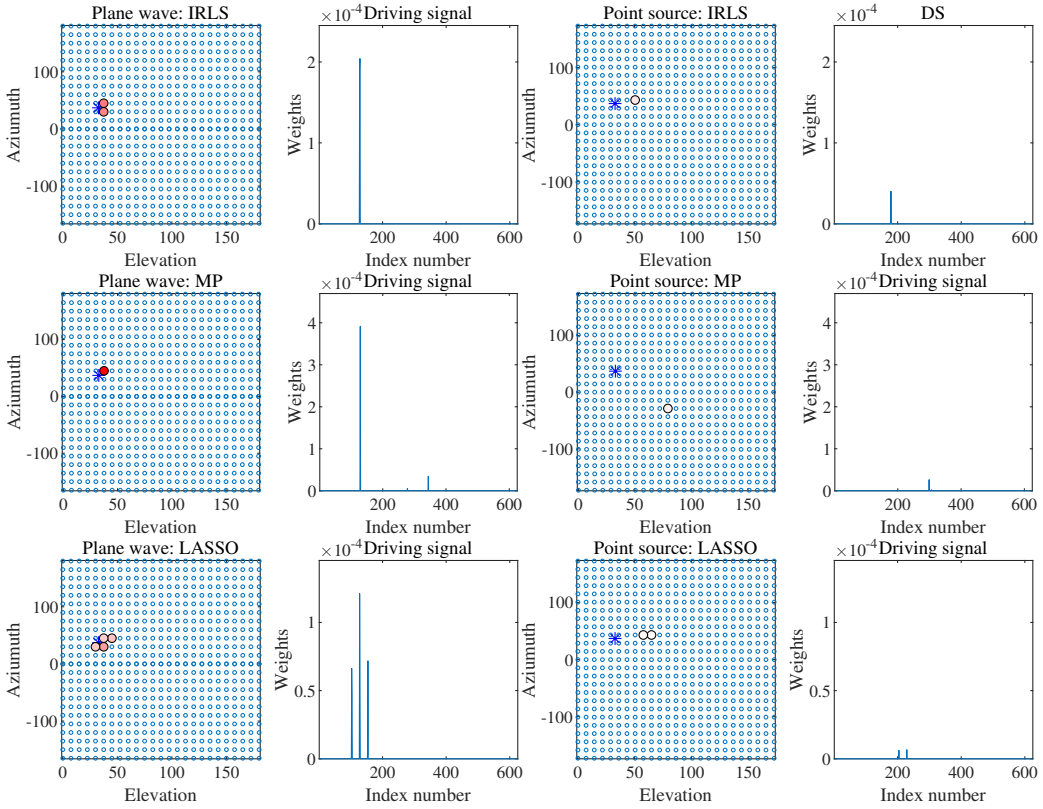


Figure 6.6: Sparse distributions of selected candidates along with corresponding magnitude of driving signal for mixed source model at 600Hz.

ical errors, it does not select an optimal sparse distribution due to that the global minimum of numerical error does not necessarily coincide with the optimal sparse solutions.

6.7 Summary

This chapter has proposed several acoustic models to model the inherent characteristics of commercial loudspeakers used in soundfield reproduction systems. Experimental results with promising performance have shown that both the plane wave model and mixed source model perform possess an ability to characterize

commercial speakers with acceptable accuracy. The current mixed source model is a combination of the plane wave and point source over the whole frequency band. However, characteristic of the real sound sources depends on the frequency band, i.e., the sound sources perform differently over low, medium, and high-frequency band. Hence, a promising direction to improve current work is to use different acoustic models to represent the real sources over different frequency bands. The differentiated modeling of the sound sources over the frequency bands shall be more consistent with the case in reality.

In this chapter, these developed acoustic models suit more for real-life loudspeakers, which can be used in real-life sound source localization systems as well as other spatial acoustic techniques, such as soundfield reproduction and active noise control systems. However, this chapter has not yet developed related techniques to incorporate the acoustic models into those acoustic applications/systems. Hence, another future direction is to apply the acoustic models for improved performance in practical acoustic applications, such as source localization under real-life scenarios.

6.8 Related Publications

This chapter's work has ever been published in the following conference proceedings.

- Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling the interior response of real loudspeakers with finite measurements", in 2018 *IEEE 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 16-20.

- Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-domain approaches", in 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561-565.

Chapter 7

Summary

***Overview**: This chapter first concludes the work in this thesis, then summarizes the main contributions achieved by the research, and finally outlines some future research directions.*

7.1 Conclusions

Chapter 1 presented the problem to be addressed and introduced the outline of this thesis.

Chapter 2 reviewed the up-to-date progress of sound source localization in the literature review. This chapter also introduced some essential background knowledge of spherical harmonics representation of a measured soundfield using higher-order microphone arrays. Finally, we introduced the relative transfer function that motivated us to develop the spherical harmonics domain source feature in this thesis.

Chapter 3 presented the novel source feature called the *relative harmonic coefficients*. We utilized this source feature to develop two single source localization algorithms that have significantly reduced computational complexity. Additionally, this feature is useable for acoustic enhancement for the noisy microphone array recordings. The localization approach in Chapter 3 is only suited for a single sound source in a generally mild environment. By contrast, Chapter 4 used the relative harmonic coefficients to develop a semi-supervised multi-source localization

algorithm in a complex acoustic environment. We highlight that this data-driven method has a strong potential to be implemented in practice as only a limited number of labeled measurements is required. The proposed multi-source localization algorithm in Chapter 4 had an inherent limit as it required the availability of single-source components in the time domain. Thus, it is unusable in the cases where the original recordings have a limited number of single-source components (i.e., simultaneous multi-source recordings). To fix this issue, we developed a novel MUSIC framework based approach that directly exploited the simultaneous multi-source recordings. This developed MUSIC approach showed to be more suitable in noisy environments than the traditional MUSIC method.

There still exist many practical factors that influence the localization accuracy of the proposed localization techniques. Generally, the assumption of omnidirectional behavior in the loudspeakers, which is hardly true with commercial loudspeakers, actually degrades the localization approaches under real-life scenarios. Hence, Chapter 6 analyzed such characteristics of loudspeakers by deriving equivalent theoretical acoustic models. Several acoustic models were investigated, including plane waves decomposition, point source decomposition, and mixed source decomposition.

To conclude, this thesis has developed several spherical harmonics domain approaches to address the sound source localization in adverse acoustic environments. By the end, we highlight the following three significant contributions in this thesis: (i) To the author's best knowledge, this thesis is the first comprehensive investigation of the relative harmonic coefficients: including the feature definition, estimator in noisy environments, analysis of its directivity pattern as well as a source feature selector over space and an overlapped frame detector to find the single-source components. (ii) The developed sound source localization techniques that are suitable under different acoustic scenarios: the decoupled localization method with a fast speed for single source tracking, the Multi-Mode Gaussian Process modeling based semi-supervised multi-source localization performing with sufficient accuracy in a severely noisy and reverberant environment, and the improved MUSIC methods using simultaneous multi-source recordings in noisy environments. (iii) A compact solution to model the characteristics of commercial loudspeakers: in addition to sound source localization, this study also contributes to many other spatial acous-

tic techniques, such as soundfield reproduction and active noise control systems.

7.2 Contributions

Several key contributions of this thesis are summarized as follows:

- *A systematic investigation of relative harmonic coefficients*: one of the main contributions by this thesis is the systematic study of relative harmonic coefficients. It performs to be a promising source feature for sound source localization using the high order array microphone recordings. We explored several remarkable properties of the relative harmonic coefficients desired for source localization, including (1) its independence from the time-varying source signal and sole dependence on the source position even in a reverberant environment; (2) easy estimations in noisy environments from the higher-order microphone array recordings; (3) a significant spatial resolution due to its unique directivity pattern over space; and (4) exploitable by an overlapped frame detector to simplify the challenging localization of multiple sources into single source localization issues.

- *A low complexity DOA estimation approach*: this developed approach overcomes the inherent drawback of the traditional DOA estimations that require an exhaustive search over the two-dimensional (2-D) space (i.e., elevation and azimuth space). The algorithm exploits the fact that the elevation and azimuth components in the estimated relative harmonic coefficients are decoupled so that the source elevation and azimuth are localized separately. The reduced computational complexity dramatically improves the processing speed, which well suits for sound source tracking.

- *Application for an acoustic enhancement approach for the noisy microphone array recordings*: the properties of the relative harmonic coefficients also usable to develop an acoustic enhancement approach in the spherical harmonics domain. This approach addresses the common problem when acquiring spatial soundfield recordings in noisy environments because the environmental and thermal noise hinders an accurate acquisition of the desired soundfield.

Hence, this developed approach can be used as a preprocessing tool by current spatial acoustic processing techniques.

- *A semi-supervised localization approach that suits for complex environments*: this approach overcomes the drawback of traditional data-driven source localization techniques that require a large training set. Instead, it uses the relative harmonic coefficients estimated from a limited number of labeled measurements. This approach extends the Multi Gaussian Process modeling to the spherical harmonics domain (called as Multi-Mode Gaussian Process (MMGP)) to merge/fuse the relative harmonic coefficients over the selected spherical harmonic modes. Finally, this algorithm formulates the mapping function, revealing the underlying relation between the source feature(s) and source position(s), using MMGP based Gaussian Process Regression (GPR) to recover the unknown source position in complex environments.

- *single-source components detection*: this thesis develops a single-source frame detector that simplifies localization of multi-source into repetitive single source localization problems. The single-source frame detector exploits a distinguished characteristic of relative harmonic coefficients against source presence. Namely, the relative harmonic coefficients, due to the frames containing only a single source, belong to a unique source feature set. The proposed single-source frame detector is appliable to different types of multi-source recordings, with a promising application in the events that have multiple sound sources.

- *A MUSIC approach suitable in noisy environments*: although multiple signal classication (MUSIC) has become one of the most popular multi-source DOA estimators, its localization accuracy is vulnerable to noise. This thesis overcomes this inherent drawback by developing a relative sound pressure based MUSIC algorithm that is more suitable in noisy environments. We also decompose this method into the spherical harmonics domain, where a frequency smoothing technique is allowed to de-correlate the coherent source signals for improved localization accuracy. The proposed algorithms achieved better performance in comparison with traditional MUSIC methods.

- *A compact solution to model the characteristics of commercial loudspeakers*: lots of spatial acoustic processing techniques, such as the proposed source localization approaches by this thesis, generally assume the sound sources follow an omnidirectional behavior, which is not satisfied with the case for commercial loudspeakers. However, to the best of our knowledge, there exist very few techniques investigating this issue. This thesis analyzes real loudspeakers characteristics by deriving several equivalent theoretical models, whose performance is finally validated using the metrics proposed by this thesis.

7.3 Future Research

There remain some further research problems to be addressed by this thesis. Some promising future directions are listed as follows:

Relative harmonic coefficients estimations for multi-source scenarios in complex environments

Accurate estimations of the relative harmonic coefficients in a complex environment are vital to the proposed localization algorithms. This thesis proposed a biased feature estimator that neglects the noise power spectral density. Thus, the accuracy degrades severely in strongly background noise environments. One possible solution is to use state-of-art noise power spectral density estimators [146, 147]. Another limit of the current feature estimator is the limitation to single-source scenarios. Hence, an interesting topic is how to estimate the relative harmonic coefficients in noisy environments where multiple simultaneous sources are active.

Decoupled multi-source localization in more dynamic environments

The decoupled localization approach in Chapter 3 mainly considered a single source scenario. A promising direction is to extend the underlying theory into the multi-source localization. A possible solution to address this issue is to use single source

detection, developed in Chapter 4, to find the single-source components. Then, repeated decoupled single source localization is applied to achieve the multi-source localization. Not that Chapter 3 considered a far-field scenario and failed to localize the source range. Hence, another possible direction is to investigate the decoupled source localization under a near-field case where the source to microphone array distance is also estimated.

Apply the relative harmonic coefficients into deep learning schemas based source localization

Chapter 4 studies the multi-source localization issue in severely noisy and reverberant environments. We proposed a data-driven technique to learn the characteristic of the training measurements. However, the developed approach used a Multi-mode Gaussian process regression model, which only suits to localize continuous positional variables, such as the Cartesian coordinates of the sources. Thus, it is unsuitable for DOA estimations as this belongs to a classification problem. In terms of data-driven DOA estimations, a promising direction is to use advanced deep learning schema, such as CNNs [23] and CRNNs [22], to achieve sufficient localization accuracy in a severely noisy and reverberant environment.

Acoustic enhancement for multi-source scenarios in a dynamic environment

Chapter 2 applied the relative harmonic coefficients to an acoustic enhancement approach for cleaning the noisy higher-order microphone recordings. The enhancement approach estimates the spherical harmonic coefficients by joint estimation of relative harmonic coefficients and received signal at the microphone array in noisy environments. However, it is currently only limited to a single static sound source under a far-field scenario. Hence, one potential future work is to extend the underlying theory into a denoising scheme for multi-source cases in a more dynamic environment, such as a near-field propagation.

Application for voice activity detection

Voice activity detection [175, 176] refers to a family of methods that perform segmentation of an audio signal into parts that contain speech and silent periods. This technique benefits many speech-based applications such as speech and speaker recognition, speech enhancement, emotion recognition, and dominant speaker identification. Our initial research in [1] studied the characteristics of relative harmonic coefficients against source presence. We see the relative harmonic coefficients of the frames/bins containing only a single source have unique characteristics. Hence, a promising future direction is to apply the relative harmonic coefficients into voice activity detection (VAD) for sufficient detection accuracy.

Implement the proposed techniques using other types of microphone arrays

Although the approaches by the thesis use a spherical microphone array, the underlying theories are equally applicable for other structured microphone arrays, such as first-order microphone arrays [177], planar microphone arrays [140] and multiple circular microphone arrays [139]. More interestingly, we expect to implement the techniques developed by this thesis into practical microphone arrays, such as a circular microphone array using a limited number of microphones [178]. Alternatively, it is a promising direction to design a practical microphone setup suitable for the developed approaches, convenient for practical applications while achieving satisfying performance.

Bibliography

[1] Y. Hu, P. N. Samarasinghe, T. D. Abhayapala, and S. Gannot, "Unsupervised multiple source localization using relative harmonic coefficients," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 571–575.

[2] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, "Surround by sound: A review of spatial audio recording and reproduction," *Applied Sciences*, vol. 7, no. 5, pp. 532, 2017.

[3] C. Evers, H. W. Lllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 1620–1643, 2020.

[4] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation and source separation in a noisy reverberant environment using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1594–1607, 2018.

[5] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[6] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent

neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.

[7] H. Sun, S. Yan, and U. P. Svensson, "Space domain optimal beamforming for spherical microphone arrays," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 117–120.

[8] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[9] Y. Hu, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 909–912.

[10] S. Gannot, M. Haardt, W. Kellermann, and P. Willett, "Introduction to the issue on acoustic source localization and tracking in dynamic real-life scenes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 3–7, 2019.

[11] H. W. Lllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The locata challenge data corpus for acoustic source localization and tracking," in *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 410–414.

[12] 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.

[13] B. Jo and J. Choi, "Direction of arrival estimation using nonsingular spherical ESPRIT," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 181–187, 2018.

[14] L. I. Birnie, T. D. Abhayapala, and P. N. Samarasinghe, "Reflection assisted sound source localization through a harmonic domain music frame-

work," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 279–293, 2020.

[15] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer, 2007.

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

[17] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–224.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[19] 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.

[20] H. Teutsch and W. Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5276–5279.

[21] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 117–120.

[22] L. Perotin, R. Serizel, E. Vincent, and A. Gurin, "CRNN-based multiple DOA estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[23] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Multi-source DOA estimation through pattern recognition of the modal coherence of a reverberant soundfield," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 605–618, 2020.

[24] 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 (TASLP)*, vol. 22, no. 10, pp. 1494–1505, 2014.

[25] 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.

[26] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[27] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[28] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, 1976.

[29] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 375–378.

[30] S. Gannot and T. Dvorkind, "Microphone array speaker localizers using spatial-temporal information," *EURASIP Journal on Applied Signal Processing*, pp. 174–174, 2006.

[31] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A hybrid approach for speaker tracking based on TDOA and data-driven models," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 4, pp. 725–735, 2018.

[32] K. Yao, J. C. Chen, and R. E. Hudson, "Maximum-likelihood acoustic source localization: experimental results," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 2949–2952.

[33] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science, 2013.

[34] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. I–121–I–124.

[35] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," *IEEE Signal Processing Letters*, vol. 18, no. 1, pp. 71–74, 2011.

[36] W. Zhang, *Measurement and Modelling of Head-Related Transfer Function for Spatial Audio Synthesis*, Ph.D. thesis, Australian National University, 2010.

[37] Z. Zohny, S. M. Naqvi, and J. A. Chambers, "Variational em for clustering interaural phase cues in messl for blind source separation of speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3966–3970.

[38] G. R. Karthik and P. K. Ghosh, "Binaural speech source localization using template matching of interaural time difference patterns," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5164–5168.

[39] S. T. Birchfield and R. Gangishetty, "Acoustic localization by interaural level difference," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. iv/1109–iv/1112 Vol. 4.

[40] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space," *Journal of the Audio Engineering Society*, vol. 49, pp. 231–249, 2001.

[41] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Binaural localization of speech sources in 3-D using a composite feature vector of the HRTf," pp. 2654–2658.

[42] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Spatial feature learning for robust binaural sound source localization using a composite feature vector," pp. 6320–6324.

[43] J. Braasch and K. Hartung, "Localization of distracted sound sources: Determining the role of binaural cues using unilaterally attenuated and interaurally uncorrelated signals," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1151–1151, 1999.

[44] A. Kulkarni and H. S. Colburn, "Infinite impulse response models of the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 115, pp. 1714–28, 2004.

[45] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised source localization on multiple manifolds with distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 7, pp. 1477–1491, 2017.

[46] N. Ma, J. A. Gonzalez, and G. J. Brown, "Robust binaural localization of a target sound source by combining spectral source models and deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2122–2131, 2018.

[47] B. Laufer-Goldshtein, R. Talmon, I. Cohen, and S. Gannot, "Multi-view source localization based on power ratios," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75.

[48] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C. Deledalle, "Machine learning in acoustics: Theory and applications," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3590–3628, 2019.

[49] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2814–2818.

[50] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6.

[51] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6403–6413, 2018.

[52] S. Chakrabarty and Emanul A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 136–140.

[53] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting cnns for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 103–116, 2018.

[54] W. He, P. Motlicek, and J. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," in *2018 INTERSPEECH*, pp. 312–316.

[55] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1462–1466.

[56] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 405–409.

[57] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Semi-supervised sound source localization based on manifold regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 8, pp. 1393–1407, 2016.

[58] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Manifold-based bayesian inference for semi-supervised source localization," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6335–6339.

[59] R. Opochinsky, B. Laufer-Goldshtein, S. Gannot, and G. Chechik, "Deep ranking-based sound source localization," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 283–287.

[60] Y. Dorfan, O. Schwartz, B. Schwartz, E. A. Habets, and S. Gannot, "Multiple DOA estimation and blind source separation using estimation-maximization," in *2016 IEEE Conference on the Science of Electrical Engineering (ICSEE)*, pp. 1–5.

[61] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[62] 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.

[63] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2136–2147, 2008.

[64] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[65] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *2003 Seventh International Symposium on Signal Processing and Its Applications.*

[66] S. Rennie, P. Aarabi, T. Kristjansson, B. J. Frey, and K. Achan, "Robust variational speech separation using fewer microphones than speakers," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I–88.

[67] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[68] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1085–1094, 2017.

[69] Kevin Wilson, *Speech Source Separation by Combining Localization Cues with Mixture Models of Speech Spectra*, vol. 1, 2007.

[70] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[71] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[72] P. Stoica and A. Nehorai, "Music, maximum likelihood, and cramer-rao bound," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 5, pp. 720–741, 1989.

[73] B. A. Johnson, Y. I. Abramovich, and X. Mestre, "MUSIC, G-MUSIC, and maximum-likelihood performance breakdown," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3944–3958, 2008.

[74] X. Mestre and M. A. Lagunas, "Modified subspace algorithms for DOA estimation with large arrays," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598–614, 2008.

[75] J. Liang and D. Liu, "Passive localization of mixed near-field and far-field sources using two-stage MUSIC algorithm," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 108–120, 2010.

[76] J. Vinogradova, R. Couillet, and W. Hachem, "Statistical inference in large antenna arrays under unknown noise pattern," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5633–5645, 2013.

[77] P. Vallet, X. Mestre, and P. Loubaton, "Performance analysis of an improved MUSIC DOA estimator," *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6407–6422, 2015.

[78] Q. Huang, L. Zhang, and Y. Fang, "Two-stage decoupled DOA estimation based on real spherical harmonics for spherical arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 11, pp. 2045–2058, 2017.

[79] W. Zuo, J. Xin, H. Ohmori, N. Zheng, and A. Sano, "Subspace-based algorithms for localization and tracking of multiple near-field sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 156–171, 2019.

[80] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.

[81] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Relative transfer function modeling for supervised source localization," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4.

[82] F. Ribeiro, D. Florencio, and D. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1781–1792, 2010.

[83] Y. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[84] P. Svaizer, A. Brutti, and M. Omologo, "Use of reflected wavefronts for acoustic source localization with a line array," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 165–169.

[85] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 1999.

[86] L. Birnie, T. D. Abhayapala, H. Chen, and P. N. Samarasinghe, "Sound source localization in a reverberant room using harmonic based MUSIC," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 651–655.

[87] L. I. Birnie, T. D. Abhayapala, and P. N. Samarasinghe, "Reflection assisted sound source localization through a harmonic domain MUSIC framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 279–293, 2020.

[88] P. Samarasinghe, T. Abhayapala, M. Poletti, and T. Betlehem, "An efficient parameterization of the room transfer function," *IEEE/ACM Transactions*

on Audio Speech and Language Processing (TASLP), vol. 23, no. 12, pp. 2217–2227, 2015.

[89] P. N. Samarasinghe, T. D. Abhayapala, Y. Lu, H. Chen, and G. Dickins, "Spherical harmonics based generalized image source method for simulating room acoustics," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1381–1391, 2018.

[90] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 24, no. 11, pp. 2171–2186, 2016.

[91] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1997–2012, 2017.

[92] L. Madmoni and B. Rafaely, "Direction of arrival estimation for reverberant speech based on enhanced decomposition of the direct sound," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 131–142, 2019.

[93] H. Beit-On and B. Rafaely, "Speaker localization using the direct-path dominance test for arbitrary arrays," in *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pp. 1–4.

[94] W. Xue, Y. Tong, G. Ding, C. Zhang, T. Ma, X. He, and B. Zhou, "Direct-path signal cross-correlation estimation for sound source localization in reverberation," in *2019 INTERSPEECH*, pp. 2693–2697.

[95] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[96] M. B. teli, O. Olgun, and H. Hachabibolu, "Multiple sound source localization with steered response power density and hierarchical grid refine-

ment," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2215–2229, 2018.

[97] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, 2019.

[98] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *2014 IEEE 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 258–262.

[99] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, *Spherical microphone array beamforming*, pp. 281–305, Springer, 2010.

[100] 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.

[101] T. D. Abhayapala and H. Bhatta, "Coherent broadband source localization by modal space processing," *2003 10th International Conference on Telecommunications (ICT)*, pp. 1617–1623.

[102] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *2010 18th European Signal Processing Conference (EUSIPCO)*, pp. 442–446.

[103] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[104] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2100–2111, 2005.

[105] P. N. Samarasinghe, M. A. Poletti, S. M. A. Salehin, T. D. Abhayapala, and F. M. Fazi, "3D soundfield reproduction using higher order loudspeakers," pp. 306–310.

[106] J. Wenyu and W. B. Kleijn, "Theory and design of multizone soundfield reproduction using sparse methods," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2343–2355, 2015.

[107] J. Ahrens and S. Spors, "An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 988–999, 2008.

[108] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 8, pp. 2038–2050, 2010.

[109] A. Gupta and T. D. Abhayapala, "Three-dimensional sound field reproduction using multiple circular loudspeaker arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1149–1159, 2011.

[110] H. Chen, J. Zhang, P. N. Samarasinghe, and T. D. Abhayapala, "Evaluation of spatial active noise cancellation performance using spherical harmonic analysis," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5.

[111] F. Ma, W. Zhang, and T. D. Abhayapala, "Active control of outgoing broadband noise fields in rooms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 529–539, 2020.

[112] F. Ma, W. Zhang, and T. D. Abhayapala, "Active control of outgoing noise fields in rooms," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1589–1599, 2018.

[113] J. Zhang, W. Zhang, T. D. Abhayapala, J. Xie, and L. Zhang, "2.5D multizone reproduction with active control of scattered sound fields," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 141–145.

[114] J. Zhang, T. D. Abhayapala, W. Zhang, P. N. Samarasinghe, and S. Jiang, "Active noise control over space: A wave domain approach," *IEEE/ACM*

Transactions on Audio, Speech and Language Processing (TASLP), vol. 26, no. 4, pp. 774–786, 2018.

[115] W. Zhang, C. Hofmann, M. Buerger, T. D. Abhayapala, and W. Kellermann, "Spatial noise-field control with online secondary path modeling: A wave-domain approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, no. 12, pp. 2355–2370, 2018.

[116] X. Li, S. Yan, X. Ma, and C. Hou, "Spherical harmonics MUSIC versus conventional MUSIC," *Applied Acoustics*, vol. 72, no. 9, pp. 646–652, 2011.

[117] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 117–120.

[118] 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 (ICASSP)*, pp. 346–349.

[119] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1539–1551, 2015.

[120] Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling the interior response of real loudspeakers with finite measurements," in *2018 IEEE 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 16–20.

[121] Y. Hu, P. N. Samarasinghe, G. Dickins, and T. D. Abhayapala, "Modeling characteristics of real loudspeakers using various acoustic models: Modal-domain approaches," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561–565.

[122] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 261, 2012.

[123] R. E. Tiana, F. Jacobsen, and E. Fernandez-Grande, "Beamforming with a circular array of microphones mounted on a rigid sphere," *Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1095, 2011.

[124] 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, 2011.

[125] C. C. Lai, S. Nordholm, and Y. H. Leung, "Design of steerable spherical broadband beamformers with flexible sensor configurations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 427–438, 2013.

[126] T. D. Abhayapala, R. A. Kennedy, and Ro. C. Williamson, "Nearfield broadband array design using a radially invariant modal expansion," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 392, 2000.

[127] M. Schneider and W. Kellermann, "A wave-domain model for acoustic mimo systems with reduced complexity," in *2011 IEEE Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 133–138.

[128] M. Schneider and W. Kellermann, "Adaptive listening room equalization using a scalable filtering structure in thewave domain," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13–16.

[129] A. Fahim, P. Samarasinghe, and T. D. Abhayapala, "Extraction of exterior field from a mixed sound field for 2D height-invariant sound propagation," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5.

[130] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "Sound field separation in a mixed acoustic environment using a sparse array of higher order

spherical microphones," *2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA)*, pp. 151–155, 2017.

[131] 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.

[132] M. A. Poletti and T. D. Abhayapala, "Interior and exterior sound field control using general two-dimensional first-order sources," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 234–244, 2011.

[133] M. Poletti, T. D. Abhayapala, and P. N. Samarasinghe, "Interior and exterior sound field control using two dimensional higher-order variable-directivity sources," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 3814–3823, 2012.

[134] P. A. Martin, *Multiple scattering: interaction of time-harmonic waves with N obstacles*, Cambridge University Press, 2006.

[135] P. Morse and H. Feshbach, *Methods of theoretical physics*, 1953.

[136] G. N. Watson, *A treatise on the theory of Bessel functions*, p. 361, Cambridge University Press, London, UK, 2 edition, 1995.

[137] T. D. Abhayapala, *Modal analysis and synthesis of broadband nearfield beamforming arrays*, Ph.D. thesis, Australian National University, 1999.

[138] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1949–1952.

[139] T. D. Abhayapala and A. Gupta, "Spherical harmonic analysis of wavefields using multiple circular sensor arrays," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 6, pp. 1655–1666, 2010.

[140] H. Chen, T. D. Abhayapala, and W. Zhang, "Theory and design of compact hybrid microphone arrays on two-dimensional planes for three-dimensional soundfield analysis," *Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 3081–3092, 2015.

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

[142] B. Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[143] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[144] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[145] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "A study on manifolds of acoustic responses," in *2016 International Conference on Latent Variable Analysis and Signal Separation*. pp. 203–210, Springer.

[146] J. K. Nielsen, M. S. Kavalekalam, M. G. Christensen, and J. Boldt, "Model-based noise PSD estimation from speech in non-stationary noise," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5424–5428.

[147] K. Niwa, T. Kawase, K. Kobayashi, and Y. Hioka, "PSD estimation in beamspace using property of m-matrix," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5.

[148] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments,"

IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 1, pp. 88–103, 2019.

[149] A. Brendel, B. Laufer-Goldshtein, S. Gannot, R. Talmon, and W. Kellermann, "Localization of an unknown number of speakers in adverse acoustic conditions using reliability information and diarization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7898–7902.

[150] Y. Hu, P. N. Samarasinghe, and T. D. Abhayapala, "Sound source localization using relative harmonic coefficients in modal domain," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 348–352.

[151] 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.

[152] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and applications of spherical microphone array processing*, vol. 9, Springer, 2017.

[153] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using wigner-d weighting," *IEEE Signal Processing Letters*, vol. 15, pp. 417–420, 2008.

[154] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smalloom acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[155] C. E. Rasmussen and C. K. Williams, *Gaussian process for machine learning*, MIT press, 2006.

[156] I. Pazsit and Y. Yamane, "The variance-to-mean ratio in subcritical systems driven by a spallation source," *Annals of Nuclear Energy*, vol. 25, no. 9, pp. 667–676, 1998.

[157] L. Kumar, G. Bi, and R. M. Hegde, "The spherical harmonics root-music," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3046–3050.

[158] J. K. Thomas, L. L. Scharf, and D. W. Tufts, "The probability of a subspace swap in the SVD," *IEEE Transactions on Signal Processing*, vol. 43, no. 3, pp. 730–736, 1995.

[159] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 727–732, 2007.

[160] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[161] M. Poletti, T. D. Abhayapala, and P. N. Samarasinghe, "Interior and exterior sound field control using two dimensional higher-order variable-directivity sources," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 3814–3823, 2012.

[162] N. Radmanesh, I. S. Burnett, and B. D. Rao, "A LASSO-LS optimization with a frequency variable dictionary in a multizone sound system," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 583–593, 2016.

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

[164] N. Murata, S. Koyama, N. Takamune, and H. Saruwatari, "Sparse representation using multidimensional mixed-norm penalty with application to sound field decomposition," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3327–3338, 2018.

[165] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Audio Speech and Language Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[166] H. Khalilian, I. V. Bajic, and R. G. Vaughan, "Comparison of loudspeaker placement methods for sound field reproduction," *IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP)*, vol. 24, no. 8, pp. 1364–1379, 2016.

[167] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, vol. 55, Courier Corporation, 1964.

[168] R. Chartrand and Y. Wotao, "Iteratively reweighted algorithms for compressive sensing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3869–3872.

[169] Y. Maeno, Y. Mitsufuji, and T. D. Abhayapala, "Mode domain spatial active noise control using sparse signal representation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[170] S. G. Mallat and Z. F. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[171] T. T. Wu and K. Lange, "Coordinate descent algorithms for LASSO penalized regression," *The Annals of Applied Statistics*, pp. 224–244, 2008.

[172] P. N. Samarasinghe, H. Chen, A. Fahim, and T. D. Abhayapala, "Performance analysis of a planar microphone array for three dimensional soundfield analysis," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 249–253.

[173] L. Birnie, P. N. Samarasinghe, and T. D. Abhayapala, "3D exterior soundfield capture using pressure and gradient microphone array on 2D plane," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1225–1229.

[174] L. Birnie, T. Abhayapala, and P. Samarasinghe, "Loudspeaker 3D directivity estimation with first order microphone measurements on a 2D plane," in *2017 143rd Audio Engineering Society Convention*, pp. 1–7.

[175] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 254–264, 2019.

[176] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, 2019.

[177] H. C. Chen, T. D. Abhayapala, P. N. Samarasinghe, and W. Zhang, "Direct-to-reverberant energy ratio estimation using a first-order microphone," *IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP)*, vol. 25, no. 2, pp. 226–237, 2017.

[178] A. Fahim, P. N. Samarasinghe, T. D. Abhayapala, and H. Chen, "A planar microphone array for spatial coherence-based source separation," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6.