ENABLING SPEECH APPLICATIONS USING AD HOC MICROPHONE ARRAYS

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 6589 (2015)
présentée le 27 March 2015
à la Faculté Sciences et Techniques de l'Ingénieur
laboratoire de LIDIAP
programme doctoral en Génie Électrique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Mohammad Javad Taghizadeh

acceptée sur proposition du jury:

Prof. Jean-Philippe Thiran, président du jury
Prof. Hervé Bourlard, Dr. Philip N. Garner, directeurs de thèse
Prof. Emanuel Habets, rapporteur
Prof. Dietrich Klakow, rapporteur
Dr. Hervé Lissek, rapporteur

Lausanne, EPFL, 2015

As you start to walk out on the way, the way appears
– *Rumi*

To my loving wife, Afsaneh

Acknowledgements

I would like to acknowledge my advisor, Prof. Hervé Bourlard for the great PhD opportunity at Idiap Research Institute and directing my research on the very interesting and challenging subject of ad hoc microphone arrays. I also would like to acknowledge my co-advisor, Dr. Phil Garner for his support and encouragement along with the independence that he gave me in my research endevors. Special thanks to the jury committee members, Prof. Emanuel Habets, Prof. Dietrich Klakow, Dr. Hervé Lissek and Prof. Jean-Philippe Thiran.

I extend my gratitude to the administrative team at Idiap. In particular, I would like to thank Nadine and Sylvie, Francois, Frank, Norbert, Cédric, Ed, Louis-Marie, Vincent, Christophe and Bastian.

I extend my acknowledgment to dear colleagues, Reza, Saeid, Sucheta, Ramya, Raphael, Marzieh, Samira, Milos, Maryam, Hassan, Alexandros, Pierre-Edouard, Petr, Marc, Vahid, Amin, Ali, Farid, Hamed, Shirin, Mohsen, Armin, Ahmad, Hesam, Hamid, Gelareh, Samuel, Gwenole, Elham, Lakshmi, Ivana, Marco, André Anjos, Leonidas, Elie, Kenneth, Murali, Harsha, Serena, Benjamin and Particia.

Special thanks to my mother and father, Homayoun and Mohammad Reza for their love and prayers. I also acknowledge my sisters, Shahed, Hoda and Katayoun for their encouragements. I also would like to acknowledge my mother and father in law, Maryam and Mohsen and bother-in-law, Ehsan. And above all, I would like to express my gratitude to my wife, Afsaneh for her true love and support in all circumstances. Her positive attitude and determination is always a great source of hope and strength. This dissertation is dedicated to her.

1 May 2015 Mohammad Javad Taghizadeh

Abstract

Microphone arrays are central players in hands-free speech interface applications. The main duty of a microphone array is capturing distant-talking speech with high quality. A microphone array can acquire the desired speech signals selectively by leading the beampattern towards the desired speaker. The foreseen application of ubiquitous sensing motivated by the abundance of microphone-embedded devices, such as notebooks and smart phones, raises the importance of research on ad hoc microphone arrays. The key challenges pertain to the unknown geometry of the microphones and asynchronous recordings.

The goal of this PhD thesis is to address the issues of microphone and source localization to enable beamforming for higher level speech processing tasks. To that end, we exploit the prior knowledge of the acoustical and geometrical structures underlying the ad hoc distributed nodes to devise novel algorithms for microphone array calibration and source localization, as well as beamforming techniques for distant speech applications.

To address the problem of ad hoc microphone array calibration, the analytic diffuse sound field coherence model is investigated and its fundamental properties are studied. This model enables pairwise distance estimation for calibration of a relatively compact microphone array. We derive the mathematical framework for estimation of long pairwise distances exploiting the low-rank properties of the Euclidean distance matrix and develop a novel matrix completion algorithm for ad hoc microphone array calibration along with theoretical guarantees.

Furthermore, the problem of source localization using ad hoc microphones in a reverberant enclosure is addressed. We incorporate the image model of multipath propagation for construction of a Euclidean distance matrix. The low-rank structure of the distance matrix is exploited to identify the support of the room impulse response function and its unique map to the source location. This approach enables single-channel and distributed source localization from asynchronous recordings provided by ad hoc microphones. Along this line, we address the problem of robust microphone array placement to optimize the localization performance.

Finally, spatial filtering techniques relying on beamforming are investigated for high quality speech acquisition and higher level applications. We develop beamformers for joint multi-speaker localization and voice activity detection. In addition, the broadband beampattern of a microphone array is characterized and its relation to predict the speech recognition accuracy is desired.

Key words: Ad hoc microphone array, Microphone calibration, Distributed source localization, Diffuse sound field , Euclidean distance matrix, Broadband beampattern, Voice activity detection, Reverberation, Distant speech recognition

Résumé

Les réseaux de microphones jouent un rôle central pour les applications utilisant des interfaces mains libres contrôlées par la voix. La tâche principale d'un réseau de microphones est de capter la parole distante en haute qualité. Un réseau de microphones peut acquérir de manière sélective les signaux de parole souhaités, en guidant le faisceau vers le locuteur désiré. Le développement prévisible d'applications d'intelligence ambiante motivées par l'abondance de périphériques disposant de microphones intégrés, tels que les ordinateurs ou téléphones portables, augmente l'importance de la recherche dédiée aux réseaux de microphones ad hoc. Les principales difficultés sont liées à la géométrie inconnue [du réseau de] microphones et aux enregistrements asynchrones.

Le but de cette thèse de doctorat est de traiter des problèmes relatifs aux localisations de microphone et de source, pour permettre la formation de faisceaux nécessaires aux tâches de traitement de parole de plus haut niveau. À cette fin, nous exploitons la connaissance a priori des structures acoustiques et géométriques au sein desquelles les nds du réseau ad hoc sont distribués, de faon à développer de nouveaux algorithmes de calibration de réseau de microphones et de localisation de source, ainsi que des techniques de formation de faisceaux pour les applications de parole distante.

Pour traiter du problème de la calibration de réseaux de microphones ad hoc, le modèle analytique de cohérence du champ sonore diffus est examiné, et ses propriétés fondamentales étudiées. Ce modèle permet l'estimation de distances par paires pour la calibration d'un réseau de microphones relativement compact. Nous développons le cadre mathématique pour l'estimation de longues distances par paires en exploitant les propriétés de bas rang de la matrice de distance euclidienne, et un algorithme novateur de complétion de matrice pour la calibration de réseaux de microphones ad hoc offrant des garanties théoriques.

Nous traitons également du problème de localisation de source en utilisant un réseau de microphones ad hoc dans une enceinte réverbérante. Nous incorporons le modèle d'image de la propagation multivoie pour la construction d'une matrice de distance euclidienne. La structure de bas rang de la matrice de distance est exploitée pour identifier le support de la fonction de réponse impulsionnelle de la salle, et sa relation unique à la position de la source. Cette approche permet la localisation de sources simples ou distribuées, à partir des enregistrements asynchrones fournit par un réseau de microphones ad hoc. De faon similaire, nous traitons le problème du positionnement robuste de réseau de microphones de faon à optimiser les performances de localisation. Enfin, des techniques de filtrage spatial reposant sur la formation de faisceaux sont examinées pour de l'acquisition de parole de haute

qualité et des applications de plus haut niveau. Nous développons des modeleurs de faisceaux (beamformers) pour effectuer de manie conjointe de la localisation de locuteurs multiples et de la détection d'activité vocale. De plus, nous caratérisons le diagramme de directivité (beampattern) à large bande d'un réseau de microphones, et nous établissons une relation permettant de prédire à partir de ses caratéristiques la précision de la reconnaissance vocale.

Mots-clés réseau de microphones ad hoc, calibration de microphones, localisation de sources distribuées, champ sonore diffus, matrice de distance euclidienne, diagramme de directivité large bande, détection d'activité vocale, réverbération, reconnaissance de parole distante

— French version translated by Blaise Potard as per the English version

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Microphone arrays are widely used to enable high quality distant audio acquisition. They are an essential part of a plethora of applications ranging from source localization [47, 22, 109, 6] and separation [8, 121] to distant speech recognition [104, 10, 7] and from sound field analysis and monitoring to virtual reality and surveillance [62, 120].

The spatial configuration of a microphone array enables high-quality acquisition of a desired speech signal relying on the principle of directional beam steering. This electronically steerable feature eliminates the need for an equipment to perform manual directional setting towards the speaker and offers several advantages for hands-free technologies. It can lead to mitigation of the noise and interferes regardless of the signal nature.

Resent advances in mobile computing and communication technologies motivates using cell phones, PDA's or tablets as a ubiquitous sensing platform leading to the emergence of ad hoc microphone arrays. Ad hoc microphone arrays consist of a set of sensor nodes spatially distributed over the acoustic field, in an ad hoc fashion thus eliminating the restrictions on microphone placement.

The distributed acquisition provides a flexible infrastructure for high quality sound acquisition, and requires to be calibrated to function. In this context, the following issues have made the application of ad hoc microphones challenging: (1) unknown geometry of the microphone array, (2) communication constraints between sensors and (3) asynchronous data acquisition. These issues pose technical challenges for array signal processing algorithms and they must be addressed from a research point of view to enable realistic applications and technological advances. Compared to the ad hoc sensor array, ad hoc acoustic microphones is a relatively new field of research while in many aspects it is inspired from the radar, sonar and antenna applications. Thereby, we approach the fundamental problems of ad hoc microphones considering the reverberant acoustics of an enclosure due to multipath propagation.

1.2 Big Picture

At the core of steered high quality acquisition schemes, the conventional beamforming techniques are impractical without sufficient prior information on microphones positions. Hence, in order to enable the effective use of the ad hoc microphones for speech applications, we need to perform calibration of the microphone positions and source localization to design beamforming for higher level speech applications.

The broad goal of this research is thus twofold:

1. Addressing the fundamental problems in effective aggregation of ad hoc microphone array data; these problems include specifically
   - ◇ Ad hoc microphones position calibration
   - ◇ Source localization
   - ◇ Design of microphone placement
2. Devising algorithms for microphone array processing to enable higher level applications where we focus on beamforming techniques for
   - ◇ Multi-source localization
   - ◇ Voice activity detection
   - ◇ Characterizing the performance for speech recognition

Figure 1.1 illustrates the building blocks of this research in a broad picture. The goal is to develop an ad hoc microphone front-end processing to enable higher level speech applications.

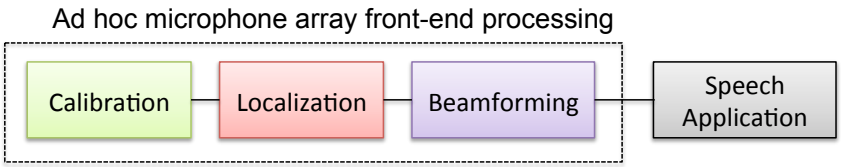

Figure 1.1: Broad picture of the research presented in this dissertation.

We assume that the recordings are synchronized and address the problem of finding the microphone positions; this problem is referred to as *calibration*. The precise knowledge of the microphones positions is then exploited for *localization* of the source. Spatial filtering via *beamforming* is applied to enable higher level applications with a focus on distant speech recognition problems in multi-source environments.

1.3 Thesis Statement

The goal of the present research is to develop an ad hoc microphone front-end processing to enable speech applications. As the ad hoc microphones are deployed in a chamber, we rely on the prior knowledge of the structures underlying the acoustic and geometry of the microphone array to address the fundamental problems of microphone calibration and source localization. We further propose novel methods for microphone array beamforming to enable multi-source localization and voice activity detection and characterize the speech recognition performance after beamforming.

The acoustic of the sound field in a reverberant room can be modeled as a diffuse field which possesses elegant mathematical properties. A key property of a diffuse field is that the correlation of the close-by microphones is a sinc function of the microphones' pairwise distance. Hence, we exploit this property to estimate the distances among the microphones. We study the fundamental limitation of a diffuse sound field for microphone pairwise distance estimation and derive the relation between the acoustic parameters of a room and the applicability of this model for distance estimation.

To estimate the pairwise distances of the far-apart microphones, we exploit the low-rank structure of the Euclidean distance matrix. This approach enables calibration of an arbitrary size ad hoc microphone array. To that end, novel algorithms are proposed for Euclidean distance matrix completion and the theoretical error bounds are rigorously studied for ad hoc microphone array calibration taking into account the connectivity of the network and the ratio of noise and missing distances.

Furthermore, we exploit the structure underlying multipath propagation to develop a distributed localization scheme. We show that the source-microphone distances can be estimated from the initial support of the room impulse response function and the low-rank structure of the Euclidean distance matrix enables finding the unique map between the source position and its echos while compensating the time offset for synchronization.

This research provides a novel perspective to incorporate the prior knowledge on a reverberant acoustic along with the underlying geometrical structure to devise novel methods for ad hoc microphone array calibration and localization. We further elucidate the theoretical implications of these structures for robust microphone placement. The acquired knowledge on the microphones' positions is essential for an effective beampattern steering and interference suppression in a multi-channel acquisition set-up.

To enable higher level applications, spatial filtering is the final necessary step of an ad hoc microphone front-end processing. Once the microphone array geometry is calibrated and assuming that the channels are synchronized and there is no communication constraint, it is possible to perform beamforming to capture the high-quality speech signal and mitigate the effect of noise and reverberation. Hence, we develop the approach for multi-source localization and voice activity detection. In this context, we characterize the broadband

beampattern to predict the speech recognition performance and to make some adjustments to the ad hoc microphones selection to achieve the highest acquisition quality.

The theoretical advancements offered in this thesis are often supported rigorously. We further compare and contrast our proposed methods with the state-of-the art counterparts. The empirical evaluation is based on the ground-truth of the microphones and speakers positions; the ad hoc microphone array data recordings were conducted at the Idiap instrumented meeting room.

1.3.1 Objectives

The key themes involved in this research are the followings

1. Applicability of diffuse field coherence model for microphone array calibration, its fundamental constrains and assumptions.
2. Euclidean distance matrix completion algorithm for calibration and its theoretical guarantees.
3. Joint localization and synchronization via Euclidean distance matrix recovery.
4. Microphone placement and a robust design to minimize the localization error.
5. Development of beamforming techniques for higher level speech application in multi source scenario.
6. Characterization the speech beampattern and quantifying the beamformer performance for higher level speech applications such as speech recognition.

The objective of this thesis is to enable speech application using ad hoc microphone array. To that end, our strategy is to incorporate the prior knowledge underlying the acoustic and geometrical structure of the problem to address the fundamental problems of microphone and source localization.

We rely on the properties of a diffuse sound field for estimation of the microphones pairwise distances. We provide a deep analysis of this approach and quantify its applicability and constrains for the real acoustics. The diffuse field coherence model enables calibration of compact microphone arrays and the pairwise distances of far-apart microphones remain missing.

To recover the missing distances, we study the application of matrix completion to exploit the low-rank structure of a squared distance matrix to reconstruct the missing components. This approach leads to the derivation of novel Euclidean distance matrix completion algorithm

and it demands theoretical support to bound the calibration error in a general scenario of any arbitrary design.

To tackle the problem of source localization, we further study the structure of the multipath propagation for single-channel and distributed source localization framework. In this context, the issue of asynchronous recording is addressed. Furthermore, we work out the robust design of microphone placement to minimize the localization error.

Finally, we study the methods for applying beamforming techniques for higher level speech applications where we focus on multi-source localization, voice activity detection and characterizing the broadband beampattern for quantifying its relation to speech recognition performance.

1.3.2 Contributions

The research presented in this dissertation features the following contributions:

- ◇ Quantifying the adequacy of diffuseness for pairwise distance estimation along with the fundamental limitation of diffuse field coherence model.
- ◇ Euclidean distance matrix completion algorithm for calibration of an ad hoc microphone array using partial measurements of the pairwise distances and its theoretical performance guarantees.
- ◇ Distributed and single-channel source localization algorithms exploiting the multipath propagation model for Euclidean distance matrix recovery in asynchronous recording scenario.
- ◇ Robust microphone placement algorithm for near-optimal source localization.
- ◇ Development of beamforming techniques for multi speaker localization and voice activity detection in a diffuse sound field.
- ◇ Characterization the broadband beampattern for speech acquisition and deriving its relation with speech recognition performance.

These contributions are communicated through the following publications:

JOURNAL PAPERS

1. *Enhanced diffuse field model for ad hoc microphone array calibration*, Mohammad J. Taghizadeh, Philip N. Garner and Hervé Bourlard, Signal Processing journal, volume 101, pages 242–255, 2014.

2. *Ad Hoc Microphone Array Calibration: Euclidean Distance Matrix Completion Algorithm and Theoretical Guarantees*, Mohammad J. Taghizadeh, Reza Parhizkar, Philip N. Garner and Hervé Bourlard, Afsaneh Asaei, Signal Processing, In Press, available online August 2014.

3. *Source Localization via Multipath Matrix Recovery*, Mohammad J. Taghizadeh, Afsaneh Asaei, Saeid Haghighatshoar, Philip N. Garner and Hervé Bourlard, IEEE Journal of Selected Topics in Signal Processing, accepted for publication, February, 2015.

CONFERENCE PAPERS

1. *Robust Microphone Placement for Source Localization from Noisy Distance Measurements*, Mohammad J. Taghizadeh, Saeid Haghighatshoar, Philip. N. Garner and Hervé Bourlard, submitted to IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

2. *Ad Hoc Microphone Array Calibration from Partial Distance Measurements*, Mohammad J. Taghizadeh , Afsaneh Asaei, Philip N. Garner and Hervé Bourlard, IEEE-ISCA Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 – **Nominated for the Best Student Paper Award**.

3. *Euclidean Distance Matrix Completion for ad hoc Microphone Array Calibration*, Mohammad J. Taghizadeh, Reza Parhizkar, Philip N. Garner and Hervé Bourlard, IEEE International Conference on Digital Signal Processing (DSP), 2013.

4. *Microphone Array Beam-pattern Characterization for Hands-free Speech Applications*, Mohammad J. Taghizadeh, Philip N. Garner and Hervé Bourlard, Proceeding of the 7th Sensor Array and Multi-channel Signal processing workshop (SAM), Hoboken, NJ, USA, 2012.

5. *An Integrated Framework for Multi-Channel Multi-Source Localization and Voice Activity Detection*, Mohammad J. Taghizadeh, Philip N. Garner, Hervé Bourlard, Hamid R. Abutalebi and Afsaneh Asaei, The third Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), Edinburgh, UK, 2011.

1.4 Thesis Outline

The introduction provided a review of the problem and motivations of this research. We outline specific objectives along with the contributions that feature this work.

In Chapter 2, we study the characteristics of a diffuse sound field. In particular, we study the relation between the measure of diffuseness and accuracy in pairwise distance estimation. In addition, we show how the size of the room affects the accuracy in distance estimation. Furthermore we propose methods for enabling the long pairwise microphone distance

estimation with reasonable accuracy. The proposed theories and algorithms are supported with simulations and experiments on real data.

In Chapter 3, we use the information learned from chapter 2 on pairwise distances for extracting the geometry of ad hoc microphone array. We propose a Euclidean distance matrix completion algorithm to recover the missing distances based on the known components of a distance matrix and its low-rank structure. The key objective is to extract the upper bound of the calibration error in noisy conditions and partially known pairwise distances.

In Chapter 4, we address the problem of source localization using a single microphone. Although, this may sound impossible, we show that by exploiting the multipath propagation model characterized by the image method, it is indeed possible to localize the source. In this chapter, we also discuss the synchronization problem and propose a new method for finding the correct impulse response of the room in an asynchronous scenario.

In chapter 5, a design problem is studied for robust placement of the microphones to minimize the error of source localization using time-of-arrival information. As the problem needs combinatorial search, optimal array geometry design is not feasible. Thus, we introduce a robust solution by minimizing the worst-case error. The proposed algorithm can be implemented efficiently and we verify its performance is the robust optimal design for source localization.

In chapter 6, spatial filtering techniques for beamforming are studied to enable speech applications. We develop a novel approach for multi-source localization which is capable of voice activity detection in a diffuse sound field and evaluate the joint performance on real data recordings.

In chapter 7, we address the speech recognition accuracy of beamforming data. This study lead to characterization of the broadband beampattern for speech acquisition by generalization of the narrowband beampattern. We extend this concept for superdirective and delay-and-sum beamforming, and finally show its clear relation to the speech recognition accuracy.

2 Enhanced Diffuse Field Model for Ad Hoc Microphone Array Calibration

In this chapter, we investigate the diffuse field coherence model for microphone array pairwise distance estimation. We study the fundamental constraints and assumptions underlying this approach and propose evaluation methodologies to measure the adequacy of diffuseness for microphone array calibration. In addition, an enhanced scheme based on coherence averaging and histogramming, is presented to improve the robustness and performance of the pairwise distance estimation approach. The proposed theories and algorithms are evaluated on simulated and real data recordings for calibration of microphone array geometry in an ad hoc set-up. The content of this chapter has been published in Elsevier journal of Signal Processing [113].

2.1 Introduction

State of the art calibration techniques can be grouped into three categories. The first approach relies on transmitting a known signal to perform microphone calibration. It may be noted that the knowledge of the source signal simplifies the estimation problems. If the signal is known beforehand, the time of arrival (ToA) of the source signal for each individual microphone is obtained through cross-correlation with the given signal. Hence, the negative effects of noise and reverberation are reduced as only one of the signals is noisy. Sachar et al. [97] presented an experimental setup using a pulsed acoustic excitation generated by five domed tweeters. The transmit times between speakers and microphones were used to find the relative geometry. Recent advances in mobile computing and communication technologies suggest using cell phones, PDA's or tablets as an ad hoc microphone array. Raykar et al. [94] used a maximum length sequence or chirp signal in a distributed computing platform. The time difference of arrival of the microphone signals were then computed by cross-correlation and used for estimating the microphone locations. Iterative nonlinear least-squares optimization removes noise and reverberation effect. Since the original signal is known, these techniques are robust to noise and reverberation.

The second category enables using an unknown signal. If the source signal is unknown,

the time difference of arrival (TDoA) for a pair of microphones is estimated through cross-correlation of the two microphone signals. Although TDoA-based methods can alleviate the need for activating a specific source signal or prior knowledge on the original signal, they may be more sensitive to noise and reverberation. The microphone calibration is usually integrated with source localization. Flanagan and Bell [53] proposed a method using the Weiss-Friedlander technique, where the sensor location and direction of arrival of the sources are estimated alternately until the algorithm converges. In an alternative approach to alleviate the requirement for a specific source signal, another approach was proposed by Chen et al. [32] where they introduced an energy-based method for joint microphone calibration and speaker localization. The energy of the signal is computed and a nonlinear optimization problem is formulated to perform maximum likelihood estimation of the source-sensor positions. This method requires several active sources for accurate localization and calibration.

McCowan et al. [78] proposed a calibration method based on the characteristics of a diffuse sound field model. This approach alleviates the need for activating several sources. A diffuse field can be roughly described as an acoustic field where the signals propagate with equal probability in all directions with the same power. The diffuse field is verified for meeting rooms and car environments [16] and it enables application of well-defined mathematical models for analysis of the acoustic field recordings. A particular property related to diffuse field recordings is the coherence function between pairwise microphone signals which is defined by a sinc function of the distance between the two microphones. Thereby, we can estimate the pairwise distances by least-squares fitting the computed coherence with the sinc function. This procedure is accomplished for each frame independently. To increase the robustness, the frame-based estimates are combined using k-means clustering [78]. This approach is applicable in a general room without the need for any explicit initialization or activating calibration signals. The study presented in this chapter is built on the idea of incorporating the properties of a diffuse field for ad hoc microphone array calibration.

The diffuse field has been studied rather extensively by many researchers with the aim of developing practical strategies for determining sound power, absorption measurements, and transmission loss. However, very few studies consider applicability of the associated models for microphone calibration. The purpose of this chapter is to investigate the fundamental hypotheses of the diffuse field model and to elucidate the limitations and the scope of its applicability. The study of sound fields in lightly damped enclosed spaces can be approached in two different ways. One is based on solving the wave equation with known boundary conditions, which leads to descriptions in terms of the modes of the room. The other approach relies on statistical models for analysis of the field and requires far less information about the room geometry. We apply both of these methods to highlight the requirements for application of the diffuse field model to enable microphone array calibration.

The chapter is organized as follows: The fundamentals of diffuse fields are studied in Section 2.2. We overview the characteristics and models of the diffuse field and the measurement for diffuseness. The methods to enhance the diffuse sound model are proposed

in Section 2.3 and applied in the framework of microphone array calibration in Section 2.4. The fundamental limitation of the diffuse model are explained in Section 2.5. The experimental analyses are presented in Section 2.6 and the conclusions are drawn in Section 2.7.

2.2 Diffuse Field Fundamentals

2.2.1 Definition of Diffuse Field

A diffuse field is defined as an acoustic field consisting of a superposition of an infinite number of sound waves traveling with random phases and amplitudes such that the energy density is equivalent at all points. In other words, all points in the field radiate equal power and random phase sound waves, with the same probability for all directions, and the field is homogeneous and isotropic [101]. A diffuse field can be realized if a point source is active in a highly echoic room. By removing the direct sound and the initial reflections from a recording of the sound, the remaining part consists of diffuse reflections. In addition, ambient distributed sound sources yield a diffuse field, while the interference phenomena near the room boundaries and corners raise the energy level and reduce the diffuseness. In a free space, having many uncorrelated sources distributed at long distances can generate a diffuse field.

The diffuse sound field at its theoretical level does not exist in practice. However, in many cases, a diffuse sound field can be a useful approximation of the real acoustic field in an enclosure. The important point is then to measure the amount of diffuseness and evaluate its adequacy for different applications. The analytic studies consider two points of view: (1) the wave equation based approach that describes diffuse field through the modes in a room and (2) the statistical approach by considering an infinite number of free propagation plane waves, referred to as the plane wave model.

2.2.2 Diffuse Field Model

Mode Model

This theory analyzes a room as a pack of resonators with bandwidth proportional to the absorption of the walls [36, 35]. The **3** dB bandwidth of the mode is given by

$$\boldsymbol{B}_{\mathbf{3}\mathrm{dB}} = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{\pi}\boldsymbol{\tau}}, \tag{2.1}$$

where $\boldsymbol{\tau}$ corresponds to the decaying time constant of the sound field energy [24].

By solving the equations of a homogeneous sound field with boundary conditions, we extract normal modes for the room. Each mode indicates a resonance frequency, and the distribution of these frequencies is determined by the shape and dimension of the room [100, 126]. At high

frequencies f, the mode density depends solely on the room volume V as expressed through

$$\gamma(f) = \frac{4\pi V}{c^3} f^2, \tag{2.2}$$

where c denotes the speed of sound. The modal overlap is defined as the average number of modes excited by a pure tone, and it is given by

$$\eta(f) = \gamma(f) B_{3dB} = \frac{4\pi V}{c^3} f^2 \frac{1}{2\pi\tau} \tag{2.3}$$

If the pure tone is close to the frequency of the mode, within a bandwidth of $2.2/T_{60}$, the adjacent mode is excited; T_{60} is equal to the time required for the level of a steady sound to decay by 60 dB after the sound has stopped. If $\eta(f) \geq 3$[1], there are enough excited modes to generate a diffuse field in the room [85], hence the critical frequency to achieve diffuseness is obtained as

$$f_s = \sqrt{\frac{3c^3\tau}{2V}} \tag{2.4}$$

This frequency is known as the Schroeder frequency [99, 98].

Plane Wave Model

An alternative analysis approach, which does not need acoustic information, relies on a statistical model. In the plane wave model or the statistical model, a diffuse field is defined by the superposition of a large set of plane waves impinging from all directions. We consider the steady state sound field generated by a pure tone source in a reverberant room. The time domain sound pressure $P(t)$ at a point far from the walls and the source is expressed as

$$P(t) = \lim_{q\to\infty} q^{-1/2} \sum_{i=1}^{q} b_i \cos(\omega t + \varphi_i), \tag{2.5}$$

where b_i and φ_i are random variables and independent of each other; φ_i has a uniform distribution in $[0, 2\pi]$ and b_i has a normal distribution; ω denotes the angular frequency and q is the number of plane waves. Each point in the field receives sound pressure from all directions [99]. Considering this spatial uniformity, we can compute an average sound pressure through

$$P(t) = \lim_{q,m\to\infty} (qm)^{-1/2} \sum_{j=1}^{m} \sum_{i=1}^{q} b_{ij} \cos(\omega t + \varphi_{ij}), \tag{2.6}$$

[1] Deriving the 3D modes in a rectangular room, a decomposition of an oblique mode into eight plane waves can be obtained. Hence, for Υ model overlap, we get 8Υ plane waves. Some heuristics indicate that 24 plane waves is a lower bound for generating diffuse sound, therefore $\Upsilon = 3$ is the smallest value to achieve diffuseness as considered in Schroeder frequency (2.4).

where m is the number of different directions from which plane waves impinge on a point in the field. In three dimensions, the distribution of the plane waves is such that there is at least one plane wave at each $4\pi/m$ steradian. The plane wave model is particularly useful at medium to high frequencies; it requires no details about the room geometry. The accuracy however, degrades at low frequencies and the effects of interference is ignored. Waterhouse [127] extended this approach by considering the interference phenomena that occur near the walls. The studies in this chapter rely on the basic mode model and the plane wave model.

2.3 Enhanced Diffuse Field Model

2.3.1 Averaging the Coherence Function

Cross Correlation

The correlation function of the sound pressures at two points in an acoustic field is defined as

$$C = \frac{\int_0^T P_1(t)P_2(t)dt}{\sqrt{\int_0^T P_1^2(t)dt \int_0^T P_2^2(t)dt}}. \tag{2.7}$$

The cross correlation function in a diffuse field has a closed form analytic solution [34, 92]. Suppose a plane wave passes two points located on the z-axis with separation d, the correlation function would be $\cos(\kappa d \cos\phi)$ where κ is the wavenumber and ϕ is the polar angle defined as the angle between the wave front and the line connecting the two points [82]. The value of C for a diffuse field can be obtained by averaging the cross correlation for all directions, as

$$\begin{aligned} C &= \int_0^{\pi} \int_0^{2\pi} \cos(\kappa d \cos\phi) \sin\phi \, d\theta \, d\phi/4\pi \\ &= \sin(\kappa d)/(\kappa d), \end{aligned} \tag{2.8}$$

where θ is the azimuth angle.

Coherence Averaging

We consider a scenario in which n microphones record a diffuse field pressure signal. Suppose that S_i and S_l represent the spectral representation of the signals in Fourier domain at microphones i and l respectively. The cross spectral density is

$$\Phi_{il}(\omega) = S_i(\omega)S_l^*(\omega), \tag{2.9}$$

where "$*$" is the conjugate transpose operator. The coherence of two signals is the cross spectrum normalized by the square roots of the auto spectra, defined concisely as

$$\mathbf{\Gamma}_{il}(\omega) = \frac{\mathbf{\Phi}_{il}(\omega)}{\sqrt{\mathbf{\Phi}_{ii}(\omega)\mathbf{\Phi}_{ll}(\omega)}}. \tag{2.10}$$

In a perfect diffuse field, at each frequency component, the coherence is a sinc function, which holds if long time averaging (2.7) is taken [38]. As the frequency analysis is conducted on short frames, we propose to collect several frames and take an average over the frame-based coherence to achieve an estimate conforming to the sinc model. Therefore, we define an average coherence function as

$$\tilde{\mathbf{\Gamma}}_{il}(\omega) = \frac{\mathbf{1}}{\boldsymbol{J}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{J}} \Re\left(\mathbf{\Gamma}_{il}^{\boldsymbol{j}}(\omega)\right) = \operatorname{sinc}\left(\frac{\omega \boldsymbol{d}_{il}}{\boldsymbol{c}}\right), \tag{2.11}$$

where the operator $\Re(.)$ takes the real part of its argument; $\boldsymbol{d}_{il}$ is the distance between the two microphones, $\boldsymbol{j}$ denotes the frame index and $\boldsymbol{J}$ is the total number of frames. Based on this model, estimation of the distance between two microphones is possible by fitting a sinc function to the coherence of their signals. The conventional approach applies sinc function fitting on a frame-basis [78]. The theory asserted in this section suggests that an averaging method can improve pairwise distance estimation. We elaborate on the empirical evidence to verify this hypothesis in Section 2.6.

2.3.2 Boosting the Power

The theory of diffuse field analysis is developed under the assumption that the contribution of air absorption to the total enclosure absorption is negligible. In a silent room, where a diffuse field is generated by the ambient sources such as running devices, computers, etc., the amplitude of the source signal is very weak. Therefore the prohibitive cost of air absorption affects the energy distribution. This condition tends to violate the necessary assumption of negligible energy loss during a mean free propagation. Hence, we propose to provide additional sources in a particular set up to boost the sound field power. The diffuse field is better realized for high frequencies, as more modes are excited leading to an increase in the number of plane waves (Table 2.1). However the air absorption also increases with frequency; the acoustic intensity[2] of a plane wave as a function of the propagation distance $\boldsymbol{r}$ is expressed as

$$\boldsymbol{I}(\boldsymbol{r},\omega) = \boldsymbol{I_0}(\omega)\mathrm{e}^{-\boldsymbol{r}/\boldsymbol{\xi}(\omega)}, \tag{2.12}$$

where $\boldsymbol{I}(\boldsymbol{r},\omega)$ is the intensity $\boldsymbol{r}$ meters from the source, $\boldsymbol{I_0}(\omega)$ is the original intensity of the source with frequency ω and $\mathbf{1}/\boldsymbol{\xi}(\omega)$ is the attenuation factor, which increases with frequency. Therefore, if the source has a very low power, the high frequencies can diminish and the low

[2] Sound power per unit area.

band index	$f_1 - f_2$ Hz	#modes/band (medium)	#modes/band (large)	#modes/band (very large)
1	4.44-5.6	0	0	1
2	5.6-7.1	0	1	2
3	7.1-9	0	0	3
4	9-11.2	0	1	6
5	11.2-14	0	1	6
6	14-18	0	4	20
7	18-22.4	1	6	25
8	22.4-28	0	6	53
9	28-35.5	1	19	100
10	35.5-45	2	29	205
11	45-56	3	50	340
12	56-71	7	100	734
13	71-90	9	205	1458
14	90-112	18	340	2589
15	112-140	30	684	5054
16	140-180	68	1508	11127
17	180-224	115	2589	15754
18	224-280	206	5054	39132
19	280-355	440	10611	82724

Table 2.1: Number of modes in the one-third-octave bands for medium size room ($\mathbf{8} \times \mathbf{5.5} \times \mathbf{3.5}\,\mathrm{m}^3$), large size room ($\mathbf{24} \times \mathbf{16.5} \times \mathbf{10.5}\,\mathrm{m}^3$) and very large size room ($\mathbf{48} \times \mathbf{33} \times \mathbf{21}\,\mathrm{m}^3$).

frequencies, which do not excite enough resonance modes, remain in the sound field. This phenomenon reduces the diffuseness. Hence, we speculate that increasing the energy of the sound field yields higher diffuseness, and enables more accurate distance estimation. This idea has been evaluated empirically through the experiments conducted in Section 2.6.3.

2.3.3 Diffuseness Evaluation

Broadband Power Pattern

We consider a well-designed symmetric and regular spherical array of $\boldsymbol{n}$ microphones. The spectral representation of the signals recorded by microphone array in Fourier domain is denoted by $\mathscr{S}(\omega) = [S_1(\omega), S_2(\omega), \ldots, S_n(\omega)]^T$. Suppose that the beamformer weights steered towards direction $\boldsymbol{a}(\theta, \phi)$ is represented by

$$\mathscr{F}(\omega, \boldsymbol{a}(\theta, \phi)) = [F_1(\omega, \boldsymbol{a}(\theta, \phi)), F_2(\omega, \boldsymbol{a}(\theta, \phi)), \ldots, F_n(\omega, \boldsymbol{a}(\theta, \phi))], \tag{2.13}$$

the response of the array by applying the beamformer would be

$$\boldsymbol{Y}(\omega, \boldsymbol{a}(\theta, \phi)) = \mathscr{F}(\omega, \boldsymbol{a}(\theta, \phi))\mathscr{S}(\omega). \tag{2.14}$$

Given $\boldsymbol{Y}$, the directional power can be measured as $\boldsymbol{Y^2(\omega, a(\theta,\phi))}$. The directional power can be used to evaluate the level of diffuseness. As stated in Section 2.2, the power in a diffuse field is isotropic, which indicates equal power accumulated from all directions.

In a broadband diffuse field, we can apply a filter to improve the model fitting by restricting the broadband processing to frequencies conforming to the theoretical diffuseness bounds. The enhanced model can then be evaluated in terms of the isotropic power distribution using a broadband beamformer. Given $\boldsymbol{Y}$, the beamformer output for the spectrum of signal, the broadband beampattern is given by

$$\boldsymbol{B(a(\theta,\phi)) = \Lambda\sqrt{\int_{\Omega} Y^2(\omega, a(\theta,\phi))\, d\omega},} \tag{2.15}$$

where $\boldsymbol{\Omega = [\omega_{\min}, \omega_{\max}]}$ is the frequency band of the signal and $\boldsymbol{\Lambda}$ is a normalization factor given by

$$\boldsymbol{\Lambda^{-1} = \max_{\theta,\phi} \sqrt{\int_{\Omega} Y^2(\omega, a(\theta,\phi))\, d\omega}\,.} \tag{2.16}$$

We can see that the broadband pattern can be interpreted as a weighted average of the beamformer's output over the broadband spectrum [110]. Accordingly, the broadband power-pattern would be

$$\boldsymbol{\mathscr{P}(a(\theta,\phi)) = |B(a(\theta,\phi))|^2.} \tag{2.17}$$

Diffuseness Evaluation Measure

The appropriate application-specific criterion is necessary to evaluate the adequacy of the diffuseness. In this section, we propose a novel approach for evaluating the diffuseness in the room to assess the diffuseness adequacy for estimating pairwise distances. For the particular application of microphone calibration, a *pointwise* diffuseness is important, which indicates that the angular distribution of the power at any point is equal in all directions.

To measure the signal power, we propose to use a superdirective beamformer by steering the beam toward several representative directions of the space. In real scenarios, the ambient sound source in the environment does not have the same power at all frequencies, so it is crucial to consider the broadband power-pattern as explained in Section 2.3.3. After normalization, we have to compare the three-dimensional (3D) pattern to a sphere of radius one. To obtain the broadband power-pattern at a particular point $\boldsymbol{A}$ in space, the microphone array has to be centered at $\boldsymbol{A}$. Hence, the diffuseness level is defined as

$$\boldsymbol{\mathscr{X}_A = \frac{3}{4\pi} \iiint_{\mathscr{P}_A(\theta,\phi)} \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta,} \tag{2.18}$$

where $\mathscr{P}_A(\theta,\phi)$ is the measure of the power stated in (A.7) that is received from a direction with azimuth θ and polar angle ϕ in the diffuse field at point A; ρ denotes the radial distance in the Spherical coordinate system. $\mathscr{X}_A$ equals **1** if we have a complete diffuse field at point A.

Computation of $\mathscr{P}_A(\theta,\phi)$ is not easy and we need a 3D microphone array with a carefully designed symmetric and regular geometry. To simplify this computation, we consider reducing the 3D pattern to 2D by averaging over all angles ϕ. By defining $Q_A(\theta)$ as a 2D approximation of the 3D pattern $\mathscr{P}_A(\theta,\phi)$ through

$$Q_A(\theta) = \int_0^{\pi} \int_0^{\mathscr{P}_A(\theta,\phi)} \rho \, d\rho \, d\phi, \tag{2.19}$$

and

$$\nu = \max_{\theta} Q_A(\theta), \tag{2.20}$$

we derive $\widetilde{\mathscr{X}_A}$ as an approximation of $\mathscr{X}_A$ through

$$\widetilde{\mathscr{X}_A} = \frac{1}{\pi \nu^2} \int_0^{2\pi} \int_0^{Q_A(\theta)} r \, dr \, d\theta. \tag{2.21}$$

The approximated quantity $\widetilde{\mathscr{X}_A}$ is more practical, and it has enough accuracy for our application as we investigate numerically in Sections 2.6.3 and 2.6.4. The conventional methods consider mere sphericity and roundness to measure the level of diffuseness [55] whereas the proposed method is capable of directly measuring the isotropic sound field power at any given point in space; hence, the proposed diffuseness measure yields more accurate results.

2.4 Ad Hoc Microphone Array Calibration

2.4.1 Conventional Method

The following objective measure has been used to fit a sinc function for a broadband spectrum of coherence function and estimate the pairwise distance [78]

$$\delta_{il}^{j}(d) = \int_{\omega_{min}}^{\omega_{max}} \left| \Re\{\Gamma_{il}^{j}(\omega)\} - \mathrm{sinc}\left(\frac{\omega d}{c}\right) \right|^2 d\omega, \tag{2.22}$$

By minimizing $\delta_{il}^{j}(d)$ over d, we obtain an estimate $\tilde{d}_{il}^{j}$ per frame

$$\tilde{d}_{il}^{j} = \arg\min_{d} \delta_{il}^{j}(d). \tag{2.23}$$

The pairwise distance has been estimated for each frame of the sound signal. To improve the estimation accuracy, the estimates of multiple frames are combined using k-means clustering

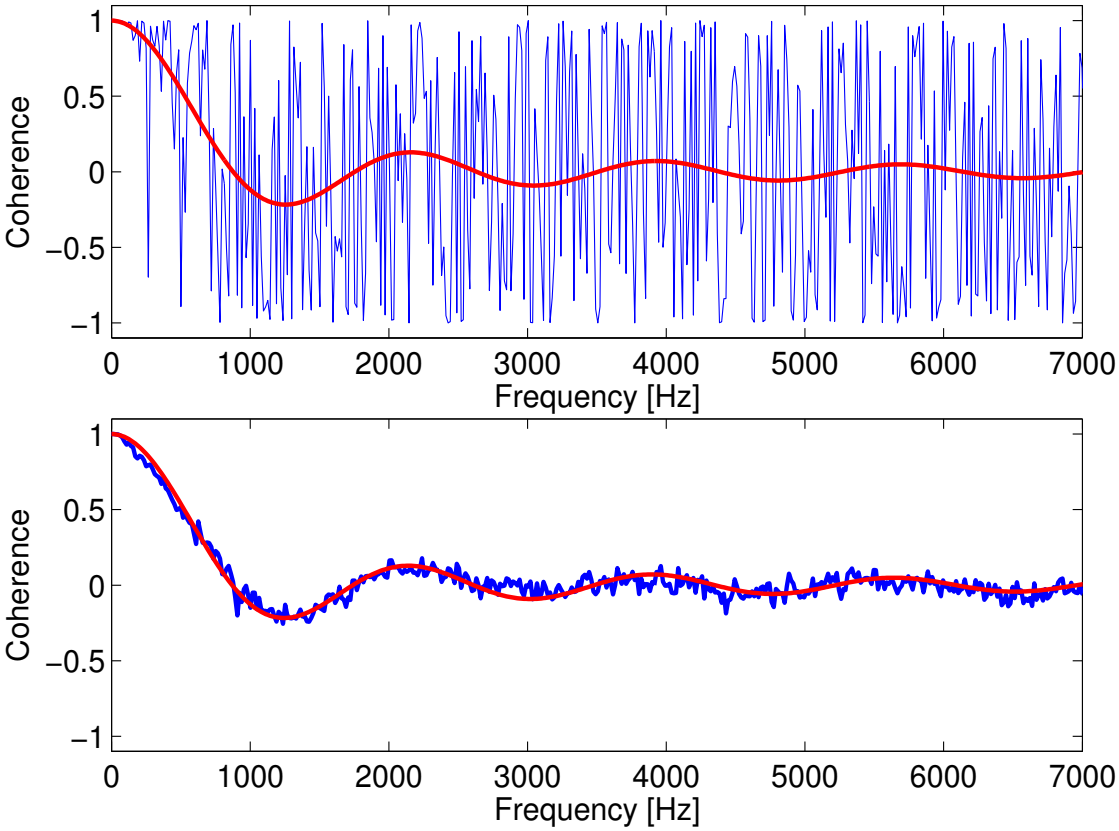

Figure 2.1: (Top) Fitting a sinc function (red) on one frame of diffuse field coherence (blue); the correct distance is 20 cm and the estimated distance is 19.3 cm. (bottom) Fitting a sinc function on average of 100 frames of diffuse sound field coherence; the estimated distance is 19.8 cm.

to remove the large-error estimates by grouping the points in two clusters. The clustering step is costly and requires long recorded signals to enable accurate estimation.

2.4.2 Proposed Averaging Method

The theoretical analysis carried out in Section 2.3.1 showed that the coherence function of a long segment is a sinc function and this model is not exact for a single frame. To obtain a sinc function model, we need to average over a sequence of frames. Figure 2.1, shows empirical evidence for this argument, and supports the requirement for averaging prior to fitting. The nonlinear characteristic and quick damping of the sinc function can lead to huge errors by only slight deviation from a diffuse field.

The averaging of the coherence of multiple frames prior to fitting the sinc function requires fewer frames than the clustering approach, and is very effective to improve the pairwise distance estimation performance. To state it more precisely, we consider J frames to extract

the distance between two microphones $\boldsymbol{i}$ and $\boldsymbol{l}$. The averaging method is expressed as

$$\begin{aligned} \boldsymbol{\delta}_{il}(\boldsymbol{d}) &= \int_{\omega_{min}}^{\omega_{max}} \left| \left[\frac{\mathbf{1}}{\boldsymbol{J}} \sum_{j=1}^{J} \Re\left(\boldsymbol{\Gamma}_{il}^{j}(\boldsymbol{\omega})\right) \right] - \operatorname{sinc}\left(\frac{\boldsymbol{\omega d}}{\boldsymbol{c}}\right) \right|^{\mathbf{2}} \boldsymbol{d\omega}\,, \\ \tilde{\boldsymbol{d}}_{il} &= \underset{\boldsymbol{d}}{\mathbf{arg\,min}}\ \boldsymbol{\delta}_{il}(\boldsymbol{d}) \end{aligned} \tag{2.24}$$

2.4.3 Outlier Detection Techniques

In practice, there is no complete diffuseness and the characteristic of the sound field changes due to irregularities and acoustic ambiguities. This phenomenon results in outlier observations in the coherence function which lead to a high error in pairwise distance estimation. Hence, we propose to apply an outlier detection technique after averaging the coherence of multiple frames. The goal of outlier detection is to increase the quality and robustness of a data analysis approach.

We consider statistical outlier detection techniques based on k-means (parametric-based) as well as histogram (non-parametric) methods. In the parametric approach, we consider a profile and unsupervised learning with certain criteria to identify the outliers in pairwise distance estimation. More experiments show that the erroneous estimates do not conform to a specific parametric model. Hence, we resort to a non-parametric histogram-based approach. In the histogramming method, the outliers are identified through a fixed threshold. In addition, this method requires less memory and computational cost, although finding the optimal size of the bins for a large number of attributes is a challenging task. The experimental analyses conducted in sections 2.6.2 and 2.6.4 confirm the validity of the averaging method followed by outlier removal using histogram-based clustering for robust estimation of the pairwise distance. Furthermore, we show that histogram clustering outperforms the k-means clustering approach. At the final step in calibration of the microphones, the geometry is extracted using the s-stress method [40].

2.5 Fundamental Limitation of Diffuse Model

This section explains the fundamental limitations and the performance bound of distance estimation using a diffuse field coherence model. As we have already seen earlier in the chapter, the spatial coherence of two signals in a diffuse field is a sinc function of the pairwise distance (2.11). This function decreases quickly and, as shown in Figure 2.1, it disappears after one cycle. Hence, the coherence measured in the first cycle is vital in estimation accuracy.

We consider three scenarios, being a medium size room ($\mathbf{8}\times\mathbf{5.5}\times\mathbf{3.5}\,\mathrm{m}^3$), a large size room ($\mathbf{24}\times\mathbf{16.5}\times\mathbf{10.5}\,\mathrm{m}^3$) and a very large size room ($\mathbf{48}\times\mathbf{33}\times\mathbf{21}\,\mathrm{m}^3$). The second zero crossings on the sinc function as expressed in (2.11) occur at **343** Hz, **114** Hz and **57** Hz for pairwise distances of **1** m, **3** m and **6** m, respectively. Hence, diffuseness at frequencies lower than these

frequencies are important.

On the other hand, the Schroeder frequency is obtained as $f_s = \sqrt{6c^2/\alpha Z}$ where α is the average absorption coefficient of the walls with a surface area of Z [98]; therefore, for an average absorption coefficient $\alpha = 0.07$ and $c = 343$ m/s, the Schroeder frequencies for these three rooms are **235** Hz, **78** Hz and **39** Hz respectively. As indicated in Section 2.2.2, a diffuse field cannot be generated in a room with a monochromatic source under the Schroeder frequency.

The mode model can be used for computing the acoustic pressure in modal behavior. Diffuseness at each frequency band can be illustrated by expansion modes. Table 2.1 summarizes the number of modes for each one-third-octave band in three room sizes. Based on theory, we hypothesize that increasing the dimension of the room increases the diffuseness, in particular at low frequencies which are highly effective in distance estimation. In addition, by increasing the pairwise distances, the number of discrete frequencies below the second zero crossing decreases linearly so we speculate that a linear regression can illustrate the relationship between the errors and distances. The empirical evaluations carried out in Sections 2.6.3–2.6.6 confirm the validity of these hypotheses. These experiments enable formulating a relation between room dimension and achievable distance estimation.

2.6 Experimental Analysis

This section presents the numerical results to evaluate the proposed theories and hypotheses. The microphone calibration performance measure must be robust to rigid transformations (translation, rotation and reflection). Hence, we use the distance between the actual locations X and estimated locations $\hat{X}$ as defined in [18]

$$\begin{aligned} \text{dist}(X, \hat{X}) &= \frac{1}{n} \left\| LXX^T L - L\hat{X}\hat{X}^T L \right\|_{\text{F}} , \\ L &= \mathbb{I}_n - (1/n)\mathbf{1}_n \mathbf{1}_n^T, \end{aligned} \tag{2.25}$$

where $\|\cdot\|_{\text{F}}$ denotes the matrix Frobenius norm. The $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector, $\mathbb{I}_n$ is the $n \times n$ identity matrix and $X, \hat{X} \in \mathbb{R}^{n \times \eta}$, where η is the dimension of the space. The distance measure stated in (3.15) is useful to compare the performance of different methods when the microphone array geometry is fixed.

2.6.1 Data Recording Set-up

Simulation Scenarios

We simulate a medium size room of dimensions $8 \times 5.5 \times 3.5\,\text{m}^3$, which has the same dimension of the room in the real scenario. The room is equipped with 48 omni-directional loudspeakers playing independent white Gaussian noise. These are divided into 3 uniform

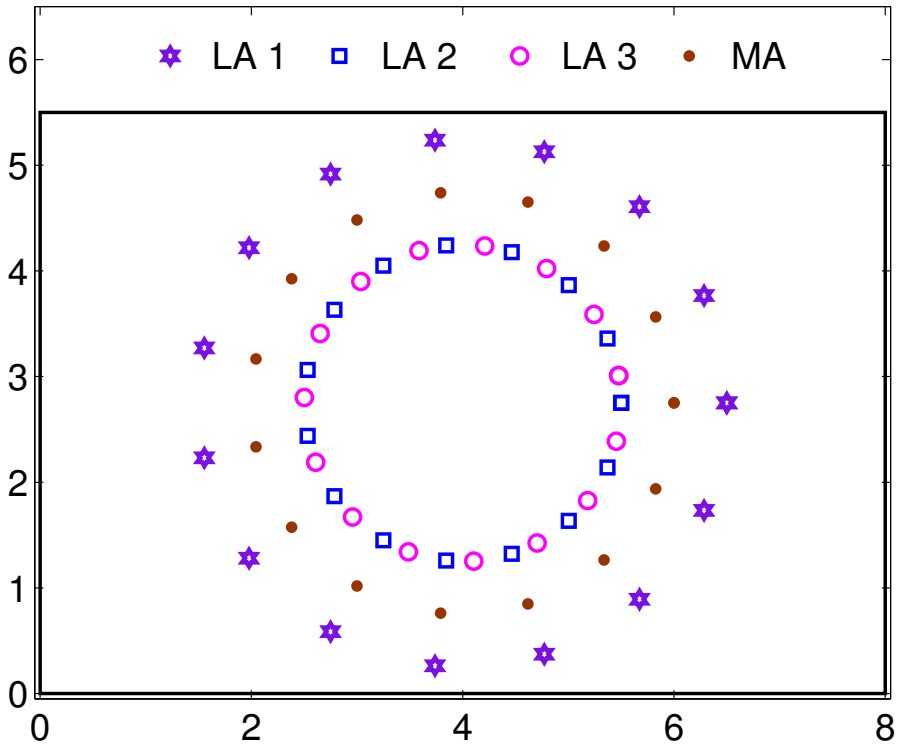

Figure 2.2: Top view of the simulated medium size room scenario: This scenario consists of three circular 16-element omni-directional loudspeaker arrays (LA) and one circular microphone array (MA) with the following set-up parameters: LA1 has diameter=2.5 m located at height=1.75 m; LA2 and LA3 have diameters=1.5 m located at height=0.1 m and 3.4 m respectively. A 16-element microphone array is depicted with diameter=2 m and it is located at height=1.75 m. All arrays are parallel to the floor. The number of microphones and the diameter of the MA are varied as explained in Section 2.6.1 to generate various pairwise distances.

circular arrays with diameters of **1.5** m, **2.5** m and **1.5** m, producing the sound field. The three circular loudspeaker arrays are parallel to the floor and located at the center of the planar area of the room at **0.1** m, **1.75** m and **3.4** m height. A uniform 8-channel circular microphone array located at center of the room is used to record the sound field. The diameter of the array is adjusted such that the pairwise distance between the microphones is equal to $\{\mathbf{0.1}, \mathbf{0.2}, \mathbf{0.3}, \ldots, \mathbf{0.8}\}$ m; that corresponds to the microphone array diameters of $\{\mathbf{0.26}, \mathbf{0.52}, \mathbf{0.78}, \mathbf{1.04}, \mathbf{1.30}, \mathbf{1.57}, \mathbf{1.83}, \mathbf{2.10}\}$ m. To enable evaluations for larger distances beyond **0.8** m, the 8-channel array is replaced with a 16-channel uniform circular microphone array with a diameter equal to $\{\mathbf{0.9}, \mathbf{1}, \ldots, \mathbf{2}\}$ m. Figure 2.2 depicts a top view of the simulated scenario. In addition, for investigation of the effect of room dimension on diffuseness of the field, and distance estimation, a large room as well as a very large room of dimensions $\mathbf{24} \times \mathbf{16.5} \times \mathbf{10.5}\,\mathrm{m}^3$ and $\mathbf{48} \times \mathbf{33} \times \mathbf{21}\,\mathrm{m}^3$ such that each dimension is **3** and **6** times bigger than real scenario are simulated. The same set-up of loudspeakers are used where the diameters are expanded by a factor of **3** and **6**. The same microphone array is used to record the sound field.

The diameter of the array is adjusted such that the pairwise distance between the microphones are varied from $\mathbf{0.1}$m to $\mathbf{10}$m.

The room impulse responses are generated with the image source model [4] using intra-sample interpolation up to $\mathbf{15}^{th}$ order reflections. The corresponding reflection ratio, $\boldsymbol{\beta}$ used by the image model was calculated via Eyring's formula:

$$\boldsymbol{\beta} = \mathbf{exp}(-\mathbf{13.82}/[\boldsymbol{c} \times (\boldsymbol{L}_x^{-1} + \boldsymbol{L}_y^{-1} + \boldsymbol{L}_z^{-1}) \times \boldsymbol{T}_{\mathbf{60}}]), \tag{2.26}$$

where $\boldsymbol{L}_x, \boldsymbol{L}_y$ and $\boldsymbol{L}_z$ are the room dimensions. The temperature of the room is assumed to be $\mathbf{20}^\circ$ Celsius, thus $\boldsymbol{c} = \mathbf{343}$m/s. In our experiments, $\boldsymbol{T}_{\mathbf{60}} = \mathbf{300}$ms for the medium size room. The direct-path propagation is discarded from the impulse response for generating a diffuse sound field [109].

Real Data Scenario

In addition to the simulated recordings, we use the geometrical setup of the MONC corpus to record the sound field in a meeting room [1]. The enclosure is a $\mathbf{8} \times \mathbf{5.5} \times \mathbf{3.5}\,\mathrm{m}^3$ rectangular room and it is moderately reverberant. It contains a centrally located $\mathbf{4.8} \times \mathbf{1.2}\,\mathrm{m}^2$ rectangular table. Twelve microphones are located on a planar area parallel to the floor at height of $\mathbf{1.15}$m: Eight of them are located on a circle with diameter $\mathbf{20}$cm and one microphone is at the center. There are three additional microphones at a $\mathbf{70}$cm distance from the central microphone. The microphones are Sennheiser MKE-2-5-C omnidirectional miniature lapel microphones. The floor of the room is covered with carpet and surrounded with plaster walls and two big windows.

The recordings were made in two scenarios: (1) Collecting the diffuse sound field of ambient noise in the room without any additional source and (2) playing extra sounds by putting two small loudspeaker under the table, and covering them with anti-acoustic material, so that the direct paths between loudspeaker and microphones are prohibited to ensure diffuseness. The microphone placement is depicted in Figure 2.3. The sampling rate is $\mathbf{48}$kHz while the processing applied for microphone calibration is based on a down-sampled signal at rate $\mathbf{16}$kHz to reduce the computational cost. The experiments are conducted using $\boldsymbol{c} = \mathbf{343}$m/s that corresponds to $\mathbf{20}^\circ$ Celsius temperature of the room.

2.6.2 Averaged Coherence Function

Figure 2.1 shows a real data example of the coherence of one frame (top) and the coherence function averaged over 100 frames (bottom) along with the fitted sinc function. As we can see, averaging is crucial prior to fitting the model by least square regression. The conventional method [78] fitted a sinc function on a single frame followed by k-means clustering of multiple frames to determine the distance. The numerical results show that the error of fitting a sinc function on the averaged coherence function is $\mathbf{35}$ times smaller than the conventional

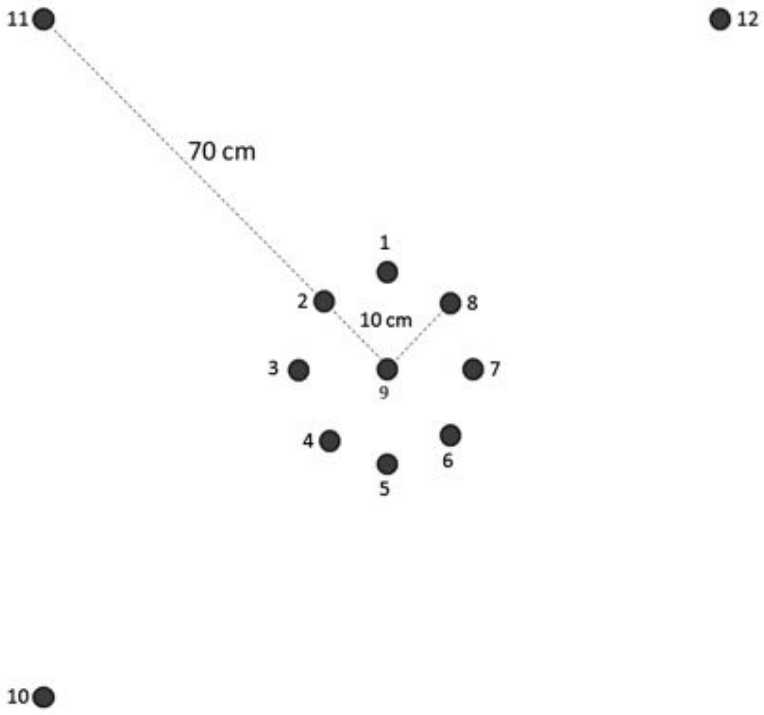

Figure 2.3: Microphone placement for real data recording scenario.

method for small distances. Furthermore, this method speeds up the calibration by a factor of **60** compared to the k-means clustering method in terms of CPU time using the same number of frames.

2.6.3 Diffuseness Evaluation

The first experiments consider measuring the diffuseness with the method proposed in section 2.3.3. A superdirective beamformer was used for measuring the power of the received signal from all directions. Figure 2.4 shows the patterns for the simulated very large room at distances 2 m (top) and **5** m (bottom) from the room center. We can see that a more isotropic power is obtained if the point of measurement is closer to the room center. Based on the definition stated in (2.21), the diffuseness levels at 2m and 5m distances from the center of the room are **.92** and **.84** respectively, which shows that the diffuseness reduces as we get closer to the borders. The diffuseness for the real data recorded at the meeting room without additional sources is measured as **0.70**. We increased the power by playing white Gaussian noise from the two small loudspeakers put under the table. Figure 2.5 shows the pattern with the proposed sound field augmenting method compared to the initial recordings. A more isotropic sound field is obtained as the pattern is closer to a circle. Quantitatively, the diffuseness is improved to **0.83**, that indicates a **19%** increase in diffuseness level.

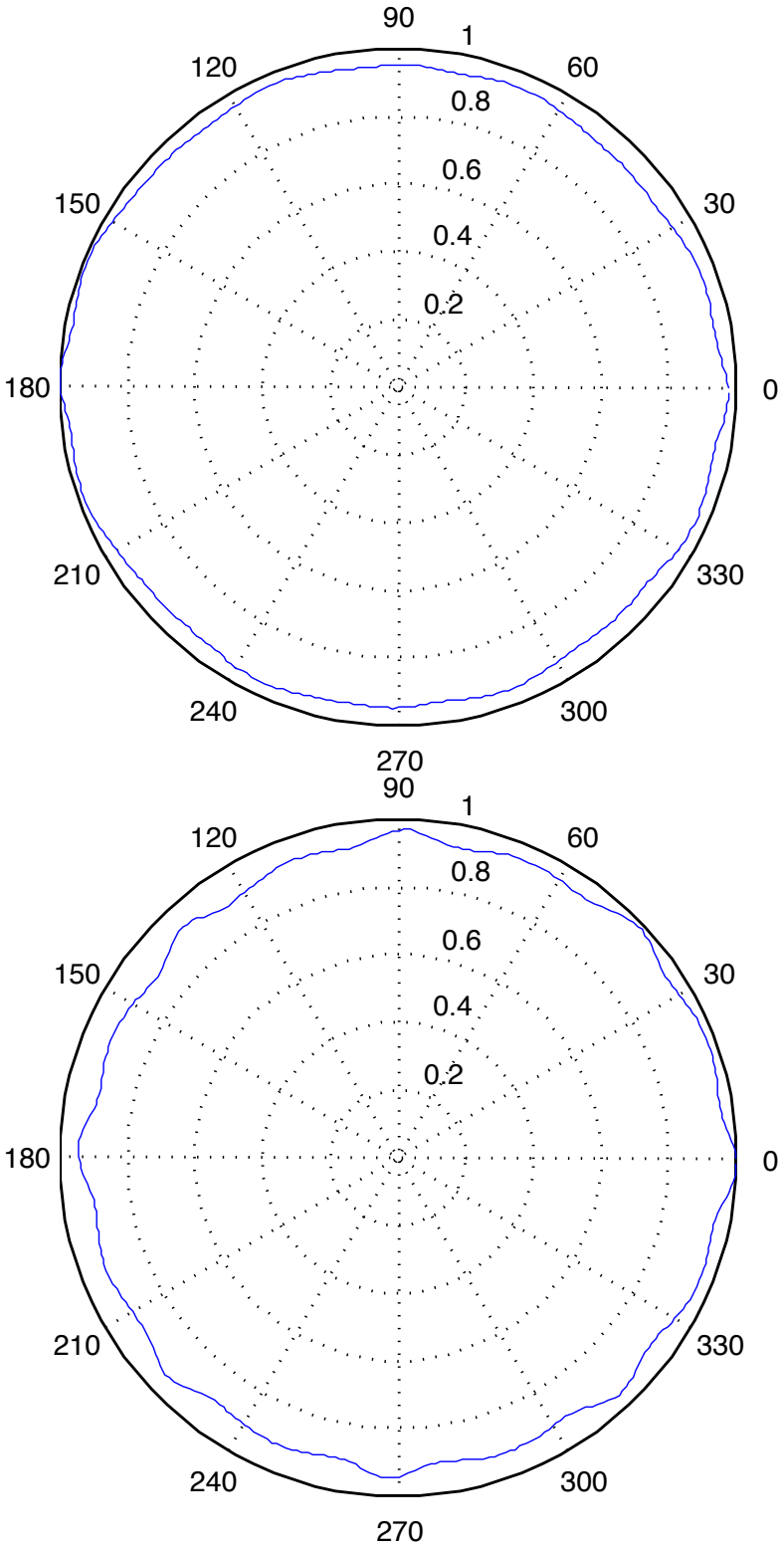

Figure 2.4: Broadband power-pattern obtained at 2m (top) and 5m (bottom) from center of the room by averaging over all polar angles; the scenario is synthesized in a very large room using 48 loudspeakers.

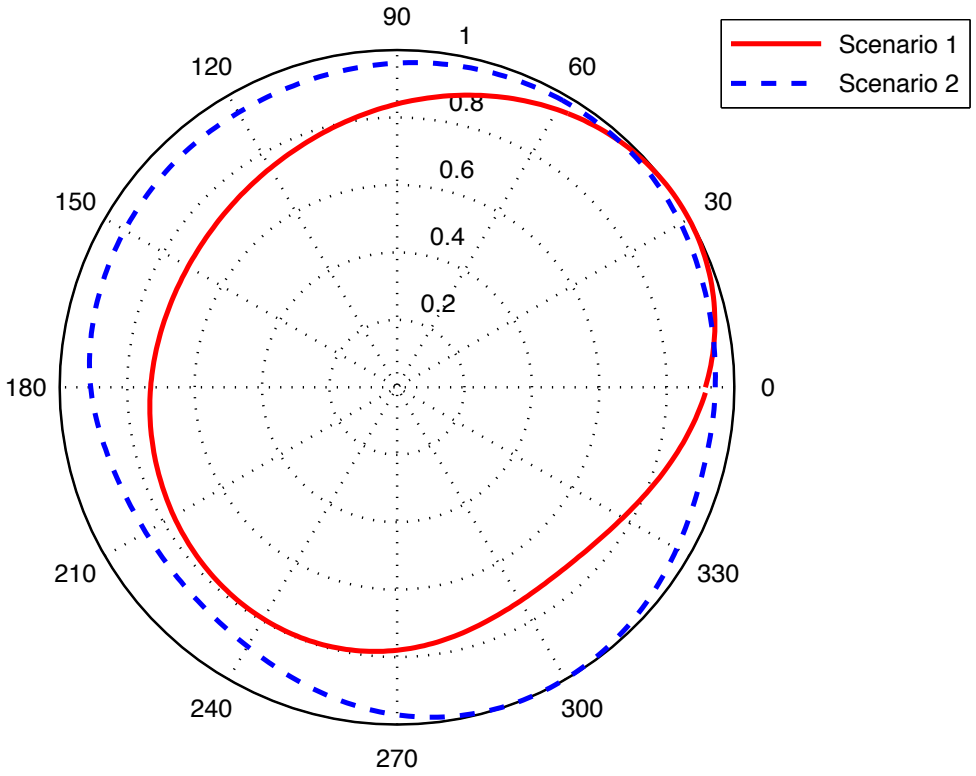

Figure 2.5: Diffuseness assessment using broadband power-pattern; scenario 1: ambient source diffuse field and scenario 2: boosted power diffuse field by adding additional sources.

Based on the diffuseness level quantified in this section and the real data distance estimation results listed in Table 2.2 (explained further in the next Section 2.6.4), we can see that a diffuseness level around 0.7 is a reasonably adequate diffuseness as we can estimate the pairwise distances with less than 5% relative error.

As discussed in Section 2.3.3, estimation of the directional power can be accomplished by a symmetric and uniform microphone array; that implies a carefully designed spherical (3D) or circular (2D) array. The 2D approximation reduces and simplifies some of the computations. We consider this level of approximation reasonable as the obtained calibration error and distance measurement are not very sensitive to the quantified diffuseness [101]. Furthermore, the numerical results confirm that the quantified diffuseness levels are in agreement with the distance estimation results (Table 2.2).

2.6.4 Distance Estimation Performance

In order to estimate the pairwise distances, two microphone signals are processed using a short time Fourier transform of **64**ms frames obtained by applying the Tukey window with parameter = **0.25**. The total length of each microphone signal is **30**s. For each frame, we

Distance (cm)	Baseline	BP	AVG+HIS	AVG+BP+HIS	Corresponding microphone pairs as depicted in Figure 2.3
7.65	.3	.26	.24	.20	{**(1,2),(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,1)**}
10	.37	.35	.31	.26	{**(1,9),(2,9),(3,9),(4,9),(5,9),(6,9),(7,9),(8,9)**}
14.14	.38	.36	.33	.29	{**(1,3),(2,4),(3,5),(4,6),(5,7),(6,8),(7,1),(8,2)**}
18.48	.44	.4	.36	.32	{**(1,4),(2,5),(3,6),(4,7),(5,8),(6,1),(7,2),(8,3)**}
20	.55	.47	.45	.35	{**(1,5),(2,6),(3,7),(4,8)**}
60	8.4	6.3	3.3	2.7	{**(4,10),(2,11),(8,12)**}
70	10.4	9.6	3.5	3.0	{**(9,10),(9,11),(9,12)**}
80	14.1	13.5	3.8	3.2	{**(8,10),(6,11),(4,12)**}
99	25.2	21.3	4.3	3.6	{**(10,11),(11,12)**}

Table 2.2: Root mean squared error of pairwise distance estimation using diffuse field coherence model evaluated on real data recordings. The presented techniques include the baseline formulation [78], enhanced model by averaging coherence function (AV), using histogram (HIS) for removing the outliers as well as boosting the power (BP) of the sound field.

compute the coherence function through (2.10) and estimate the pairwise distance by fitting a sinc function as stated in (2.22) and (2.23). In the baseline approach, each frame is processed independently, which yields **468** point estimates of pairwise distances. To obtain a single estimate of the distance between the two microphones, clustering is applied on the point estimates. Based on k-means clustering, the center of the cluster with the smaller error determines the pairwise distance [78]. Using our enhanced model elaborated in Section 2.3.1, a sinc function is fitted to the averaged coherence function.

We conduct the evaluations using simulated data in a controlled (almost ideal) diffuse field in the medium and large size rooms as described in Section 2.6.1. Figure 2.6 illustrates the results. We can see in Figure 2.6 (top: averaging method) that in the medium size room, the pairwise distances smaller than **1** m can be estimated with less than or equal to **0.02** m error (the **90%** confidence interval is **0.03** m). The estimates become highly erroneous beyond **1** m. By using the conventional method (Figure 2.6 top: k-means clustering), observations show that this method is only applicable when the microphones are located in close proximity to each other (i.e., the pairwise distance less than **30** cm). Figure 2.6 (bottom) illustrates that in the large size room, the averaging method is effective for estimation of pairwise distances up to **3** m.

The relative error for distance estimation $\boldsymbol{d_i}$ can be quantified as

$$\boldsymbol{\epsilon_i} = \sqrt{\frac{\sum_{l=1}^{N}\left(\frac{\hat{d}_{il}-d_i}{d_i}\right)^2}{N}} \tag{2.27}$$

where $\hat{\boldsymbol{d}}_{\boldsymbol{il}}$ is $\boldsymbol{l^{th}}$ estimation of distance $\boldsymbol{d_i}$ and $\boldsymbol{N}$ is the number of microphone pairs with pairwise distance $\boldsymbol{d_i}$. Figure 2.7 shows that measure of $\boldsymbol{\epsilon_i}$ is almost constant for each room and we can fit a linear regression model on the relative error. As depicted in Figure 2.7 (top), the line corresponds to **0.0164** m relative error and the residual error of the linear regression is **0.0028** for the medium size room. In addition, we performed some evaluations in the large room set-up as described in Section 2.6.1. The theories of the sinc function coherence model hold for up to **3** m pairwise distance, which is also verified through our experiments in a

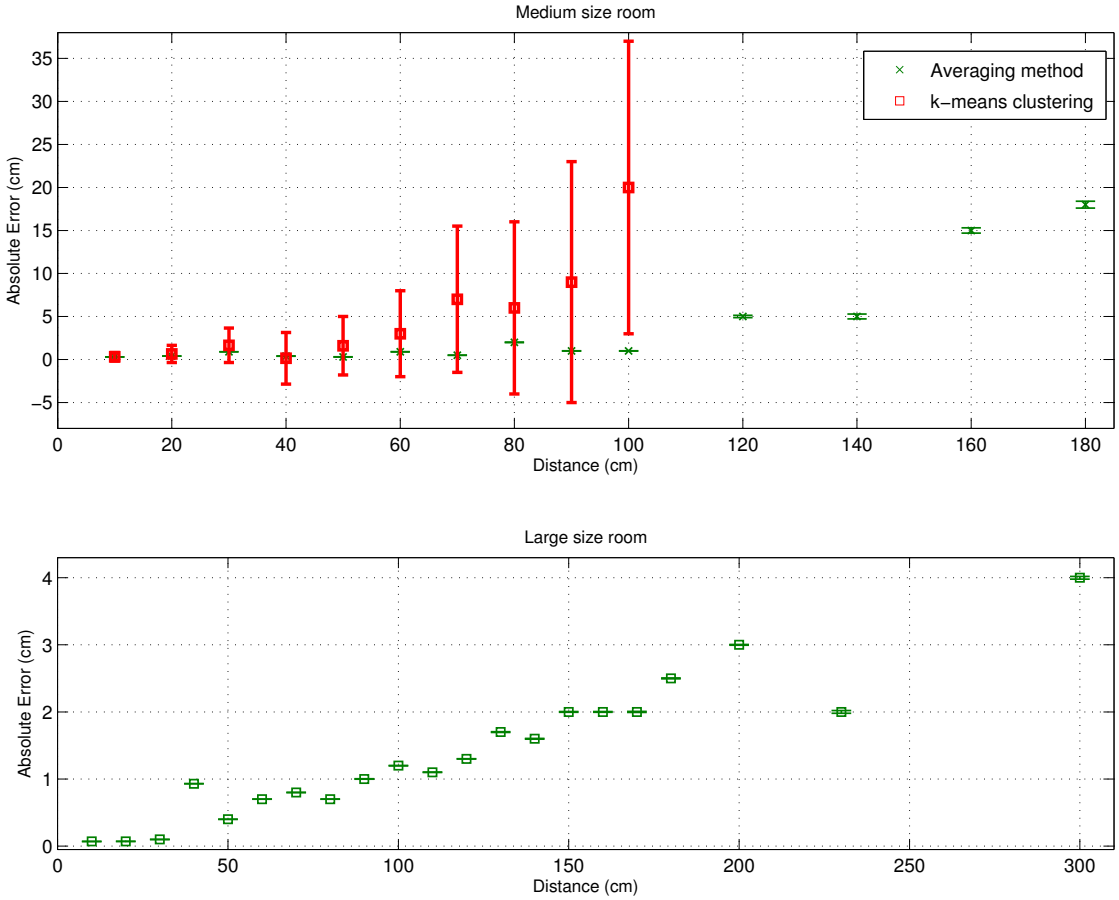

Figure 2.6: Comparison of error bars for estimation of pairwise distances in the medium (top) and large size (bottom) rooms. In the top plot, "cross" corresponds to the averaging method and "square" corresponds to the k-means clustering. The bottom plot corresponds to the averaging method.

diffuse field. Similar to the previous experiment, we can fit a linear regression model on the relative error as depicted in Figure 2.7 (bottom). The line corresponds to **0.0124** m relative error and the residual error of the linear regression is **0.0012**. We can see that the following mathematical model holds for estimation of pairwise distance

$$\hat{\boldsymbol{d}} \sim \mathcal{N}(\boldsymbol{d}, (\boldsymbol{d}\boldsymbol{\epsilon})^2) \tag{2.28}$$

where $\mathcal{N}$ denotes the normal distribution and $\boldsymbol{\epsilon}$ is the mean of the relative errors in distance estimation which is equal to **0.0164** and **0.0124** in the medium and large size rooms respectively. A smaller $\boldsymbol{\epsilon}$ indicates that diffuseness is better realized in the larger room. We further conduct some evaluations using real data recorded in a meeting room. Figure 2.8 illustrates that, for microphones 7 and 8 located **7.6** cm apart, the k-means clustering estimated distance is **8.2** cm. Figure 2.9 (top) demonstrates that for microphones 11 and 5

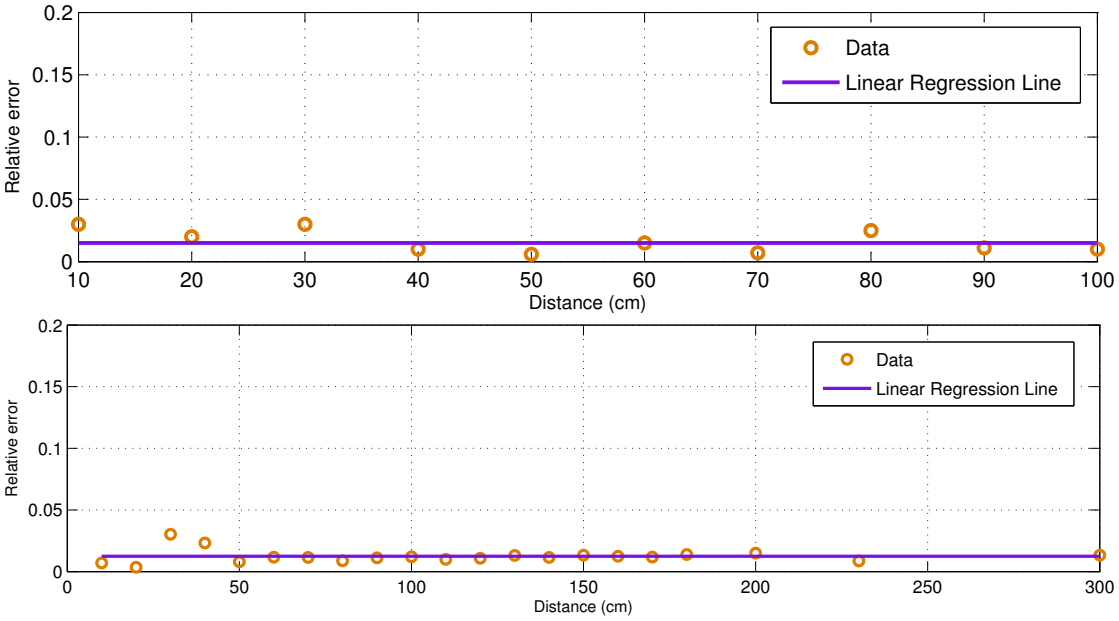

Figure 2.7: Relative error vs. distance for medium size (top) and large (bottom) rooms. The linear regression can be used to predict the relative error.

where the distance is **77.38** cm, it is not possible to provide a reliable estimate by fitting a sinc function on a single frame and k-means clustering. The estimated distance is **66** cm which shows more than **11** cm (**14%**) error.

The proposed averaging method enables more accurate point estimates with fewer outliers. Figure 2.9 (bottom) shows fusion of the k-means method with an averaging technique for estimating the distance between microphones 11 and 5. The averaging is performed on each **5** frames with 80% overlapping. The results shows that the percentage of outliers is reduced so the estimated pairwise distance is **76.6** cm, amounting to **8** mm (**1%**) error.

Although the averaging method reduces the number of outliers, the k-means clustering is not stable and it can generate the wrong winner class. Figure 2.10 shows distance estimation for microphones **11** and **6**. The estimated distance is **90.2** cm, whereas the correct distance is **80**cm. The winner class is wrong using k-means clustering. We propose to remove the outliers using a histogram clustering method, which also offers computational speed advantages over the k-means algorithm. Furthermore, as discussed in Section 2.4.3, it is a more appropriate technique for removing outliers compared to k-means clustering. The two-dimensional histogram clustering is shown in Figure 2.11. Note that the histogram represents the difference of the positions (distance) of the microphones and not the positions themselves. This method is not dependent on the absolute position of the compared microphones. In the histogram method, the bin with the largest number of estimation points is the winner used for the final estimation. The resolution of the bins is a critical parameter for construction of the histogram. We observed empirically that a $\mathbf{50} \times \mathbf{50}$ histogram provides a good estimate; it corresponds

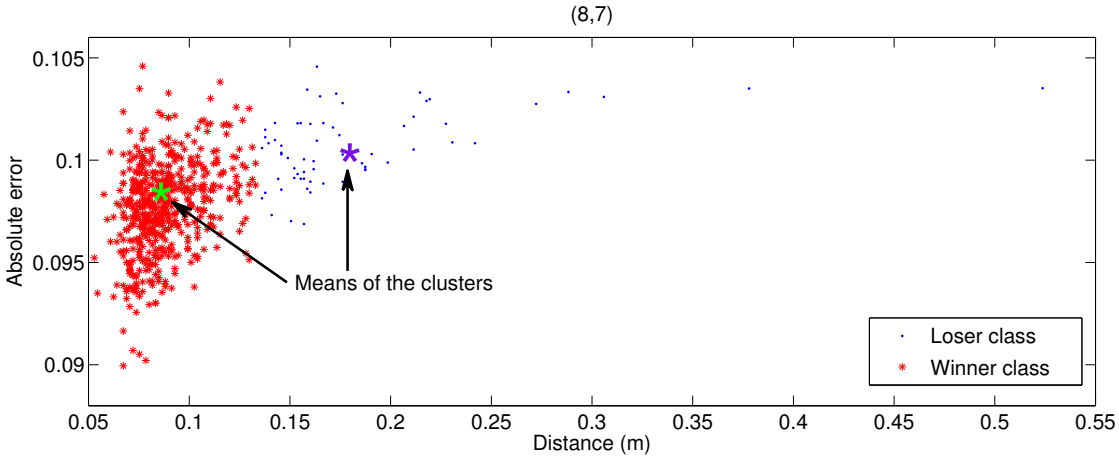

Figure 2.8: Baseline method: k-means clustering for microphones 7 and 8 located **7.6** cm apart. Blue points have high errors and red points are the winners. The estimated pairwise distance is **8.21** cm.

to a resolution of an average **7**mm in pairwise distance estimation. The two-dimensional histogram enables estimation of the pairwise distance as **80.3** cm which has only **3** mm error, equal to **0.4%**. We can see that this method is more accurate than k-means clustering and it is more robust to noisy estimates in real data evaluations.

Table 2.2 summarizes all the results for pairwise distance estimation in the real data scenario. The first column is the ground truth distances. The second column is the root mean square error (RMSE) for the baseline method, and the third column is the RMSE for the boosted power diffuse field; it shows an improvement compared to the baseline. The fourth column corresponds to the results of using the averaging and two-dimensional histogram methods, which shows noticeable improvement. Applying this method on the boosted power diffuse field shows an additional slight improvement as listed in the last column. We can see that the averaging and two-dimensional histogram are more important to achieve robust and accurate results. Furthermore, Figure 2.12 illustrates the measure of improvement using each method. We can see that although boosting the power increases the diffuseness, the improvement in pairwise distance estimation is small because measuring the diffuseness was done on all frequency bands, whereas only the low frequency part has contribution to the distance estimation. Therefore measuring the diffuseness at low frequencies is essential to predict the performance of distance estimation.

2.6.5 Array Calibration Performance

In the final section, we compare all methods for calibration of the geometry of the 9 microphones using real data. Figure 2.13 illustrates the microphone calibration results. The geometry of the array is extracted using the state-of-the-art s-stress [18] method by solving

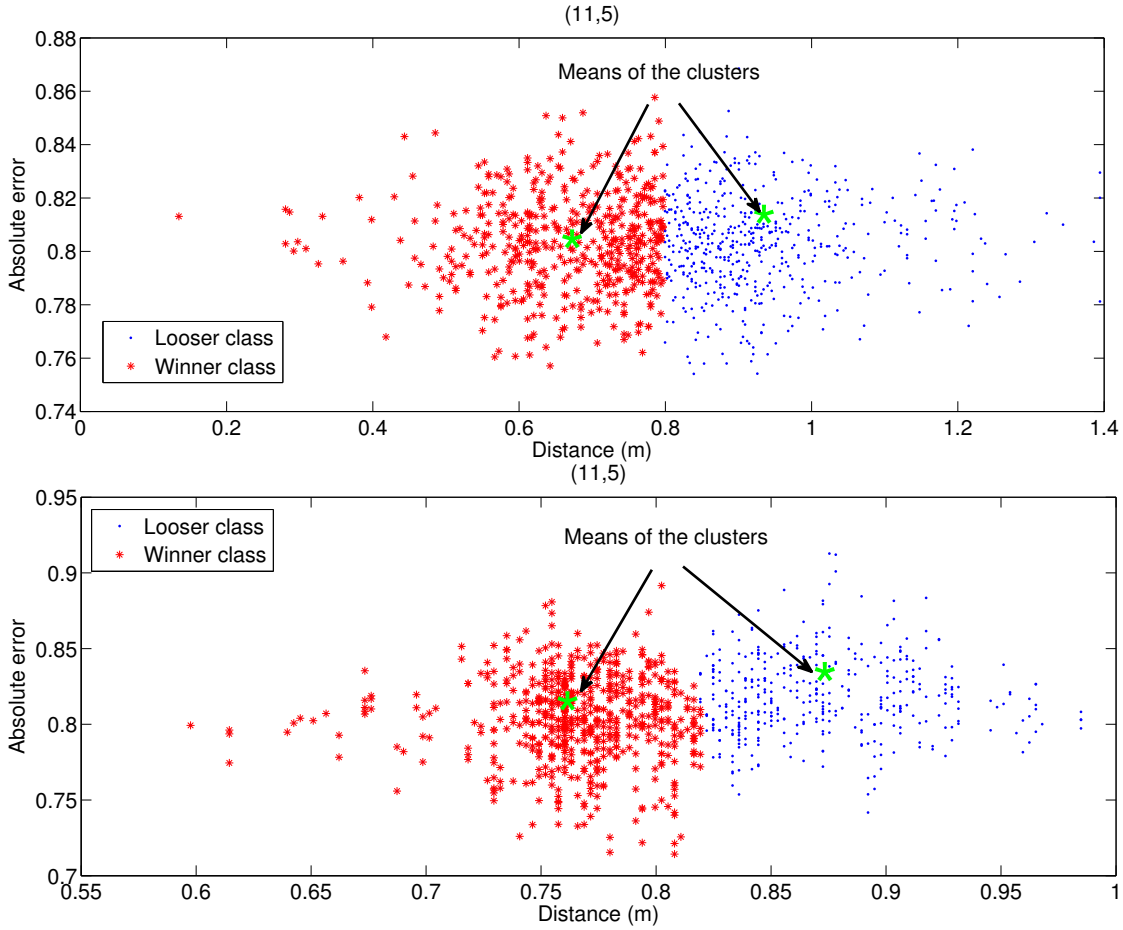

Figure 2.9: Distance estimation of microphones 11 and 5 using real data recordings. The ground truth is **77.38** cm. (top) Baseline method using k-means clustering on single frame coherence function. The estimated distance is **66** cm. (bottom) k-means clustering on averaged coherence function. The estimated distance is **76.6** cm.

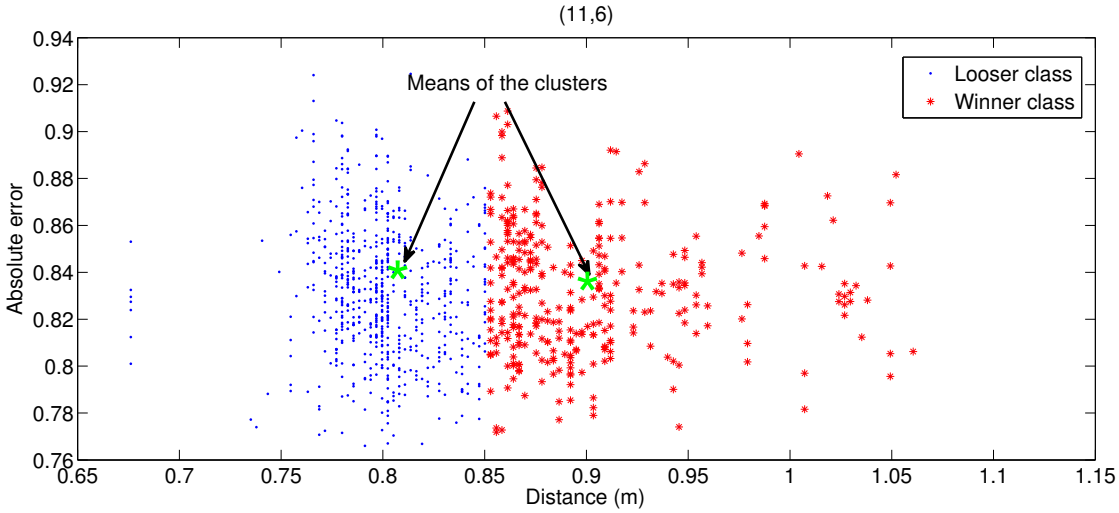

Figure 2.10: Distance estimation of microphones 11 and 6 using averaging and k-means clustering; correct distance is **80** cm and the estimated distance is **90.2** cm.

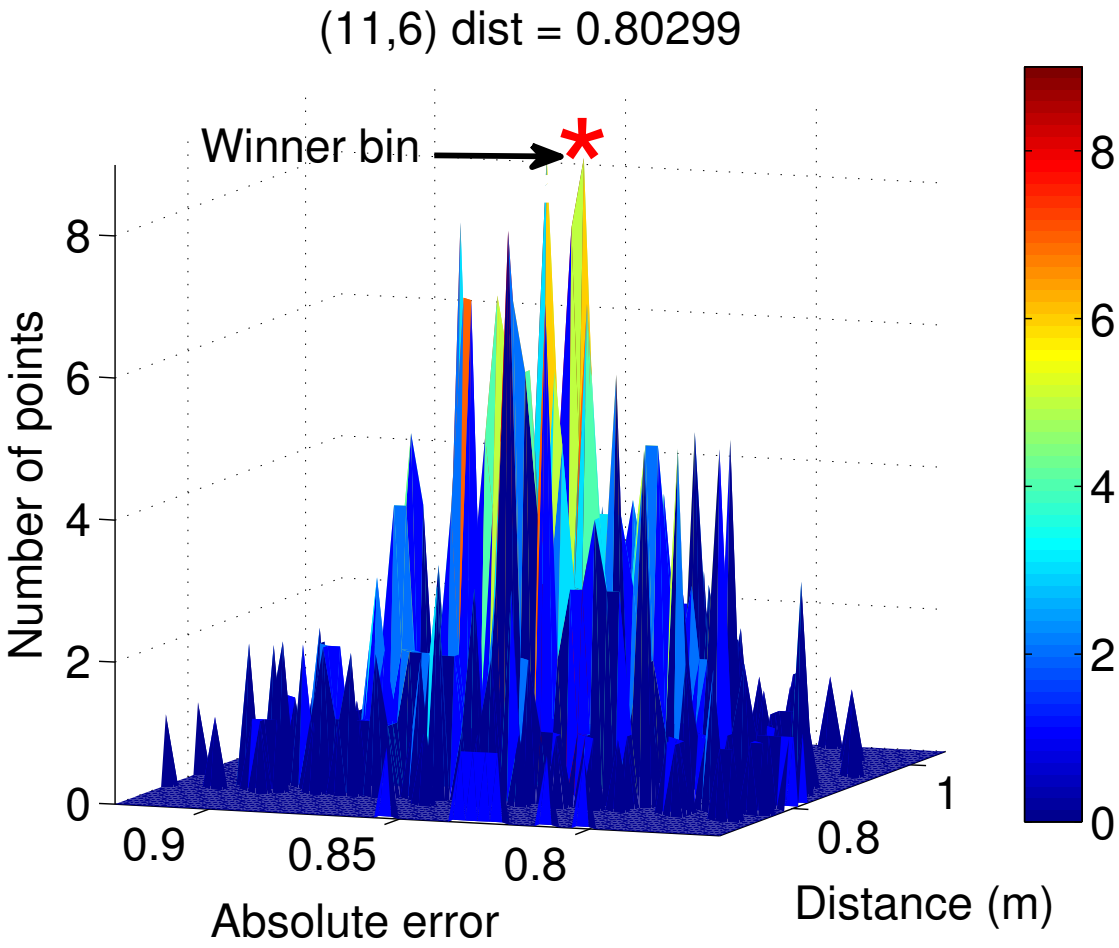

Figure 2.11: Distance estimation using averaging and two-dimensional histogram clustering; the correct distance is **80** cm and the estimated distance is **80.3** cm.

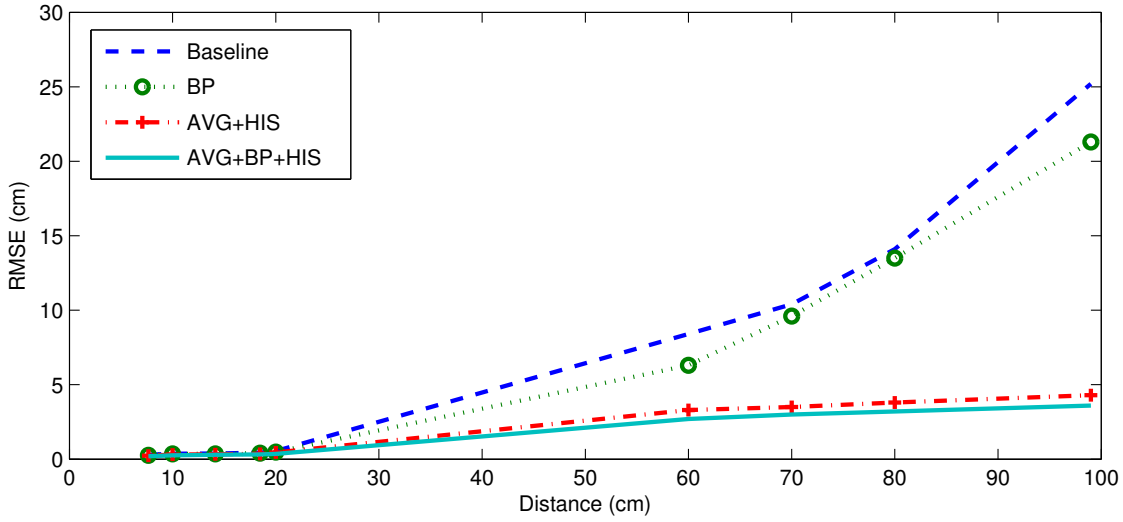

Figure 2.12: Comparing the performance of all methods using real data recordings for pairwise distance estimation. The baseline is k-means method. BP illustrates the results of using extra broadband sound. Furthermore, AVG+HIS shows big improvement by using averaging method and 2D histogram. Finally AVG+HIS+BP shows the result of applying averaging, histogram and augmenting the sound field.

Method	Error
Baseline	8.83
Averaging	8.04
Averaging + Histogram	5.00

Table 2.3: Calibration results of 9 microphones.

the following optimization problem

$$\hat{X} = \underset{X}{\arg\min} \sum_{(i,j)\in E} \left(\|x_i - x_j\|^2 - \tilde{d}_{ij}^2 \right)^2, \tag{2.29}$$

where $E \subseteq [n] \times [n]$ denotes the subset of the estimated pairwise distances and x_i represents the microphone location i. This method is a robust and accurate localization technique where the search space is constrained to the Euclidean geometry. The reconstruction error for the baseline method using the criterion stated in (3.15) is **8.83**. The estimated error based on averaging method is **8.04**.

To further improve the performance, we use the two-dimensional histogram to remove outliers. We can see the improved estimates using the hybrid of averaging method and outlier detection, where the averaging method is applied on five frames to estimate the pairwise distances and to construct the two-dimensional histogram; the estimated error is **5.00**. Table 2.3 summarizes the results. The same number of frames is used by each method.

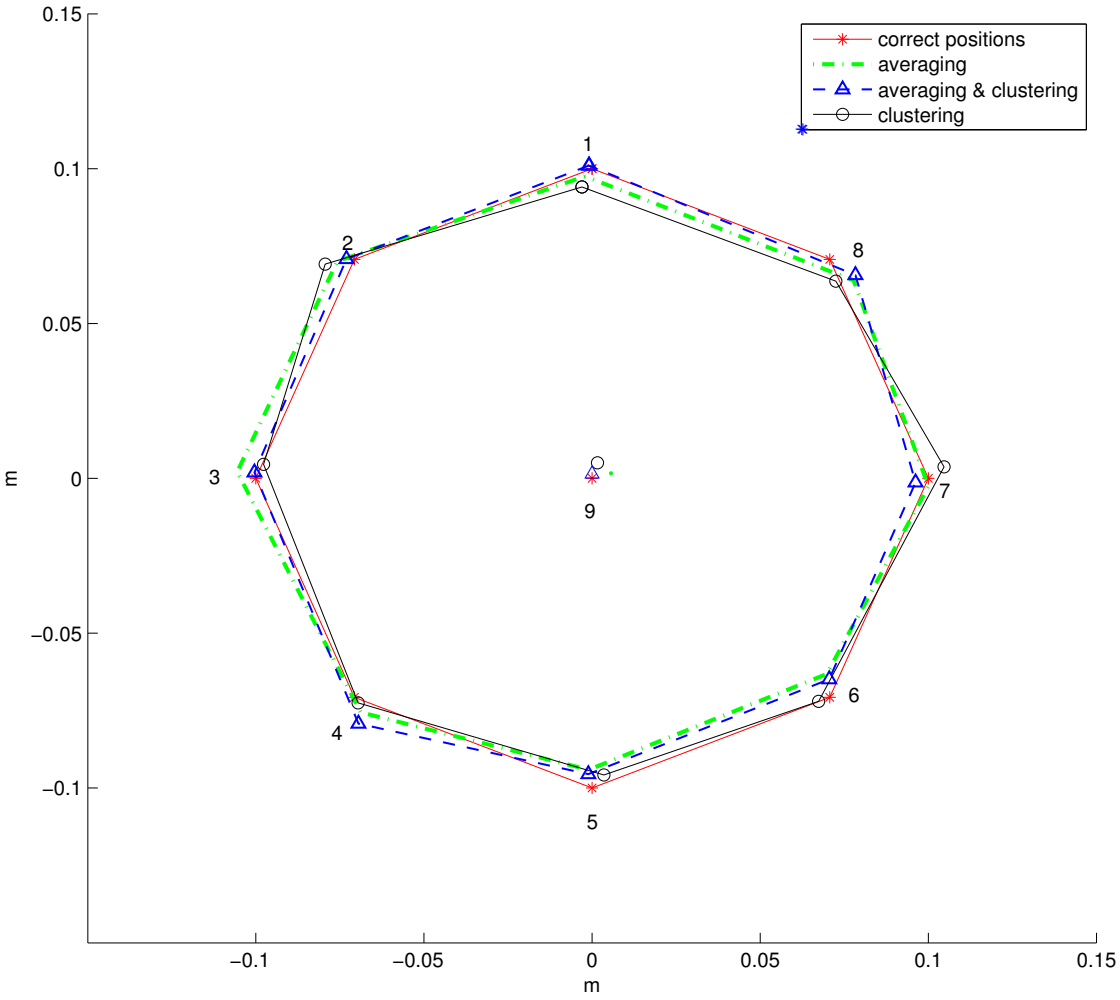

Figure 2.13: Calibration of a 9-channel microphone array on real diffuse sound field recordings using averaging and a hybrid of averaging and histogram-based clustering.

2.6.6 Diffuseness Adequacy for Pairwise Distance Estimation

The theory stated in section 2.2.2 asserts that the critical frequency to create a diffuse field is inversely proportional to the dimension of the room. Hence, as the room gets larger, the critical frequency gets smaller, and we can achieve a higher diffuseness especially at low frequencies. On the other hand, Equation (2.11) shows that by increasing the pairwise distance, the sinc function squeezes in the frequency domain; therefore, the diffuseness at low frequencies becomes highly important for fitting the coherence function and the estimated sinc function. Hence, estimation of large pairwise distances is difficult. Table 2.4 illustrates the relation between room dimension and maximum pairwise distance estimation.

As the simulation results illustrate, increasing the dimensions by a factor of 6 enables estimation of larger pairwise distances by a similar factor of 6. Therefore, in the room with dimensions $\mathbf{48} \times \mathbf{33} \times \mathbf{21}\mathrm{m}^3$, pairwise distances up to **6** m can be estimated accurately.

Room size	ϵ	Max distance (m)
Medium	0.0164	1
Large	0.0124	3
Very large	0.0103	6

Table 2.4: Maximum pairwise distance that can be estimated with relatively low error in three different rooms based on fitting the sinc function to the average coherence function.

Distance (cm)	RMSE
400	2
500	5
600	10
700	12
800	30
900	25
1000	57

Table 2.5: Root mean squared error of pairwise distance estimation using diffuse field coherence model for a very large size room (6 times greater than the medium size room).

Table 2.5 summarizes the results of distance estimation for the very large room. Comparing the simulated and real data evaluations on the medium size room shows that, in a simulated as well as real diffuse field, we can estimate pairwise distance up to **1** m (Table 2.4).

Section 2.5 showed that between the second zero crossing frequency and the Schroeder frequency for the aforementioned three rooms (medium, large and very large), there are only two bands for distances **1**m, **3**m and **6**m respectively (Table 2.1). Hence, the diffuseness generated by a tone is very weak and we may not be able to fit the sinc function to extract these pairwise distances. In our particular case of using broadband signal, all bands that have more than **25** modes generate a diffuse field [85]. Our empirical evaluations show that at least **5** diffuse field bands below the second zero crossing frequency are necessary to achieve reasonable accuracy in distance estimation. Table 2.1 shows that in the medium size room, 5 bands (15–19) generate an adequate diffuse field at frequencies below the second zero crossing; similarly, in the large room and the very large room, the bands 10–14 and 7–11 generate adequate diffuse field distances corresponding to **1**m, **3**m and **6**m respectively. Based on this theory and the second column of Table 2.1, estimation of **3** m distances in the medium size room are impossible. The experiments on real data recordings confirm this theoretical insight. Hence, it becomes straightforward to determine the maximum distance which can be estimated using the diffuse field model. The procedure requires extraction of the modes for the room. The minimum frequency ($\boldsymbol{f}^*$) to have **5** bands generating more than **25** modes lower than $\boldsymbol{f}^*$ yields the maximum resolvable distance as $\boldsymbol{d}^* = \boldsymbol{c}/\boldsymbol{f}^*$. Hence, as $\boldsymbol{f}^*$ gets smaller (i.e. room gets larger) the maximum estimated distance is increased. The $\boldsymbol{f}^*$ is equal to **355**, **112** and **56** Hz for the medium, large, and very large rooms respectively, cf. Table 2.1, **19,14** and **11** band indices. Those correspond to the maximum distances of **0.97** m, **3.06** m and **6.12** m.

2.7 Conclusions

In this chapter, we studied the diffuse field model to enable ad hoc microphone array calibration. The analyses showed the importance of averaging the coherence function prior to fitting the sinc function. The robustness was further improved using 2D histogram based clustering for outlier detection. This observation shows that the errors do not necessarily group into two clusters and it confirms the hypothesis of the effectiveness of 2D histogramming.

The enhanced model was shown to outperform the conventional method significantly. The fundamental limitations of this approach were elaborated and effective strategies were proposed to enable estimation of array geometry in an arbitrary set-up. Based on the theoretical as well as empirical studies on adequacy of diffuseness, a mathematical relationship was characterized to link the room dimensions to the maximum resolvable distance using a diffuse field model. The theory explains why larger aperture arrays can be calibrated in larger enclosures and suggests a simple procedure to figure out the maximum distance that can be estimated using a diffuse field coherence model.

3 Ad Hoc Microphone Array Calibration: Euclidean Distance Matrix Completion Algorithm and Theoretical Guarantees

In the previous chapter, we elaborated on diffuse field model and its use in pairwise distance estimation. In general, pairwise distances might be estimated with a wide range of methods. The common point between many of methods are 1) Increase of error in long pairwise distances 2) Number of estimations are unreliable . This chapter addresses the problem of ad hoc microphone array calibration where only partial information about the distances between microphones is available. Therefore based on the previous chapter or any another methods we have estimations from pairwise distance, but noisy and incomplete. We construct a matrix consisting of the pairwise distances and propose to estimate the missing entries based on a novel Euclidean distance matrix completion algorithm by alternate low-rank matrix completion and projection onto the Euclidean distance space. This approach confines the recovered matrix to the cone of Euclidean distance matrices (EDM) at each iteration of the matrix completion algorithm. The theoretical guarantees of the calibration performance are obtained considering the random and locally structured missing entries as well as the measurement noise on the known distances. This study elucidates the links between the calibration error and the number of microphones along with the noise level and the ratio of missing distances. Thorough experiments on real data recordings and simulated setups are conducted to demonstrate these theoretical insights. A significant improvement is achieved by the proposed Euclidean distance matrix completion algorithm over the state-of-the-art techniques for ad hoc microphone array calibration. The content of this chapter has been published in Elsevier journal of Signal Processing [114].

3.1 Introduction

In the previous chapter, we have seen that the properties of a diffuse acoustic field can be exploited for microphone calibration. The pairwise distances can be estimated by fitting the computed coherence with the sinc function in the least squares sense. Once the pairwise distances are estimated, the s-stress method is used to reconstruct the microphone array geometry [40]. Along similar lines, Hennecke et al. [61] proposed a hierarchical approach where the compact sub-arrays are calibrated using the coherence model of a diffuse sound field. A sound signal is activated and the relative positions of the distributed arrays are determined using steered response power based source localization.

Estimation of the pairwise distances becomes unreliable as the distances between the microphones are increased. Hence, we need to devise some methods to enable microphone calibration when some of the pairwise distances are missing.

The problem of missing data arises when the pairwise distance of only a subset of the sensors can be measured. If a source event is activated, device malfunctioning or architectural barriers (e.g. indoor calibration) may cause the signal of the emitted sounds to reach, or be acquired, by only a subset of the sensors. Furthermore, some of sensors deployed far apart may fail to capture the source energy leading to a locality constraint in distance estimation in ad hoc microphone arrays [106]. In this chapter, as an example use case, the local pairwise distances are measured based on the diffuse sound field coherence model. However, the proposed algorithm and theoretical results are applicable for calibration of a general ad hoc microphone array network. The approach proposed in this chapter imposes no constraint on the geometrical set up.

To address the problem of missing distances, we rely on the characteristics of a Euclidean distance matrix. The matrix consisting of the squared pairwise distances has very low rank (explained in Section 3.3.1). The low-rank property has been investigated in the past years to devise efficient optimization schemes for matrix completion, i.e. recovering a low-rank matrix from randomly known entries. Candès et al. [28] showed that a small random fraction of the entries are sufficient to reconstruct a low-rank matrix *exactly*. Keshavan et al. proposed a matrix completion algorithm known as OPTSPACE and showed its optimality [67]. Furthermore, they proved that their algorithm is robust against noise [68]. Drineas et al. [50] exploited the low rank property to recover the distance matrix. However, they assume a nonzero probability of obtaining accurate distances for any pair of sensors regardless of their distance. This assumption severely restricts the applicability of their result for the microphone array calibration problem.

In the present study, we first estimate the pairwise distances of the microphones in close proximity using the coherence model of the signals of the two microphones in a diffuse noise field using the improved method described in [111]; this approach implies a local connectivity constraint as the pairwise distances of the further microphones can not be estimated. We construct a matrix of all the pairwise distances with missing entries corresponding to the

unknown distances. We exploit the low-rank property of the square of this matrix to enable estimation of all the pairwise distances using matrix completion approach.

The goal of this chapter is to show that exploiting the combination of the rank condition of Euclidean distance matrices (EDMs), similarity in the measured distances, and projection on the EDM cone enables us to estimate the microphone array geometry accurately from only partial measurements of the pairwise distances. To this end, we show that matrix completion is capable of finding the missing entries in our scenario and provide theoretical guarantees to bound the error for ad hoc microphone calibration considering the local connectivity of the noisy known entries. To increase the accuracy, we incorporate the properties of EDMs in the matrix completion algorithm. We show that imposing EDM characteristics on matrix completion improves the robustness and accuracy of extraction of the ad hoc microphone geometry.

The rest of the chapter is organized as follows. In Section 3.2, we explain how pairwise distances of the microphones are estimated using the coherence model of the diffuse noise field as an example use case of the proposed method (this method is elaborated in chapter 2). Section 3.3 describes the mathematical basis and the model used for the calibration problem. The proposed Euclidean distance matrix completion algorithm is elaborated in Section 3.4. Section 3.5 is dedicated to the theoretical guarantees for ad hoc microphone array calibration based on matrix completion. The related methods are investigated in Section 3.6 and the experimental analysis are presented in section 3.7. Finally, the conclusions are drawn in Section 3.8.

3.2 Example Use Case

We consider N microphones located at random positions on a large circular table in a meeting room with homogeneous reverberant acoustics. In the time intervals that there is no active speaker, diffuse noise is the dominant signal in the room. The table is located at the center of the room, hence deviation from diffuseness near the walls can be neglected. Based on the theory of the diffuse noise model, the distance of each two close microphones can be estimated by computing the coherence of their signals $\mathbf{\Gamma}$, and fitting a sinc function with the relation expressed as

$$\Re\left(\mathbf{\Gamma}_{ij}(\omega)\right) = \operatorname{sinc}\left(\frac{\omega d_{ij}}{c}\right), \tag{3.1}$$

where ω is the frequency, operator $\Re(.)$ takes the real part of its argument, d_{ij} is the distance between the two microphones i and j, and c is the speed of sound [38]. Figure 2.1 represents an example of the coherence and the fitted sinc function.

In practice, if the distance between the sensors is large (e.g. greater than **73** cm [111, 113]) we observe deviations from the diffuse characteristics. The maximum distance that can be

computed by this method is assumed to be d_{max}. Therefore, pairwise distances greater than d_{max} are missing implying a locality structure in the missing entries in the distance matrix consisting of the pairwise distances. In addition, the computation algorithm can lead to deviation from the model resulting in unreliable estimates of the short distances causing random missing entries in the distance matrix; the random missing entries intend to model the distances which can not be measured due to mismatch or violations of the underlying pairwise distance estimation model pertained to the acoustic ambiguities. Furthermore, the known entries are noisy due to measurement inaccuracies and variations of diffuseness [113].

3.3 Problem Formulation

3.3.1 Distance Matrix

Consider a distance matrix $D_{N\times N}$ consisting of the distances between N microphones constructed as

$$D = [d_{ij}], \quad d_{ij} = \|x_i - x_j\|, \quad i,j \in \{1,\dots,N\}, \tag{3.2}$$

where d_{ij} is the Euclidean distance between microphones i and j located at x_i and x_j. Therefore, D is a symmetric matrix and it is often full rank.

Let $X_{N\times\zeta}$ denote the position matrix whose i^{th} row, $x_i^T \in \mathbb{R}^\zeta$, is the position of microphone i in ζ-dimensional Euclidean coordinates where microphones are deployed and $\cdot^T$ denotes the transpose operator. By squaring the elements of D, we construct a matrix $M_{N\times N}$ which can be written as

$$M = \mathbf{1}_N \Lambda^T + \Lambda \mathbf{1}_N{}^T - 2XX^T, \tag{3.3}$$

where $\mathbf{1}_N \in \mathbb{R}^N$ is the all ones vector and $\Lambda = (X \circ X)\mathbf{1}_\zeta$; $\circ$ denotes the Hadamard product. We observe that M is the sum of three matrices of rank 1, 1 and at most ζ respectively. Therefore, the rank of the squared distance matrix constructed of the elements $M_{ij} = \left[d_{ij}^2\right]$ is at most $\zeta + 2$ [50]. For instance, if the microphones are located on a plane or shell of a sphere, M has rank 4 and if they are placed on a line or circle, the rank is exactly 3. Hence, there is significant dependency between the elements of M and exploiting this low-rank property is the core of the proposed algorithm in this chapter.

3.3.2 Objective

The noisy estimates of the pairwise distances are modeled as

$$\tilde{d}_{ij} = d_{ij} + w_{ij} \quad ; \quad \widetilde{D} = D + W, \tag{3.4}$$

where w_{ij} is the measurement noise for distance d_{ij} and W is the corresponding measurement noise matrix. We introduce a noise matrix on the squared distance matrix as

$$Z = \widetilde{M} - M = \widetilde{D} \circ \widetilde{D} - D \circ D \,, \tag{3.5}$$

where $\widetilde{M}$ is the noisy squared distance matrix.

As described in Section 3.2, there are two kinds of missing entries. The first group consists of the structured missing entries corresponding to the distances greater than d_{max}. We denote this group by S defined as

$$S = \{(i,j) : d_{ij} \geq d_{max}\} \,, \tag{3.6}$$

These structured missing entries are represented by a matrix

$$D^s_{ij} = \begin{cases} D_{ij} & \text{if } (i,j) \in S \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

Hence, the noiseless recognized pairwise distance matrix is given by

$$D^{\bar{s}} = D - D^s \,, \tag{3.8}$$

and we obtain the corresponding known squared distance matrix as

$$\begin{aligned} M^s &= D^s \circ D^s \\ M^{\bar{s}} &= D^{\bar{s}} \circ D^{\bar{s}} = M - M^s \,. \end{aligned} \tag{3.9}$$

Considering the noise on the known entries, we obtain

$$\widetilde{M}^{\bar{s}} = M^{\bar{s}} + Z^{\bar{s}} \,, \tag{3.10}$$

where $Z^{\bar{s}}$ denotes the noise on the known entries in the squared distance matrix.

To model the random missing entries, we assume that each entry is sampled with probability p; sampling can be introduced by a projection operator on an arbitrary matrix $Q_{N\times N}$, given by

$$\Psi_E(Q)_{ij} = \begin{cases} Q_{ij} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

where $E \subseteq [N] \times [N]$ denotes the known entries after random erasing process and has cardinality $|E| \approx pN^2$. Therefore, the final known squared distance matrix is given by

$$M^E = \Psi_E(\widetilde{M}^{\bar{s}}) \,. \tag{3.12}$$

The goal of the matrix recovery algorithm is to find the missing entries and remove the noise, given matrix $\boldsymbol{M^E}$.

3.3.3 Noise Model

The level of noise in extracting the pairwise distances, $\boldsymbol{w_{ij}}$ in (3.4), increases as the distances become larger [113]. We model this effect through

$$\boldsymbol{W = \Upsilon \circ D}\,, \tag{3.13}$$

where the normalized noise matrix $\boldsymbol{\Upsilon_{N\times N}}$ consists of entries with sub-Gaussian distribution with variance $\boldsymbol{\varsigma^2}$, thus [68]

$$\mathbb{P}(|\boldsymbol{\Upsilon_{ij}}| \geq \boldsymbol{\beta}) \leq \boldsymbol{2}\,\mathrm{e}^{-\frac{\boldsymbol{\beta^2}}{\boldsymbol{2\varsigma^2}}}\,. \tag{3.14}$$

Based on (3.10), $\boldsymbol{Z_{ij}^{\bar{s}} = 2d_{ij}^2\Upsilon_{ij} + d_{ij}^2\Upsilon_{ij}^2}$; thereby $\boldsymbol{Z_{ij}^{\bar{s}}}$ is also a sub-Gaussian random variable with a bounded constant $\boldsymbol{2\varsigma d_{ij}^2}$. The physical setup confines $|\boldsymbol{Z_{ij}^{\bar{s}}}| \leq \boldsymbol{4a^2}$ where $\boldsymbol{a}$ is the radius of the table[1].

3.3.4 Evaluation Measure

Extracting the absolute position of the microphones deployed in $\boldsymbol{\zeta}$ dimensional space requires at least $\boldsymbol{\zeta + 1}$ anchor points in addition to the distance matrix. Therefore, in a scenario where the only available information is pairwise distances, the evaluation measure must quantify the error in estimation of the *relative* position of the microphones thus robust to the rigid transformations (translation, rotation and reflection). Hence, we quantify the distance between the actual locations $\boldsymbol{X}$ and estimated locations $\boldsymbol{\hat{X}}$ as [18]

$$\begin{aligned} \mathrm{dist}(\boldsymbol{X},\boldsymbol{\hat{X}}) &= \frac{\boldsymbol{1}}{\boldsymbol{N}}\left\|\boldsymbol{JXX^TJ - J\hat{X}\hat{X}^TJ}\right\|_{\mathrm{F}}\,, \\ \boldsymbol{J} &= \mathbb{I}_{\boldsymbol{N}} - \boldsymbol{(1/N)1_N1_N^T} \end{aligned} \tag{3.15}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $\mathbb{I}_{\boldsymbol{N}}$ is the $\boldsymbol{N\times N}$ identity matrix. The distance measure stated in (3.15) is useful to compare the performance of different methods in terms of microphone array geometry estimation.

Table 3.1 summarizes the set of important notation.

[1] The sub-Gaussian assumption is exploited for the proof of Theorem 3 stated in Section 3.5. This model is not restrictive in practice and a Gaussian noise is considered for the simulations conducted in Section 4.5.

Table 3.1: Summary of the notation.

Symbol	Meaning	Symbol	Meaning
N	number of microphones	D	complete noiseless distance matrix
a	radius of the circular table on which microphones are distributed	M	squared distance matrix
ς	normalized standard deviation of noise	$\widetilde{M}$	noisy squared distance matrix
Ψ_E	projection into matrices with entries on index set E	$\hat{M}$	estimated squared distance matrix
$\mathscr{P}_e$	projection to EDM cone	Z	noise matrix
p	probability of having random missing entries	M^E	observed matrix
d_{max}	radius of the circle defining structured observed entries	X	positions matrix
M^S	distance matrix with observed entries on index set S	$\hat{X}$	estimated positions matrix

3.4 Euclidean Distance Matrix Completion Algorithm

The approach proposed in this chapter exploits low-rank matrix completion and incorporates the EDM properties for recovering the distance matrix.

3.4.1 Matrix Completion

We recall our problem of having N microphones distributed on a space of dimension ζ. Hence, the squared distance matrix M has rank $\eta = \zeta + 2$, but it is only partially known. The objective is to recover $M_{N\times N}$ of rank $\eta \ll N$ from a sampling of its entries without having to ascertain all the N^2 entries, or collect N^2 measurements about M. The approach proposed through *matrix completion* relies on the fact that a low-rank data matrix carries much less information than its ambient dimension implies. Intuitively, as the matrix M has $(2N-\eta)\eta$ degrees of freedom[2], we need to know at least ηN of the row entries as well as ηN of the column entries reduced by η^2 of the repeated values to recover the entire elements of M.

Given M^E defined in (3.12), the matrix completion recovers an estimate of the distance matrix $\hat{M}$ through the following optimization

$$\begin{aligned} &\text{Minimize} && \operatorname{rank}(\hat{M}) \\ &\text{subject to} && \hat{M}_{ij} = M_{ij}, \qquad (i,j) \in E \end{aligned} \tag{3.16}$$

In this chapter, we use the procedure of OptSpace proposed by Keshavan et al. [68] for estimating a matrix given the desired rank η. This algorithm is implemented in three steps: (1) Trimming, (2) Projection and (3) Minimizing the cost function.

In the trimming step, a row or a column is considered to be over-represented if it contains more samples than twice the average number of non-zero samples per row or column. These rows or columns can dominate the spectral characteristics of the observed matrix M^E. Thus, some of their entries are removed uniformly at random from the observed matrix. Let $\widetilde{M}^E$ be

[2] The degrees of freedom can be estimated by counting the parameters in the singular value decomposition (the number of degrees of freedom associated with the description of the singular values and of the left and right singular vectors). When the rank is small, this is considerably smaller than N^2 [29].

the resulting matrix of this trimming step.

In the projection step, we first compute the singular value decomposition (SVD) of $\widetilde{\boldsymbol{M}}^{\boldsymbol{E}}$ thus

$$\widetilde{\boldsymbol{M}}^{\boldsymbol{E}} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{\sigma}_{\boldsymbol{i}}(\widetilde{\boldsymbol{M}}^{\boldsymbol{E}}) \boldsymbol{U}_{\boldsymbol{.i}} \boldsymbol{V}_{\boldsymbol{.i}}^{\boldsymbol{T}} \,, \tag{3.17}$$

where $\boldsymbol{\sigma}_{\boldsymbol{i}}(\cdot)$ denotes the $\boldsymbol{i}^{\text{th}}$ singular value of the matrix and $\boldsymbol{U}_{\boldsymbol{.i}}$ and $\boldsymbol{V}_{\boldsymbol{.i}}$ designate the $\boldsymbol{i}^{\text{th}}$ column of the corresponding SVD matrices. Then, the rank-$\boldsymbol{\eta}$ projection, $\mathscr{P}_{\boldsymbol{\eta}}(\cdot)$ returns the matrix obtained by setting to $\boldsymbol{0}$ all but the $\boldsymbol{\eta}$ largest singular values as

$$\mathscr{P}_{\boldsymbol{\eta}}(\widetilde{\boldsymbol{M}}^{\boldsymbol{E}}) = (\boldsymbol{N}^{\boldsymbol{2}}/|\boldsymbol{E}|) \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{\eta}} \boldsymbol{\sigma}_{\boldsymbol{i}}(\widetilde{\boldsymbol{M}}^{\boldsymbol{E}}) \boldsymbol{U}_{\boldsymbol{.i}} \boldsymbol{V}_{\boldsymbol{i}}^{\boldsymbol{T}} = \boldsymbol{U}_{\boldsymbol{0}} \boldsymbol{S}_{\boldsymbol{0}} \boldsymbol{V}_{\boldsymbol{0}}^{\boldsymbol{T}} \,. \tag{3.18}$$

Starting from the initial guess provided by the rank-$\boldsymbol{\eta}$ projection $\mathscr{P}_{\boldsymbol{\eta}}(\widetilde{\boldsymbol{M}}^{\boldsymbol{E}})$, $\boldsymbol{U} = \boldsymbol{U}_{\boldsymbol{0}}$, $\boldsymbol{V} = \boldsymbol{V}_{\boldsymbol{0}}$ and $\boldsymbol{S} = \boldsymbol{S}_{\boldsymbol{0}}$, the final step solves a minimization problem stated as follows: Given $\boldsymbol{U} \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{\eta}}, \boldsymbol{V} \in \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{\eta}}$, find

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{U}, \boldsymbol{V}) &= \min_{\boldsymbol{S} \in \mathbb{R}^{\boldsymbol{\eta} \times \boldsymbol{\eta}}} \mathscr{F}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{S}) \,, \\ \mathscr{F}(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{S}) &= \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{(\boldsymbol{i},\boldsymbol{j}) \in \boldsymbol{E}} (\boldsymbol{M}_{\boldsymbol{ij}} - (\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\boldsymbol{T}})_{\boldsymbol{i},\boldsymbol{j}})^{\boldsymbol{2}} \end{aligned} \tag{3.19}$$

$\boldsymbol{F}(\boldsymbol{U}, \boldsymbol{V})$ is determined by minimizing the quadratic function $\mathscr{F}$ over $\boldsymbol{S}$, $\boldsymbol{U}$, $\boldsymbol{V}$ estimated by gradient decent with line search in each iteration. This last step tries to get us as close as possible to the correct low-rank matrix $\boldsymbol{M}$.

3.4.2 Cadzow Projection to the Set of EDM Properties

The classic matrix completion algorithm as described above recovers a low-rank matrix with elements as close as possible to the known entries. However, the recovered matrix does not necessarily correspond to a Euclidean distance matrix; for example, EDMs are symmetric with zero diagonal elements. These properties are not incorporated in the matrix completion algorithm. Hence, we modify the aforementioned procedure to have, as output, matrices that are closer to EDMs [111].

To this end, we apply a Cadzow-like method. The Cadzow algorithm [27] (also known as Papoulis-Gershberg) is a method for finding a signal which satisfies a composite of properties by iteratively projecting the signal into the property sets. We modify the matrix completion algorithm by inserting an extra step at each iteration. In the classic version of this algorithm a simple rank-$\boldsymbol{\eta}$ approximation is used as the starting point for the iterations using gradient descent on (3.19). After each iteration of the gradient descent, we apply the transformation $\mathscr{P}_{\boldsymbol{c}} : \mathbb{R}^{\boldsymbol{N} \times \boldsymbol{N}} \longmapsto \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}}$ on the obtained matrix where $\mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}}$ is the space of symmetric, positive hollow

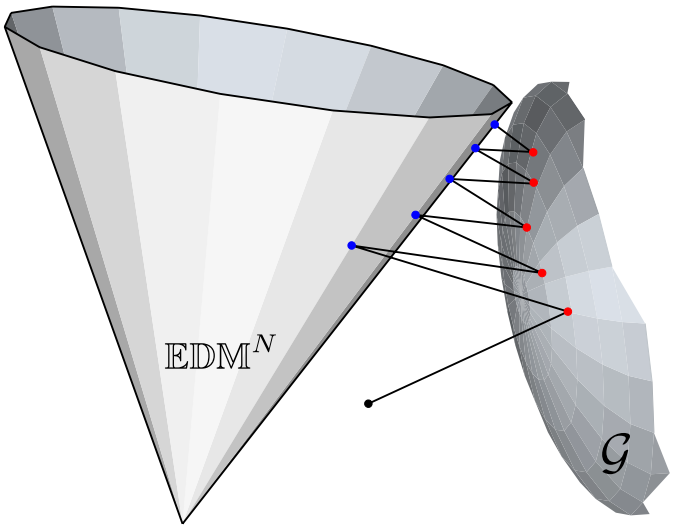

Figure 3.1: Matrix completion with projection onto the EDM cone.

matrices, to make sure that the output satisfies the following properties

$$\hat{\boldsymbol{M}} \in \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}} \Longleftrightarrow \begin{cases} \boldsymbol{d_{ij}} = \boldsymbol{0} \Leftrightarrow \boldsymbol{x_i} = \boldsymbol{x_j} \\ \boldsymbol{d_{ij}} > \boldsymbol{0},\ \boldsymbol{i} \neq \boldsymbol{j} \\ \boldsymbol{d_{ij}} = \boldsymbol{d_{ji}} \end{cases} \tag{3.20}$$

for $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{N}]$; nonnegativity and symmetry are achieved by setting all the negative elements to zero and averaging the symmetric elements.

3.4.3 Matrix Completion with Projection onto the EDM cone

In section 3.4.2, three characteristics of EDMs are employed through the Cadzow projection to reduce the reconstruction error of the distance matrix. In order to increase the accuracy even further, we propose to project to the cone of Euclidean distance matrix, $\mathbb{EDM}^{\boldsymbol{N}}$, at each iteration of the algorithm. In other words, after one step of the gradient descent method on the Cartesian product of two Grassmannian manifolds $\mathcal{G}$, we apply a projection, $\mathscr{P}_{\boldsymbol{e}} : \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}} \longmapsto \mathbb{EDM}^{\boldsymbol{N}}$ to decrease the distance between the estimated matrix and the EDM cone. This is visualized in Figure 3.1. Note that the illustration of the cone and the manifold are not mathematically accurate and only serve as visualizations (The dimension of the cone and the manifold are too large to be illustrated graphically). The projected matrix must satisfy the following EDM

properties [42]

$$\hat{\boldsymbol{M}} \in \mathbb{EDM}^{\boldsymbol{N}} \Longleftrightarrow \begin{cases} -\boldsymbol{z}^{\boldsymbol{T}} \hat{\boldsymbol{M}} \boldsymbol{z} \geq \boldsymbol{0} \\ \boldsymbol{1}^{\boldsymbol{T}} \boldsymbol{z} = \boldsymbol{0} \\ (\forall \|\boldsymbol{z}\| = \boldsymbol{1}) \\ \hat{\boldsymbol{M}} \in \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}} \end{cases} \tag{3.21}$$

The EDM properties include the triangle inequality, thus

$$\boldsymbol{d_{ij}} \leq \boldsymbol{d_{ik}} + \boldsymbol{d_{kj}}, \quad \boldsymbol{i} \neq \boldsymbol{j} \neq \boldsymbol{k}\,, \tag{3.22}$$

as well as the relative-angle inequality; $\forall \boldsymbol{i,j,l} \neq \boldsymbol{k} \in [\boldsymbol{N}], \boldsymbol{i} < \boldsymbol{j} < \boldsymbol{l}$, and for $\boldsymbol{N} \geq \boldsymbol{4}$ distinct points $\{\boldsymbol{x_k}\}$, the inequalities

$$\begin{aligned} \cos(\boldsymbol{\tau_{jkl}} + \boldsymbol{\tau_{lkj}}) \leq \cos \boldsymbol{\tau_{ikj}} \leq \cos(\boldsymbol{\tau_{ikl}} - \boldsymbol{\tau_{lkj}}) \\ \boldsymbol{0} \leq \boldsymbol{\tau_{ikl}}, \boldsymbol{\tau_{lkj}}, \boldsymbol{\tau_{ikj}} \leq \boldsymbol{\pi} \end{aligned} \tag{3.23}$$

where $\boldsymbol{\tau_{ikj}}$ denotes the angle between vectors at $\boldsymbol{x_k}$ and it is satisfied at each position $\boldsymbol{x_k}$.

The projection $\mathscr{P}_{\boldsymbol{e}}$ must map the output of matrix completion to the closest matrix on $\mathbb{EDM}^{\boldsymbol{N}}$ with the properties listed in (4.8). The projection onto $\mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}}$ is achieved by $\mathscr{P}_{\boldsymbol{c}}$ implemented via Cadzow; thereby, we define $(\boldsymbol{U_c}, \boldsymbol{V_c}, \boldsymbol{S_c}) = \mathscr{P}_{\boldsymbol{c}}(\boldsymbol{U^{k+1/2}}, \boldsymbol{V^{k+1/2}}, \boldsymbol{S^{k+1/2}})$. To achieve the full EDM properties, we search in the EDM cone using a cost function defined as

$$\mathscr{H}(\boldsymbol{X}) = \left\| \boldsymbol{1_N \Lambda^T} + \boldsymbol{\Lambda 1_N^T} - \boldsymbol{2XX^T} - \boldsymbol{U_c S_c V_c^T} \right\|_{\mathrm{F}}^{\boldsymbol{2}}. \tag{3.24}$$

To minimize the cost function, we start from the vertex of the $\mathbb{EDM}^{\boldsymbol{N}}$ thus assume that all microphones are located in the origin of the space $\mathbb{R}^{\boldsymbol{\zeta}}$. Denoting the location of microphone $\boldsymbol{i}$ with $\boldsymbol{x_i} = [\boldsymbol{x_{i1}}, ..., \boldsymbol{x_{i\zeta}}]^{\boldsymbol{T}}$, $\mathscr{H}(\boldsymbol{X})$ is a polynomial function of $\boldsymbol{x_{i1}}$ of degree $\boldsymbol{4}$. The minimum of $\mathscr{H}(\boldsymbol{X})$ with respect to $\boldsymbol{x_{i1}}$ can be computed by equating the partial derivation of equation (4.13) to zero to obtain the new estimates, thus

$$\begin{aligned} \hat{\boldsymbol{X}} &= \underset{\boldsymbol{X}}{\mathbf{argmin}}\, \mathscr{H}(\boldsymbol{X}) \\ (\boldsymbol{U^{k+1}}, \boldsymbol{V^{k+1}}, \boldsymbol{S^{k+1}}) &= \mathrm{SVD}\,(\boldsymbol{1_N \hat{\Lambda}^T} + \boldsymbol{\hat{\Lambda} 1_N^T} - \boldsymbol{2\hat{X}\hat{X}^T}) \end{aligned} \tag{3.25}$$

where $\hat{\boldsymbol{\Lambda}} = (\hat{\boldsymbol{X}} \circ \hat{\boldsymbol{X}})\boldsymbol{1_\zeta}$. The stopping criterion is satisfied when the new estimates differ from the old ones by less than a threshold.

The modified iterations can be summarized in two steps:

- iteration $k+1/2$:

$$\begin{aligned} U^{k+1/2} &= U^k + \vartheta \frac{\partial F(U^k, V^k)}{\partial U} \\ V^{k+1/2} &= V^k + \vartheta \frac{\partial F(U^k, V^k)}{\partial V} \\ S^{k+1/2} &= \arg\min_S \mathscr{F}(U^k, V^k, S) \end{aligned} \tag{3.26}$$

- iteration $k+1$:

$$(U^{k+1}, V^{k+1}, S^{k+1}) = \mathscr{P}_e(U^{k+1/2}, V^{k+1/2}, S^{k+1/2}) \tag{3.27}$$

where ϑ is the step-size found using line search.

Once the distance matrix is recovered by either classic or Cadzow matrix completion algorithms, MDS is used to find the coordinates of the microphones, $\hat{X}$, whereas the proposed Euclidean distance matrix completion algorithm directly yields the coordinates.

3.5 Theoretical Guarantees for Microphone Calibration

In this section, we derive the error bounds on the reconstruction of the positions of N microphones distributed randomly on a circular table of radius a using the matrix completion algorithm and considering the locality constraint on the known entries, i.e. $d_{ij} \leq d_{max}$, as well as the noise model with the standard deviation ςd_{ij} as stated in (3.14). Based on the following theorem we guarantee that there is an upper bound on the calibration error which decreases by the number of microphones.

Theorem 1. There exist constants C_1 and C_2, such that the output $\hat{X}$ satisfies

$$\text{dist}(X, \hat{X}) \leq C_1 \frac{a^2 \log_2 N}{pN} + C_2 \varsigma \frac{d_{max}^2}{\sqrt{pN}} \tag{3.28}$$

with probability greater than $1 - N^{-3}$, provided that the right-hand side is less than $\sigma_\eta(M)/N$.

3.5.1 Proof of Theorem 1

The squared distance matrix $M \in \mathbb{R}^{N \times N}$ with rank$-\eta$, singular values $\sigma_k(M), k \in [\eta]$ and singular value decomposition $U\Sigma U^T$ is (μ_1, μ_2)-*incoherent* if the following conditions hold.

$\mathscr{A}_1$. For all $i \in [N]$: $\sum_{k=1}^{\eta} U_{ik}^2 \leq \eta \mu_1$.

$\mathscr{A}_2$. For all $i, j \in [N]$: $\left| \sum_{k=1}^{\eta} U_{ik} (\sigma_k(M)/\sigma_1(M)) U_{jk} \right| \leq \sqrt{\eta} \mu_2$.

where without loss of generality, $U^T U = NI$.

For a (μ_1, μ_2)-incoherent matrix M, (3.29) is correct with probability greater than $1 - N^{-3}$; cf. [68]-Theorem 1.2.

$$\frac{1}{N}\|M - \hat{M}\|_{\mathrm{F}} \leq \frac{C_1' \|\Psi_E(M^s)\|_2 + C_2' \left\|\Psi_E(Z^{\bar{s}})\right\|_2}{p\,N}\,, \tag{3.29}$$

provided that

$$|E| \geq C_1' N \kappa_\eta^2(M) \max\left\{\mu_1 \eta \log N; \mu_1^2 \eta^2 \kappa_\eta^4(M); \mu_2^2 \eta^2 \kappa_\eta^4(M)\right\}\,, \tag{3.30}$$

and

$$\frac{C_1' \|\Psi_E(M^s)\|_2 + C_2' \left\|\Psi_E(Z^{\bar{s}})\right\|_2}{p\,N} \leq \sigma_\eta(M)/N\,, \tag{3.31}$$

where the condition number $\kappa_\eta(M) = \sigma_1(M)/\sigma_\eta(M)$.

To prove Theorem 1, in the first step, we show the correctness of the upper bound stated in (3.28) based on the following Theorems 2 and 3. In the second step, conditions (3.30) and (3.31) are shown to hold along with the (μ_1, μ_2)-incoherence property.

Theorem 2. There exists a constant C_1'', such that with probability greater than $1 - N^{-3}$,

$$\|\Psi_E(M^s)\|_2 \leq C_1''\, a^2 \log_2 N\,. \tag{3.32}$$

The proof of this theorem is explained in Appendix 1.

Theorem 3. There exists a constant C_2'', such that with probability greater than $1 - N^{-3}$,

$$\left\|\Psi_E(Z^{\bar{s}})\right\| \leq C_2'' d_{\max}^2 \varsigma \sqrt{pN}\,. \tag{3.33}$$

The proof of this theorem is explained in Appendix 2.

On the other hand, the following condition holds for any arbitrary network of microphones [88]

$$\mathrm{dist}(X, \hat{X}) \leq \frac{1}{N}||M - \hat{M}||_{\mathrm{F}}\,. \tag{3.34}$$

Therefore, based on Theorem 2, Theorem 3 and the relations (3.29) and (3.34), the upper bound stated in (3.28) is correct where $C_1 = C_1' C_1''$ and $C_2 = C_2' C_2''$; it is enough to investigate conditions (3.30) and (3.31) and (μ_1, μ_2)-incoherency of M to prove Theorem 1.

To show the inequality stated in (3.30), we can equivalently show that

$$Np \geq C_1' \mu^2 \eta^2 \kappa_\eta^6(M) \log N\,, \tag{3.35}$$

where $\mu = \max(\mu_1, \mu_2)$. In order to show that (3.35) holds with high probability for $N \geq \mathscr{C} \log N/p$ and some constant $\mathscr{C}$, we show that $\kappa_\eta(M)$ and μ are bounded with high probability

independent of N.

The squared distance between x_i and $x_j \in \mathbb{R}^\zeta$ is given by

$$M_{ij} = \rho_i^2 + \rho_j^2 - 2x_i^T x_j \,, \tag{3.36}$$

where ρ_i is the distance of microphone i from the center of the table. The squared distance matrix can be expressed as

$$M = A\mathcal{S}A^T \,, \tag{3.37}$$

where for a planar deployment of microphones, i.e., $\zeta = 2,\ \eta = 4$, and $x_i^T = [x_i, y_i] \in \mathbb{R}^2$, we have

$$A = \begin{bmatrix} a/2 & x_1 & y_1 & -a^2/4 + \rho_1^2 \\ \vdots & \vdots & \vdots & \vdots \\ a/2 & x_N & y_N & -a^2/4 + \rho_N^2 \end{bmatrix} \,, \tag{3.38}$$

and

$$\mathcal{S} = \begin{bmatrix} 2 & 0 & 0 & 2/a \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 2/a & 0 & 0 & 0 \end{bmatrix} . \tag{3.39}$$

Since $\mathcal{S}$ is nondefective, using eigendecomposition, there is a non-singular matrix $\mathcal{W}$ and diagonal matrix $\mathbf{\Gamma}$ such that

$$\mathcal{S} = \mathcal{W}\mathbf{\Gamma}\mathcal{W}^{-1} \,, \tag{3.40}$$

where

$$\mathbf{\Gamma} = \text{diag}\left(-2, -2, \frac{a + \sqrt{4 + a^2}}{a}, \frac{a - \sqrt{4 + a^2}}{a}\right) . \tag{3.41}$$

The largest and smallest singular values of $\mathcal{S}$ are $\sigma_1(\mathcal{S}) = \frac{a+\sqrt{4+a^2}}{a}$ and $\sigma_4(\mathcal{S}) = \min\left(2, \frac{\sqrt{4+a^2}-a}{a}\right)$ respectively. Based on (3.37), we have

$$\sigma_1(M) \leq \sigma_1(\mathcal{S})\,\sigma_1(AA^T) \,, \tag{3.42}$$

$$\sigma_4(M) \geq \sigma_4(\mathcal{S})\,\sigma_4(AA^T) \,. \tag{3.43}$$

Therefore, to bound $\kappa_4(M) = \sigma_1(M)/\sigma_4(M)$, we need to derive the bound for $\sigma_1(AA^T)$ and $\sigma_4(AA^T)$. Assuming a uniform distribution of the microphones on the circular table, we have the following distribution for ρ

$$P_\rho(\rho) = \frac{2\rho}{a^2} \qquad \text{for} \quad 0 \le \rho \le a\,. \tag{3.44}$$

Therefore, the expectation of the matrix $A^T A$ is

$$\mathbb{E}[A^T A] = \begin{bmatrix} Na^2/4 & 0 & 0 & Na^3/8 \\ 0 & Na^4/4 & 0 & 0 \\ 0 & 0 & Na^4/4 & 0 \\ Na^3/8 & 0 & 0 & 7Na^4/48 \end{bmatrix}. \tag{3.45}$$

Hence, the largest and smallest singular values of $\mathbb{E}[A^T A]$ are $N\sigma_{\max}(a)$ and $N\sigma_{\min}(a)$ respectively with $\sigma_{\max}(a)$ and $\sigma_{\min}(a)$ independent of N. Moreover, $\sigma_i(\cdot)$ is a Lipschitz continuous function of its arguments and based on the Chernoff bound [119], we get

$$\mathbb{P}(\sigma_1(AA^T) > 2N\sigma_{\max}(a)) \le e^{-\mathscr{C}'N}\,, \tag{3.46}$$

$$\mathbb{P}(\sigma_1(AA^T) < (1/2)N\sigma_{\max}(a)) \le e^{-\mathscr{C}'N}\,, \tag{3.47}$$

$$\mathbb{P}(\sigma_4(AA^T) < (1/2)N\sigma_{\min}(a)) \le e^{-\mathscr{C}'N}\,, \tag{3.48}$$

for a constant $\mathscr{C}'$. Hence, with high probability, based on relations (3.42), (3.43), (3.46) and (3.48), we have

$$\kappa_4(M) \le \frac{4\sigma_{\max}(a)\,\sigma_1(\mathscr{S})}{\sigma_{\min}(a)\,\sigma_4(\mathscr{S})} = f_{\kappa_4}(a)\,. \tag{3.49}$$

This bound is independent of N.

In the next step, we have to bound μ_1 and μ_2. The rank of matrix A is η, therefore there are matrices $B \in \mathbb{R}^{\eta\times\eta}$ and $V \in \mathbb{R}^{N\times\eta}$ such that $A = VB^T$ and $V^T V = N\mathbb{I}$. Given $M = U\Sigma U^T$ and (3.37), we have $\Sigma = Q^T B^T \mathscr{S} BQ$ and $U = VQ$ for an orthogonal matrix Q. To show the incoherence property $\mathscr{A}_1$, we show that

$$\|V_{i.}\|^2 \le \eta\,\mu_1 \qquad \forall\, i \in [N]\,, \tag{3.50}$$

where $V_{i.}$ denotes the transpose of i^{th} row of the corresponding matrix. For $\eta = 4$, since $V_{i.} = B^{-1}A_{i.}$, we have $\|V_{i.}\|^2 \le \sigma_4(B)^{-2}\|A_{i.}\|^2$ and $\sigma_4(A) = \sqrt{N}\sigma_4(B)$, therefore

$$\|V_{i.}\|^2 \le \sigma_4(A)^{-2}\|A_{i.}\|^2 N\,. \tag{3.51}$$

Moreover, $\|A_{i.}\|^2 = a^2/4 + \rho_i^2 + (-a^2/4 + \rho_i^2)^2 \leq 5a^2/4 + 9a^4/16$. Defining

$$f_{\mu_1}(a) = \frac{5a^2/2 + 9a^4/8}{\sigma_{\min}(a)}, \tag{3.52}$$

and based on (3.48) and (3.51), with high probability we have

$$\|U_{i.}\|^2 \leq f_{\mu_1}(a) \qquad \forall\, i \in [N]. \tag{3.53}$$

Therefore, the incoherence property $\mathscr{A}_1$ for $\mu_1 = f_{\mu_1}(a)/\eta$ is correct; that is independent of N.

To prove the incoherence property $\mathscr{A}_2$, it is enough to prove that $\left|M_{ij}/\sigma_1(M)\right| \leq \sqrt{\eta}\,\mu_2/N$ for all $i, j \in [N]$. The maximum value of M_{ij} is $4a^2$ and based on (3.43) and (3.48) we have

$$\sigma_1(M) \geq \sigma_4(M) \geq \frac{1}{2} N \sigma_{\min}(a)\,\sigma_4(\mathscr{S}), \tag{3.54}$$

Defining $f_{\mu_2}(a) = 8a^2/\sigma_{\min}(a)\,\sigma_4(\mathscr{S})$, we have

$$\left|M_{ij}/\sigma_1(M)\right| \leq \frac{f_{\mu_2}(a)}{N} \qquad \forall\, i, j \in [N]. \tag{3.55}$$

Therefore, the incoherence property $\mathscr{A}_2$ for $\mu_2 = f_{\mu_2}(a)/\sqrt{\eta}$ is correct; that is independent of N. Since $\kappa_4(M)$, μ_1 and μ_2 are bounded independent of N, matrix M is (μ_1, μ_2)-incoherent and the inequalities (3.30) and (3.35) are correct.

Further, (3.31) holds with high probability, if the right-hand side of (3.28) is less than $C_3\,\sigma_{\min}(a)\,\sigma_4(\mathscr{S})$, since based on (3.48), $\frac{\sigma_\eta(M)}{N} \geq \frac{1}{2}\sigma_{\min}(a)\sigma_4(\mathscr{S})$. This finishes the proof of Theorem 1.

■

The theoretical analysis elaborated in this section, elucidates a link between the performance of microphone array calibration and the number of microphones, noise level and the ratio of missing pairwise distances. In Section 4.5, thorough evaluations are conducted that demonstrate these theoretical insights. Furthermore, The theoretical error bounds of ad hoc microphone calibration established above corresponds to the classic matrix completion algorithm. We will extend the mathematical results to the completion of Euclidean distance matrices incorporating the Cadzow and EDM projections through the experiments. As we will see in Section 4.5, this bound is not tight for the Cadzow projection and the Euclidean distance matrix completion algorithm as we achieve better results than matrix completion for microphone array calibration.

3.6 Related Methods

The objective is to extract the relative (up to a rigid transformation) microphone positions $\boldsymbol{x_i}$, $\boldsymbol{i} \in \{\mathbf{1},\dots,\boldsymbol{N}\}$ from the measurements of pairwise distances. Some of the state-of-the-art methods to achieve this goal are (1) Multi-Dimensional Scaling (MDS) [107], (2) Semi-Definite Programming (SDP) [14] and S-Stress [18] discussed briefly in the following sections.

3.6.1 Classic Multi-Dimensional Scaling Algorithm

MDS refers to a set of statistical techniques used in finding the configuration of objects in a low dimensional space such that the measured pairwise distances are preserved [40]. Given a distance matrix, finding the relative microphone positions is achieved by MDSLocalize [18]. In the ideal case where matrix $\boldsymbol{M}$ is complete and noiseless, this algorithm outputs the relative positions of the microphones. At the first step, a double centering transformation is applied to $\boldsymbol{M}$ to subtract the row and column means of the distance matrix via $\boldsymbol{\Xi(M)} = \frac{-\mathbf{1}}{\mathbf{2}}\boldsymbol{JMJ}$ where $\boldsymbol{J} = \boldsymbol{I_N} - \mathbf{1}/\boldsymbol{N}\mathbf{1}_{\boldsymbol{N}}\mathbf{1}_{\boldsymbol{N}}^{\boldsymbol{T}}$. The $\boldsymbol{\zeta}$ largest eigenvalues and the corresponding eigenvectors of $\boldsymbol{\Xi(M)}$ denoted by $\boldsymbol{\Pi}_+$ and $\boldsymbol{U}_+$ are calculated and the microphone positions are obtained as $\boldsymbol{X} = \boldsymbol{U}_+\sqrt{\boldsymbol{\Pi}_+}$.

In a real scenario of missing distances, a modification called MDS-MAP [107] computes the shortest paths between all pairs of nodes in the region of consideration. The shortest path between microphones $\boldsymbol{i}$ and $\boldsymbol{j}$ is defined as the path between two nodes such that the sum of the estimated distance measures of its constituent edges is minimized. By approximating the missing distances with the shortest path and constructing the distance matrix, classical MDS is applied to estimate the microphone array geometry.

3.6.2 Semidefinite Programming

Another efficient method that can be used for calibration is the semidefinite programming approach formulated as

$$\hat{\boldsymbol{X}} = \underset{\boldsymbol{X}}{\mathbf{arg\,min}} \sum_{(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{E}} \boldsymbol{w_{ij}} \left| \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2 - \tilde{\boldsymbol{d}}_{\boldsymbol{ij}}^{\mathbf{2}} \right| , \tag{3.56}$$

where $\boldsymbol{w_{ij}}$ shows the reliability measure on the estimated pairwise distances. The basis vectors in Euclidean space $\mathbb{R}^{\boldsymbol{N}}$ are denoted by $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_N}\}$. The optimization expressed in equation (3.56) is not convex but can be relaxed as a convex minimization via

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{Y}} \sum_{(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{E}} \boldsymbol{w_{ij}} \left| (\boldsymbol{u_i} - \boldsymbol{u_j})^{\boldsymbol{T}} [\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{X}^{\boldsymbol{T}}, \boldsymbol{I_\zeta}] (\boldsymbol{u_i} - \boldsymbol{u_j})^{\boldsymbol{T}} - \tilde{\boldsymbol{d}}_{\boldsymbol{ij}}^{\mathbf{2}} \right| \\ \text{subject to} \quad [\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{X}^{\boldsymbol{T}}, \boldsymbol{I_\zeta}] \succeq \mathbf{0}, \quad \left\| \boldsymbol{X}^{\boldsymbol{T}} \mathbf{1}_{\boldsymbol{N}} \right\| = \mathbf{0} \end{aligned} \tag{3.57}$$

where $\boldsymbol{Y_{N\times N}}$ is a positive semidefinite matrix and $\succeq$ is a generalized matrix inequality on the positive semidefinite cone [19]. To further increase the accuracy, a gradient decent is applied

on the output of SDP minimization [14].

3.6.3 Algebraic S-Stress Method

The s-stress method for calibration extracts the topology of the ad hoc network by optimizing the cost function stated as

$$\hat{X} = \underset{X}{\operatorname{arg\,min}} \sum_{(i,j)\in E} w_{ij} \left(\|x_i - x_j\|^2 - \tilde{d}_{ij}^2 \right)^2 . \tag{3.58}$$

The reliability measure w_{ij} controls the least square regression stated in equation (3.58) which can be set according to the measure of $\tilde{d}_{ij}$. If $w_{ij} = \tilde{d}_{ij}^{-2}$, we have *elastic scaling* that gives importance to large and small distances. If $w_{ij} = 1$, large distances are given more importance than the small distances. In general, incorporation of $w_{ij} = \tilde{d}_{ij}^{\alpha}$, $\alpha \in \{\ldots,-2,-1,0,1,2,\ldots\}$ yields different loss functions and depending on the structure of the problem, one of them may work better than the other [26].

3.7 Experimental Analysis

3.7.1 A-priori Expectations

The simplest method that we discussed is the classical MDS algorithm. This method assumes that all the pairwise distances are known and in the case of missing entries and noise, it does not minimize a meaningful utility function. An extension of this method is MDS-MAP which replaces the missing distances with the shortest path. In many scenarios, this is considered as a coarse approximation of the true distances.

The SDP-based method on the other hand is known to perform fairly well with missing distance information. Together with its final gradient descent phase, has been shown to find good estimates of the location. However, since each distance information translates into a constraint in the semi-definite program, this approach is not scalable and becomes intractable for large sensor networks.

The alternative approach is to minimize the non-convex s-stress function. Although it is known to perform well in many conditions, in the case of missing distances, one cannot eliminate the possibility of falling into local minima using this approach. The approach that we proposed in this chapter exploits a matrix completion algorithm to recover the missing distances considering the low-rank as well as Euclidean properties of the distance matrix. The classic matrix completion does not take into account the EDM properties. By integrating the Cadzow projection, the estimated matrix has partial EDM properties, and hence we expect better reconstruction results. Further, by incorporating the full EDM structure, we achieve a Euclidean distance matrix completion algorithm and expect more fidelity in the reconstruction performance. In this section, we present thorough evaluation of ad hoc microphone array

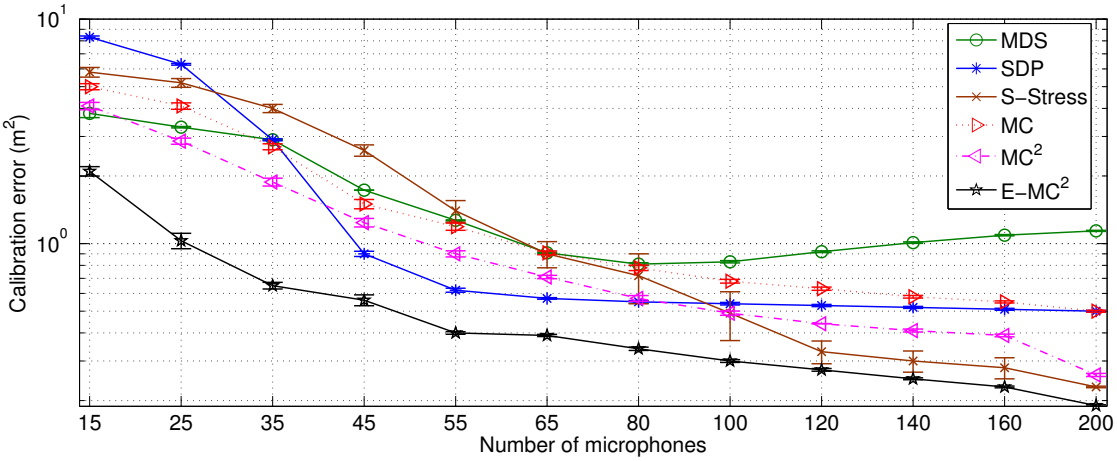

Figure 3.2: Calibration error (logarithmic scale) as defined in (3.15) versus the number of microphones. The standard deviation of noise on measured distances is ςd_{ij} where $\varsigma = \mathbf{0.0167}$. The error bars correspond to one standard deviation from the mean estimates.

calibration on simulated setups and real data recordings.

3.7.2 Simulated Data Evaluations

The simulated experiments are conducted to evaluate the performance of the proposed method and compare and contrast it against the state-of-the-art alternative approaches in different scenarios with varying number of microphones, magnitude of the pairwise distance measurements errors, percentage of missing distances as well as jitter.

The presented evaluation relies on a local connectivity assumption in pairwise distance measurements. We do not assume a particular (e.g. diffuse noise) model for pairwise distance estimation and the conclusions of this section hold for a general ad hoc array calibration framework where the pairwise distances may be provided by any other means meeting the local connectivity assumption.

Performance for Different Numbers of Microphones

In this section, we present the performance of ad hoc array calibration when the number of microphones varies from 15 to 200. The microphones are uniformly distributed on a disc of diameter 19 m. The maximum pairwise distance that can be measured is 7.5 m. In addition, 5% of the distances are assumed to be randomly missing. Hence, the total missing entries vary from 42% to 60%. The standard deviation of the noise on measured distances (expressed through (3.13)-(3.14)) between two microphones i and j is ςd_{ij} where $\varsigma = \mathbf{0.0167}$; the dependency of the noise level on the distance is due to the limitation of the diffuse noise coherence model for pairwise distance estimation as elaborated in [113].

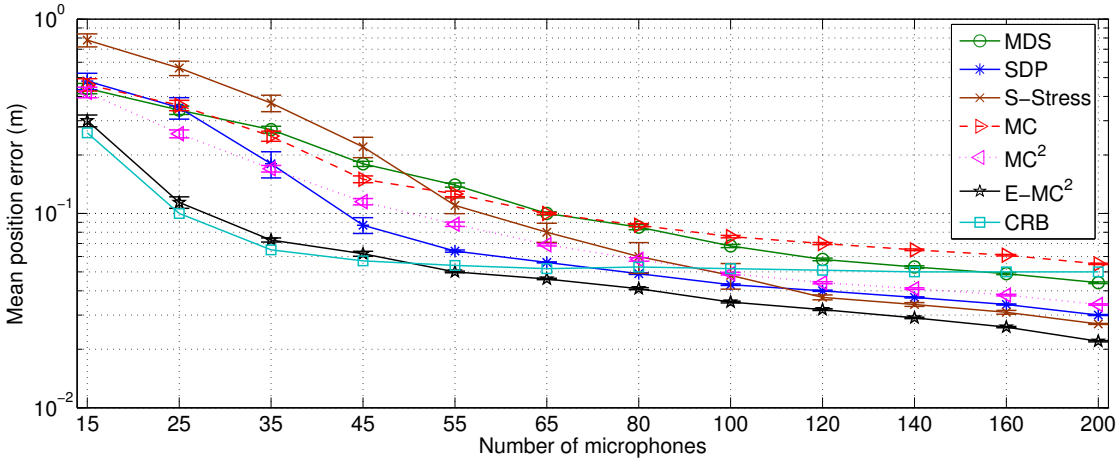

Figure 3.3: Mean position error (logarithmic scale) as defined in (3.59) versus the number of microphones. The standard deviation of the noise on measured distances is $\boldsymbol{\varsigma}\, \boldsymbol{d}_{ij}$ where $\boldsymbol{\varsigma} = \mathbf{0.0167}$. The error bars correspond to one standard deviation from the mean estimates.

The results for each number of microphones are averaged over 500 random configurations. The calibration error is quantified using the metric defined in (3.15). Furthermore, the absolute position of the microphones is estimated using the nonlinear optimization method [13] and the mean position error as defined in (3.59) over all configurations is evaluated. Figures 3.2 and 3.3 illustrate the results of matrix completion (MC), MC+Cadzow (MC^2) and the proposed Euclidean distance matrix completion (E-MC^2) algorithm compared with the related state-of-the-art methods as stated in Section 3.6; the error bars are shown for one standard deviation from the mean estimates. The Cramér rao bound (CRB) is quantified using the method elaborated in [94, 41].

The results show that the performance improves as the number of microphones increases. This observation is inline with the theoretical analysis provided in Section 3.5. The best results are achieved by the proposed E-MC^2 algorithm as it confines the search space to the Euclidean space through iterative EDM projections. We can see that for the number of microphones above 45, the error in position estimation is less than 6.2 cm and it reduces to 2.2 cm for 200 microphones. Although the mathematical proof of the unbiasedness of the proposed estimator is not achieved in this thesis, we empirically found no evidence of bias. Therefore, CRB provides a reasonable benchmark for our evaluation.

Performance for Different Noise Levels

To evaluate the effect of noise on calibration performance, similar (500) configurations of 45 microphones as generated in Section 3.7.2 are simulated. The level of white Gaussian noise added to the measured pairwise distance $\boldsymbol{d}_{ij}$ are varying as $\boldsymbol{\varsigma}\, \boldsymbol{d}_{ij}$ where $\boldsymbol{\varsigma} = \{\mathbf{0.0056}, \ldots, \mathbf{0.1}\}$. Figures 3.4 and 3.5 illustrate the results. We can see that the performance improves as the

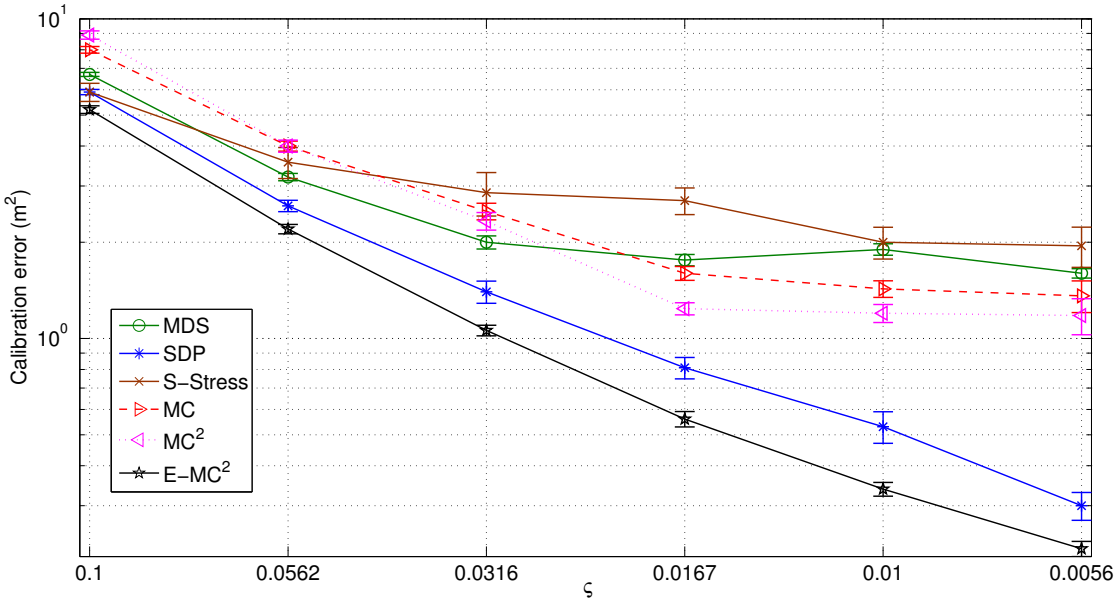

Figure 3.4: Calibration error (logarithmic scale) as quantified in (3.15) versus $\boldsymbol{\varsigma}$. The error bars correspond to one standard deviation from the mean estimates.

noise level gets smaller.

Based on the theoretical analysis of Section 3.5 as expressed in (3.28), a linear relationship between the calibration error of matrix completion and $\boldsymbol{\varsigma}$ is expected. The empirical observations are in line with this theoretical insight. As depicted in Figure 3.4, for $\boldsymbol{\varsigma < 0.0167}$, the second term in (3.28) is getting too small so the first term becomes dominant as the slope of the error reduction is reduced.

Performance for Different Missing Ratios

To study the sensitivity of the proposed algorithm to different levels of missing distances, a cubic room of unit dimensions ($\mathbf{1 \times 1 \times 1}$ $\mathbf{m^3}$) is simulated and 60 microphones are distributed uniformly at random positions. 300 random configurations are generated and the average mean position error is evaluated. As an alternative approach, the self-calibration method proposed by Crocco et al. [41] is implemented considering 30 sources and 30 sensors (thus 60 nodes in total). It may be noted that the number of nodes for calibration is equal for both approaches. The distances between all source and microphone pairs are known. Some of the distances are assumed to be missing at random. In addition, white Gaussian noise with standard deviation 0.02 m is added to the known distances. The simulated scenario mimics the evaluation setups of [41] and requires fixing the position of two microphones to derive the network position.

Figure 3.6 illustrates the errors in position estimation for different ratios of missing distances.

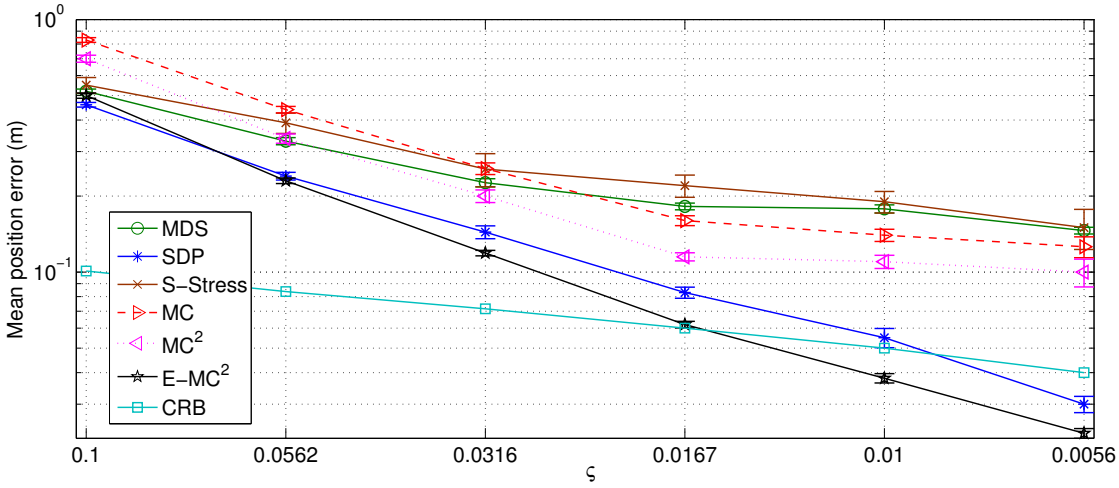

Figure 3.5: Mean position error (logarithmic scale) as defined in (3.59) versus $\boldsymbol{\varsigma}$. The error bars correspond to one standard deviation from the mean estimates.

We can see that up to 50% missing are effectively handled by the proposed algorithm. The rigorous analysis provided in Section 3.5 requires that $\boldsymbol{Np} \gg \log \boldsymbol{N}$ for the calibration error to be bounded; when the ratio of random missing entries is 60% (i.e. $\boldsymbol{p} = \mathbf{0.4}$), we have $\boldsymbol{Np}/\log \boldsymbol{N} = \mathbf{5.85}$ (violating the condition $\gg$) so the error in calibration is expected to increase significantly. The theoretical analysis is confirmed by this empirical observation.

Effect of Jitter on Calibration Performance

The study presented in this chapter assumes that the microphones are synchronized prior to calibration. If a pilot signal at sampling frequency $\boldsymbol{f}$ =16 kHz is used for synchronization, the effect of jitter can be modeled by a uniform error in distance measures as $[\frac{-\boldsymbol{c}}{2\boldsymbol{f}}, \frac{\boldsymbol{c}}{2\boldsymbol{f}}]$ where $\boldsymbol{c}$ is the speed of sound and set to 340 m/s. Hence, we can model the jitter as an additional uniform noise on the distance measures within the range of $[-\mathbf{1.065}, \mathbf{1.065}]$ cm.

The effect of jitter is evaluated for different levels of noise on the distances. The number of microphones is 45 distributed on a disc of diameter 19 m. 60% of distances are missing consisting of 5% random and 55% structured. The experiments are repeated for 300 random configurations and the average calibration error and position estimation error are quantified. Figure 3.7 illustrates the results. We can see that the effect of jitter on position estimation increases from 3 mm to 8 mm and its effect on calibration error increases from 0.01 m^2 to 0.09 m^2 as the distances are measured more accurately (smaller $\boldsymbol{\varsigma}$).

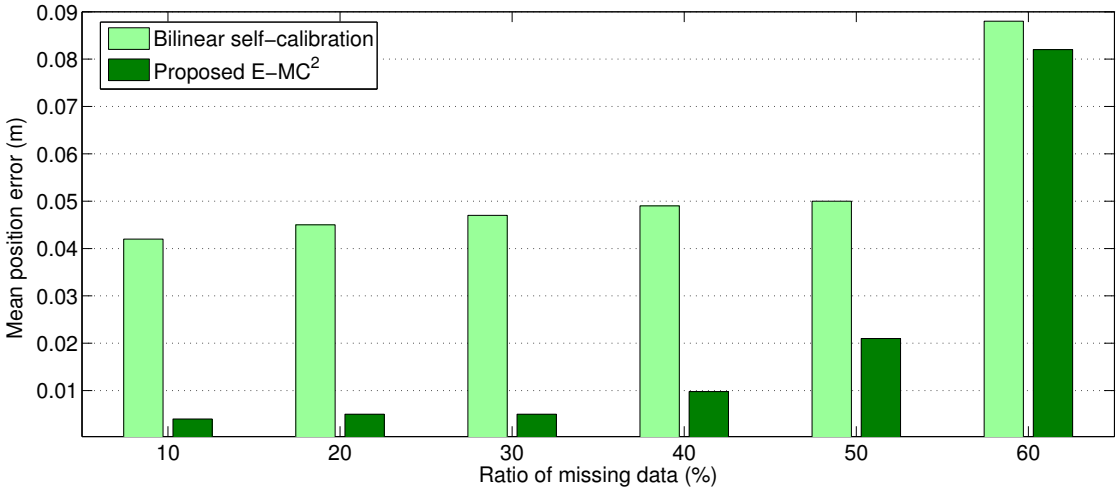

Figure 3.6: Mean position calibration error versus the ratio of missing pairwise distances for 30 sources and 30 microphones (60 nodes in total) considered in the self-calibration method [41] and 60 microphones used for the proposed E-MC^2 algorithm. The standard deviation of noise in pairwise distance estimation is 0.02.

Distributed Array Calibration

To further study the performance of the proposed approach for distributed array calibration, two scenarios are simulated. In the first scenario, a room of dimensions $\mathbf{11 \times 8 \times 5}$ m^3 is considered which yields $\boldsymbol{d_{max}} = \mathbf{101}$ cm [113] [3]. The reverberation time is about 430 ms. Two sets of 9-channel circular uniform microphone array of diameter 20 cm are simulated where the center of both compact arrays are 1 m apart. In the second scenario, a room of dimensions $\mathbf{8 \times 5.5 \times 3.5}$ m^3 is considered which yields $\boldsymbol{d_{max}} = \mathbf{73}$ cm [111]. The reverberation time is about 300 ms. A circular 9-channel microphone array of diameter 20 cm located inside another 6-channel circular array of diameter 70 cm is simulated.

The standard deviation of the noise on distance measures is $\boldsymbol{\varsigma}\, \boldsymbol{d_{ij}}$ where $\boldsymbol{\varsigma} = \mathbf{0.06}$. There are no random missing entries and all of the missing distances are due to the limitation of the diffuse noise model in distance estimation thus around 25% of the distances are missing in the first scenario (18-mic) and around 30% of the distances are missing in the second scenario (15-mic). The results are listed in Table 3.2. We can see that the positions are estimated with less than 1.6 cm error. Furthermore, we repeated the same experiment 25 times and averaged the estimates of the positions. We can see that the error after averaging is noticeably reduced.

[3] The maximum distance that can be estimated using the diffuse noise model depends on the size of the room and acoustic parameters. A linear relation between the maximum measurable distance and the room dimension has been shown rigorously (c.f. Chapter 2). Nevertheless, application of the diffuse noise method for pairwise distance estimation is just an example use case of the proposed algorithm and many alternative approaches can be exploited.

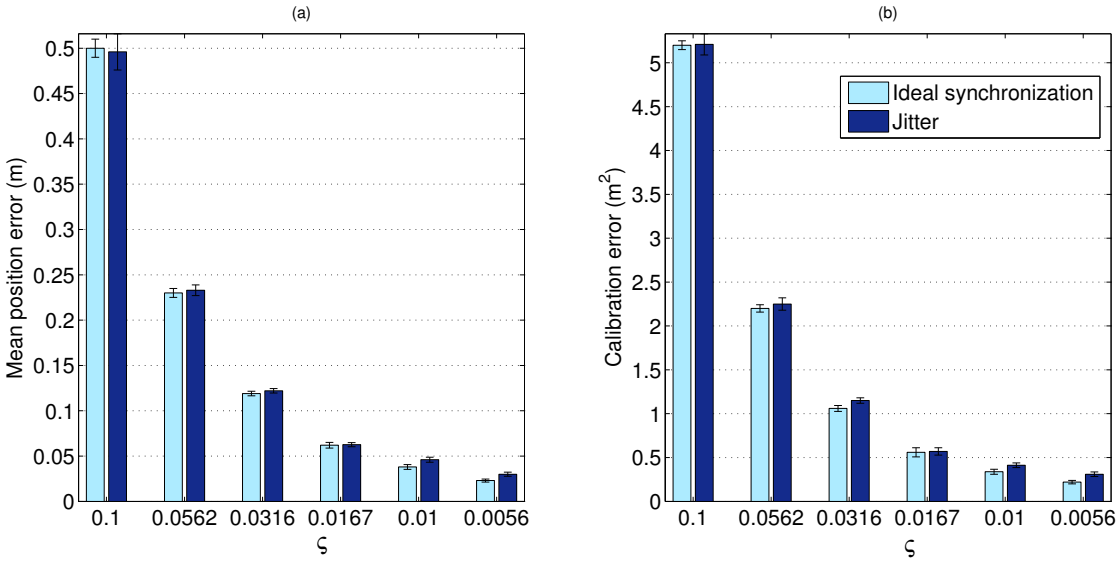

Figure 3.7: Effect of jitter on E-MC^2 algorithm quantified in terms of (a) mean position error as defined in (3.59) as well as (b) calibration error as defined in (3.15) versus $\boldsymbol{\varsigma}$. The error bars correspond to one standard deviation from the mean estimates. The number of microphones is 45 and 60% of the pairwise distances are missing.

3.7.3 Real Data Evaluation

The real data recordings are collected at Idiap's smart meeting room. The setup is similar to the real data experiments (ambient source diffuse field) presented in Section 2.6.1.

Pairwise Distance Estimation

In order to estimate the pairwise distances, we take two microphone signals of length **2.14** s, frame them into short windows of length 1024 samples using a Tukey function (parameter = 0.25) and apply Fourier transform. For each frame, we compute the coherence function. The average of the coherence functions over 1000 frames are computed and used for estimation of the pairwise distance by fitting a sinc function as stated in (3.1) using the algorithm described in [111]. This algorithm is an improved version of the distance estimation using diffuse noise coherence model which enables a reasonable estimate up to **73** cm. We empirically confirm that the distances beyond that are not reliably estimated so they are regarded as *missing*. Thereby, the following entries of the Euclidean distance matrix are missing: $\boldsymbol{d}_{10,11}, \boldsymbol{d}_{1,10}, \boldsymbol{d}_{7,10}, \boldsymbol{d}_{8,10}, \boldsymbol{d}_{5,11}, \boldsymbol{d}_{6,11}, \boldsymbol{d}_{7,11}$ (see Figure 3.8).

Geometry Estimation

In the scenario described above, microphone calibration is achieved in two steps. First, all methods are used to find the nine close microphones in order to evaluate them for geometry

Table 3.2: Performance of microphone array calibration in two scenarios. (1) Scenario 18-mic: two sets of 9-channel circular microphone array of diameter 20 cm; the center of both compact arrays are 1 m apart, and (2) Scenario 15-mic: a circular 9-channel microphone array of diameter 20 cm is located inside another 6-channel circular array of diameter 70 cm. The mean *position* error (cm) and the *calibration* error (cm^2) as defined in (3.15) are evaluated for different methods . The numbers in parenthesis corresponds to the error in position estimation if the experiments are repeated and averaged over 25 trials.

	Scenario 18-mic		Scenario 15-mic	
	Position (cm)	Calibration (cm^2)	Position (cm)	Calibration (cm^2)
MDS-MAP	3.3 (0.72)	175.8	3.18 (0.73)	170.5
SDP	2.1 (0.3)	96.3	4.64 (0.65)	258.8
S-Stress	6.8 (0.96)	265	7.05 (0.92)	281.5
MC	6.9 (1.35)	272	7.5 (1.55)	305
MC^2	6.56 (0.91)	225.1	6.8 (0.94)	274
E-MC^2	1.58 (0.37)	95.5	1.71 (0.41)	105.83

estimation when we have all distances. The geometry of these microphones is fixed and used to calibrate the rest of the network. Figure 3.10 demonstrates the results of MDS-MAP, SDP, s-stress and the proposed Euclidean distance matrix completion algorithm. The calibration error is quantified based on (3.15). The best results are achieved by the proposed algorithm with error **5.85** cm^2. The second place belongs to s-stress with error **6.14** cm^2 followed by MDS-MAP and SDP with errors **8.13** cm^2 and **8.63** cm^2 respectively (c.f. Table 3.3).

Table 3.3: Calibration errors (cm^2) as defined in (3.15) for different methods of microphone array calibration.

	Known		Missing	
	8-mic	9-mic	11-mic	12-mic
MDS-MAP	9	8.13	434.4	472
SDP	9.09	8.63	141	135
S-Stress	6.86	6.14	125	95
MC	10.6	9.75	133	115
MC^2	9.2	7.68	119	52
E-MC^2	6.5	5.85	49.6	46

Figure 3.11 provides a comparative illustration of the results of matrix completion (MC), MC+Cadzow (MC^2) and the proposed Euclidean distance matrix completion (E-MC^2) algorithm. We can see that MC^2 yields better result compared to MDS-MAP, SDP and MC, but worse than s-stress. The proposed E-MC^2 algorithm achieves the best performance (c.f.

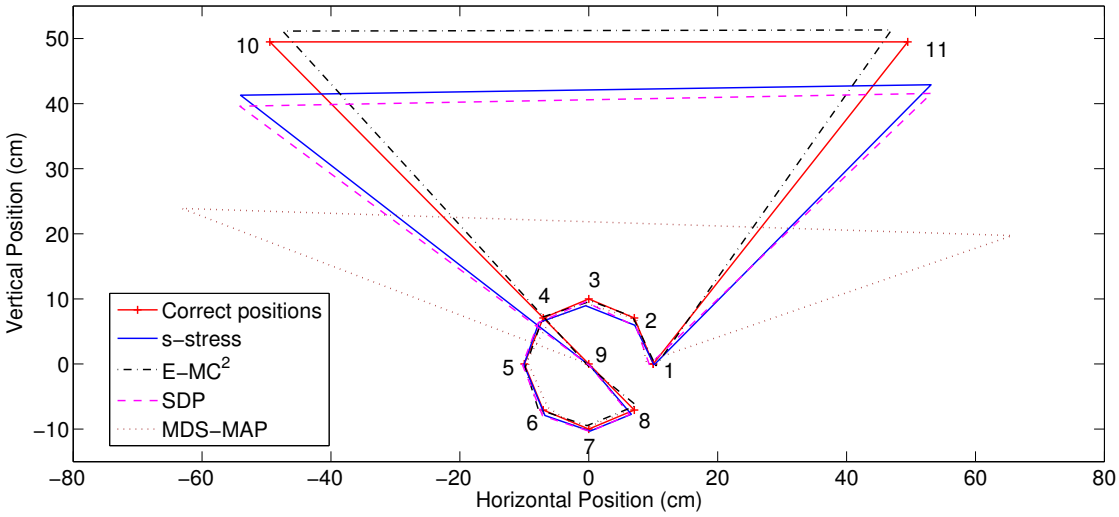

Figure 3.8: Calibration of the eleven-element microphone array while several pairwise distances are missing. The geometries are estimated using MDS-MAP, SDP, S-stress and the proposed proposed algorithm E-MC^2.

Table 3.3).

The scenario using eleven channels of microphones addresses the problem of having partial estimates of the distances for calibration of a microphone array. The experiments show that the proposed method offers the best estimation of the geometry as illustrated in Figure 3.8 and 3.9 with an error of **49.6** cm^2. As we can see, the proposed Euclidean distance matrix completion algorithm achieves less than half the error of the best state-of-the-art alternative.

The worst result belongs to MDS-MAP with error **434.4** cm^2 because the shortest path is a poor estimation of missing entries. The s-stress and SDP search the Euclidean space corresponding to the feasible positions hence, their performance are more reasonable with errors **141** cm^2 and **125** cm^2. The advantage of being constrained to a physically possible search space or close to it is considered in extensions of matrix completion in MC+Cadzow (MC^2) and the proposed method (E-MC^2) and achieves the best performance. These experimental evaluation confirm the effectiveness of the proposed algorithm and demonstrate the hypothesis that incorporating the EDM properties in matrix completion algorithm enables calibration of microphone arrays from partial measurements of the pairwise distances.

The theoretical analysis provided in Section 3.5 elucidates a link between the calibration error and the number of microphones. To demonstrate this relation, a calibration of a 8-channel circular array when the distances are all measured is performed. In addition, an extra microphone (#12) is also included which is located with a symmetry to microphone **10**. Hence, $\boldsymbol{d}_{12,11}, \boldsymbol{d}_{10,12}, \boldsymbol{d}_{3,12}, \boldsymbol{d}_{4,12}, \boldsymbol{d}_{5,12}$ are also missing. The calibration errors are listed in Table 3.3.

Furthermore, in addition to the calibration error expressed in (3.15), we apply the nonlinear

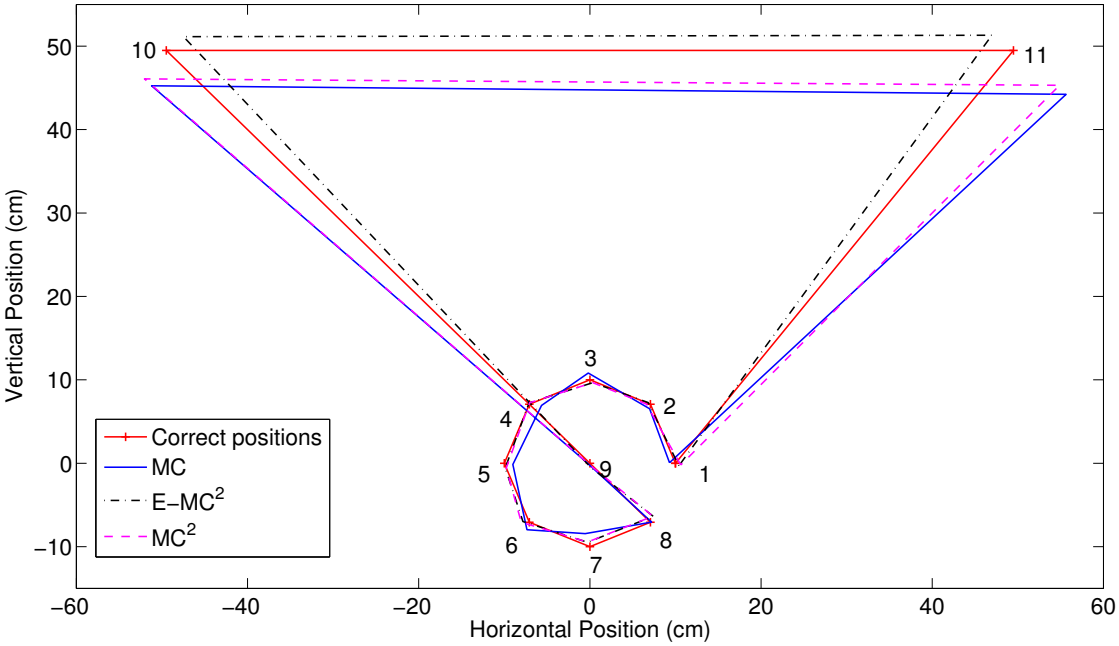

Figure 3.9: Calibration of the eleven-element microphone array while several pairwise distances are missing. The geometries are estimated using MC, MC+Cadzow (MC^2), and the proposed algorithm E-MC^2.

optimization proposed in [13] to find the best match between $\hat{\boldsymbol{X}}$ and $\boldsymbol{X}$ by considering various rigid transformations and quantify the position error as

$$\frac{1}{N}\sum_{n=1}^{N}\|\hat{\boldsymbol{x}}_n-\boldsymbol{x}_n\|_2. \tag{3.59}$$

The position errors are listed in Table 3.4. The results show that considering further microphone improves the calibration performance which is in line with the theoretical analysis of Section 3.5.

3.8 Conclusions

We proposed a Euclidean distance matrix completion algorithm for calibration of ad hoc microphone arrays from partially known pairwise distances. This approach exploits the low-rank property of the distance matrix and recovers the missing entries based on a matrix completion optimization scheme. To incorporate the properties of a Euclidean distance matrix, the estimated matrix at each iteration of the matrix completion is projected onto the EDM cone. Furthermore, we derived the theoretical bounds on the calibration error using matrix completion algorithm. The experimental evaluations conducted on real data recordings demonstrate that the proposed method outperforms the state-of-the-art techniques for ad hoc array calibration. This study confirmed that exploiting the combination of the rank condition

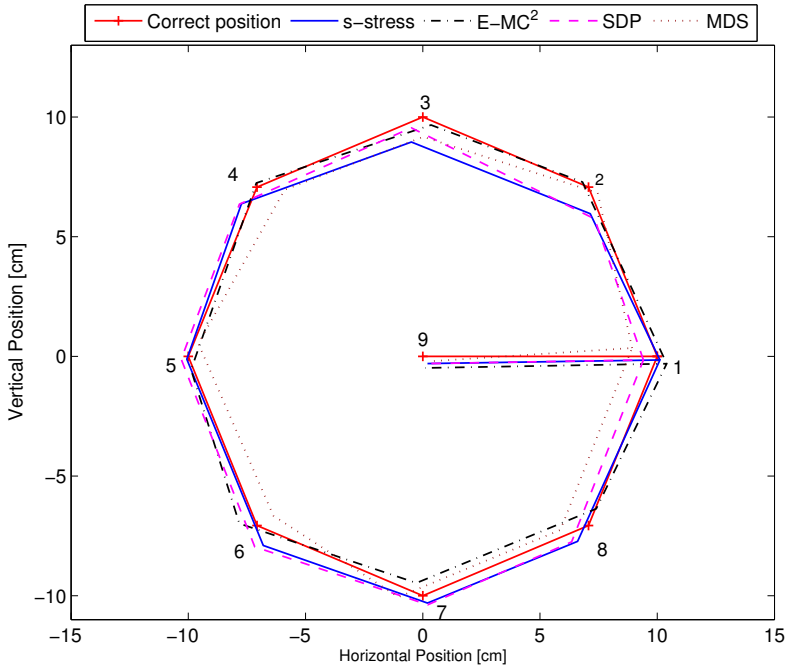

Figure 3.10: Calibration of the nine-element microphone array. The geometries are estimated using MDS-MAP, S-stress, SDP and the proposed Euclidean distance matrix completion algorithm, E-MC^2.

of EDMs, similarity in the measured distances, and iterative projection on the EDM cone leads to the best position reconstruction results. The proposed algorithm and the theoretical guarantees are applicable to the general framework of ad hoc sensor networks calibration.

Appendix 1. Proof of Theorem 2

The goal is to find the bound of the norm of the squared distance matrix with missing entries according to structures indicated by $\boldsymbol{E}$ and $\boldsymbol{S}$. Based on (3.6) and (3.11), we define matrix $\mathscr{E}$ as

$$\mathscr{E}_{ij} = \begin{cases} \mathbf{1} & \text{if } (i,j) \in \boldsymbol{E} \cap \boldsymbol{S} \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{3.60}$$

Both $\boldsymbol{E}$ and $\boldsymbol{S}$ are symmetric matrices, hence $\mathscr{E}$ is also symmetric. Due to the physical setup, we know that $\boldsymbol{\Psi}_E(\boldsymbol{M})_{ij} \leq 4a^2$ for all $i,j \in [N]$ and from the norm definition we have

$$\|\boldsymbol{\Psi}_E(\boldsymbol{M}^s)\|_2 \quad \leq \quad 4a^2 \max_{\|\mathbf{h}\|=\|\hbar\|=1} \sum_{i,j} |h_i|\,|\hbar_j|\,\mathscr{E}_{ij} = 4a^2 \|\mathscr{E}\|_2\,,$$

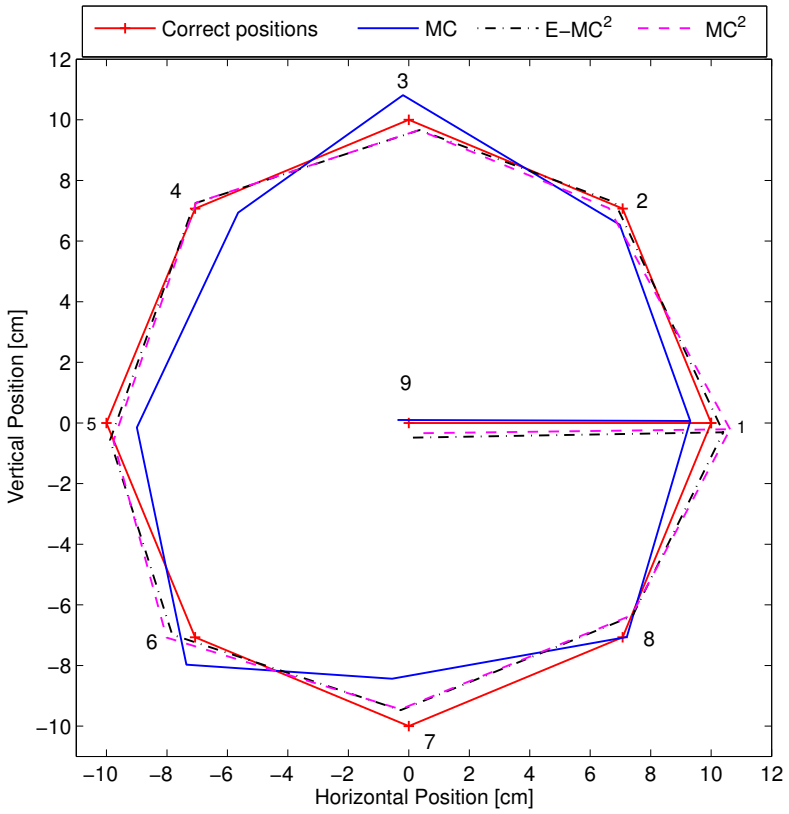

Figure 3.11: Calibration of the nine-element microphone array. The geometries are estimated using MC, MC+Cadzow (MC^2) and the proposed algorithm E-MC^2.

where $\mathbf{h} = [\boldsymbol{h_1}, \boldsymbol{h_2}, ..., \boldsymbol{h_N}]^T$ and $\hbar = [\hbar_1, \hbar_2, ..., \hbar_N]^T$ are right and left eigenvectors of matrix $\mathscr{E}$. In order to bound $\|\mathscr{E}\|_2$, we first define a binomial random variable vector $\boldsymbol{\nu} = [\boldsymbol{\nu_1}, \boldsymbol{\nu_2}, ..., \boldsymbol{\nu_N}]^T$ where

$$\boldsymbol{\nu_i} = \sum_{\boldsymbol{j} \in [\boldsymbol{N}]} |\mathscr{E}_{\boldsymbol{ij}}| \,. \tag{3.61}$$

Based on the Gershgorin circle theorem we have $\|\mathscr{E}\|_2 \leq \|\boldsymbol{\nu}\|_\infty$. Each entry in matrix $\mathscr{E}$ is one with probability $\boldsymbol{p\,q}$ where $\boldsymbol{q}$ is the probability that the entry is included in structured missing entries or

$$\boldsymbol{q} = \mathbb{P}\{|\boldsymbol{x_i} - \boldsymbol{x_j}| \geq \boldsymbol{d}_{\max}\} \,. \tag{3.62}$$

Hence, we have

$$\mathbb{E}[\boldsymbol{\nu_i}] = \boldsymbol{Npq} \,, \tag{3.63}$$

For bounding $\mathbb{E}[\boldsymbol{\nu_i}]$, it is necessary to bound $\boldsymbol{q}$. Figure 3.12.I depicts the lowest probability of missing distances if the microphone location with respect to the edge of the circular table has a distance more than $\boldsymbol{d}_{\max}$ and Figure 3.12.II depicts the highest probability if the microphone

Table 3.4: Position estimation errors (cm) as defined in (3.59) for different methods of microphone array calibration.

	Known		Missing	
	8-mic	9-mic	11-mic	12-mic
MDS-MAP	0.83	0.78	6.34	7.23
SDP	0.86	0.81	2.88	2.35
S-Stress	0.69	0.61	2.5	1.9
MC	1.1	0.97	2.6	2.1
MC^2	0.91	0.74	2.16	1.7
E-MC^2	0.64	0.59	1.06	1

is located right at the edge of the table.

The maximum of $\boldsymbol{d}_{\max}$ is $\boldsymbol{a}$. We denote the upper bound and lower bound with $\boldsymbol{q}_{\max}(\boldsymbol{a},\boldsymbol{d}_{\max})$ and $\boldsymbol{q}_{\min}(\boldsymbol{a},\boldsymbol{d}_{\max})$ respectively, therefore

$$\boldsymbol{q}_{\min}(\boldsymbol{a},\boldsymbol{d}_{\max}) \le \boldsymbol{q} \le \boldsymbol{q}_{\max}(\boldsymbol{a},\boldsymbol{d}_{\max})\,. \tag{3.64}$$

As illustrated in Figure 3.12. $\boldsymbol{q}_{\min}(\boldsymbol{a},\boldsymbol{d}_{\max}) = \mathbf{max}\{\mathbf{1} - \left(\frac{\boldsymbol{d}_{\max}}{\boldsymbol{a}}\right)^{\mathbf{2}}, \mathbf{0}\}$ and $\boldsymbol{q}_{\max}(\boldsymbol{a},\boldsymbol{d}_{\max}) = \mathbf{1} - \frac{\boldsymbol{B}}{\boldsymbol{\pi a^2}}$ where $\boldsymbol{B}$ is the intersection area between the two circles. By computing $\boldsymbol{B}$, we obtain

$$\boldsymbol{q}_{\max} = \mathbf{1} - \frac{\boldsymbol{2\gamma}}{\boldsymbol{\pi}} + \frac{\mathbf{1}}{\boldsymbol{2\pi}}\mathbf{sin}\,\boldsymbol{4\gamma} + \frac{\boldsymbol{2\xi^2}}{\boldsymbol{\pi}}[\boldsymbol{2\gamma} + \mathbf{sin}\,\boldsymbol{2\gamma}] - \boldsymbol{2\xi^2}\,, \tag{3.65}$$

where $\boldsymbol{\xi} = \boldsymbol{d}_{\max}/\boldsymbol{2a}$ and $\boldsymbol{\gamma} = \mathbf{sin}^{-1}\boldsymbol{\xi}$. Based on (3.63) and (3.64) we have

$$\boldsymbol{Npq}_{\min}(\boldsymbol{a},\boldsymbol{d}_{\max}) \le \mathbb{E}[\boldsymbol{v_i}] \le \boldsymbol{Npq}_{\max}(\boldsymbol{a},\boldsymbol{d}_{\max})\,. \tag{3.66}$$

By applying the Chernoff bound to $\boldsymbol{v_i}$ we have

$$\mathbb{P}\big(\boldsymbol{v_i} > (\mathbf{1}+\boldsymbol{\epsilon})\mathbb{E}[\boldsymbol{v_i}]\big) \le \boldsymbol{2^{-(1+\epsilon)\mathbb{E}[v_i]}}\,, \tag{3.67}$$

where $\boldsymbol{\epsilon}$ is an arbitrary positive constant. Therefore, based on (3.66) we have

$$\mathbb{P}\big(\boldsymbol{v_i} > (\mathbf{1}+\boldsymbol{\epsilon})\boldsymbol{Npq}_{\max}\big) \le \boldsymbol{2^{-(1+\epsilon)Npq}_{\min}}\,. \tag{3.68}$$

By applying the union bound we have

$$\mathbb{P}\big(\mathbf{max}_{\boldsymbol{i}\in[\boldsymbol{N}]}\boldsymbol{v_i} > (\mathbf{1}+\boldsymbol{\epsilon})\boldsymbol{Npq}_{\max}\big) \le \boldsymbol{2^{-(1+\epsilon)Npq}_{\min}+\mathbf{log}_2 N}\,. \tag{3.69}$$

We assume that $\boldsymbol{q}_{\min}$ and $\boldsymbol{q}_{\max}$ grow as $\mathcal{O}(\frac{\mathbf{log}_2 \boldsymbol{N}}{\boldsymbol{N}})$; this assumption indicates that the ratio of

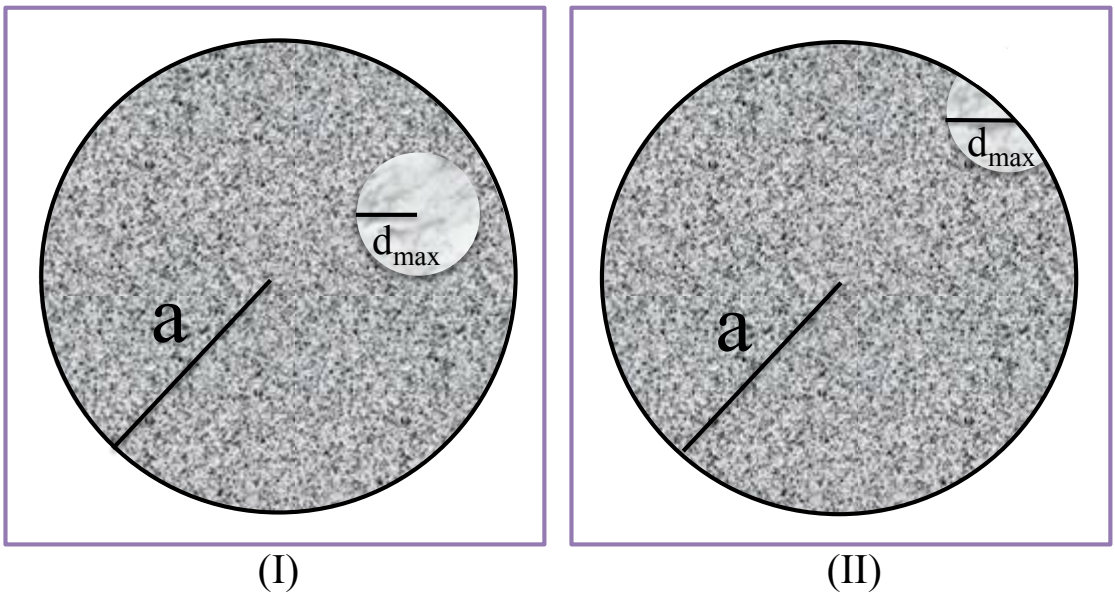

Figure 3.12: Scenario corresponding to the (I) lower bound and (II) upper bound of the probability $\boldsymbol{q}$ of structured missing distances.

the structured missing entries with respect to $\boldsymbol{N}$ decreases as $\boldsymbol{N}$ grows[4] or in other words, $\boldsymbol{d}_{\max}$ increases as the size of the network $\boldsymbol{N}$ grows. Therefore, we have

$$\mathbb{P}\big(\max_{\boldsymbol{i}\in[\boldsymbol{N}]} \boldsymbol{\nu_i} > (\mathbf{1}+\boldsymbol{\epsilon})\boldsymbol{N}\boldsymbol{p}\,\boldsymbol{q}_{\max}\big) \leq \boldsymbol{N}^{-\boldsymbol{\theta}}\,, \tag{3.70}$$

where the positive parameter $\boldsymbol{\theta} = (\mathbf{1}+\boldsymbol{\epsilon})\boldsymbol{p} - \mathbf{1}$; by choosing $\boldsymbol{\epsilon} \geq \mathbf{4}/\boldsymbol{p} - \mathbf{1}$, with probability greater than $\mathbf{1} - \boldsymbol{N}^{-\mathbf{3}}$, we have

$$\|\boldsymbol{\Psi}_{\boldsymbol{E}}(\boldsymbol{M}^{\boldsymbol{s}})\|_{\mathbf{2}} \leq \mathbf{4}\boldsymbol{a}^{\mathbf{2}} \max_{\boldsymbol{i}\in[\boldsymbol{N}]} \boldsymbol{\nu_i}\,, \tag{3.71}$$

and based on (3.70)

$$\|\boldsymbol{\Psi}_{\boldsymbol{E}}(\boldsymbol{M}^{\boldsymbol{s}})\|_{\mathbf{2}} \leq \mathbf{4}\boldsymbol{a}^{\mathbf{2}}(\mathbf{1}+\boldsymbol{\theta})\boldsymbol{q}_{\max}\boldsymbol{N}\,. \tag{3.72}$$

Therefore, we achieve

$$\|\boldsymbol{\Psi}_{\boldsymbol{E}}(\boldsymbol{M}^{\boldsymbol{s}})\|_{\mathbf{2}} \leq \boldsymbol{C}_{\mathbf{1}}''\boldsymbol{a}^{\mathbf{2}}\log_{\mathbf{2}}\boldsymbol{N}\,. \tag{3.73}$$

■

[4] This assumption can be dropped to achieve a tighter bound, but it increases the complexity of the proof.

Appendix 2. Proof of Theorem 3

Based on the noise model described in Section 3.3.3, $Z_{ij}^{\bar{s}}$ is obtained as

$$Z_{ij}^{\bar{s}} = d_{ij}^2 \Upsilon_{ij}\left(2+\Upsilon_{ij}\right) \approx 2d_{ij}^2\Upsilon_{ij}, \tag{3.74}$$

where $d_{ij} \leq d_{\max}$ and based on concentration inequality for 1-Lipschitz function $\|.\|$ on i.i.d random variables $\Psi_E(Z^{\bar{s}})$ with zero mean and sub-Gaussian tail with parameter $4\varsigma^2 d_{\max}^4$(3.14), (3.74) [117]

$$\mathbb{P}\left(\left|\left\|\Psi_E(Z^{\bar{s}})\right\| - \mathbb{E}\left(\left\|\Psi_E(Z^{\bar{s}})\right\|\right)\right| > t\right) \leq \exp\left(\frac{-t^2}{8\,\varsigma^2 d_{\max}^4}\right). \tag{3.75}$$

By setting $t = 2d_{\max}^2\sqrt{6\varsigma^2 \log N}$ we have

$$\left\|\Psi_E(Z^{\bar{s}})\right\| \leq \mathbb{E}\left(\left\|\Psi_E(Z^{\bar{s}})\right\|\right) + 2d_{\max}^2\sqrt{6\varsigma^2\log N} \tag{3.76}$$

with probability bigger than $1-N^{-3}$. So we need to extract bound for expectation of $\Psi_E(Z^{\bar{s}})$ that has symmetric random enties. By using Theorem 1.1 from [103],

$$\mathbb{E}\left(\left\|\Psi_E(Z^{\bar{s}})\right\|\right) \leq C_4\,\mathbb{E}\left(\max_{j\in[N]}\left\|\Psi_E(Z_{.j}^{\bar{s}})\right\|\right) \tag{3.77}$$

Furthermore by using union bound and with apply Chernoff bound on the sum of independent random variables [68]

$$\mathbb{E}\left(\max_{j\in[N]}\left\|\Psi_E(Z_{.j}^{\bar{s}})\right\|^2\right) \leq C_5 d_{max}^4\,\varsigma^2 pN \tag{3.78}$$

Since

$$\mathbb{E}\left(\max_{j\in[N]}\left\|\Psi_E(Z_{.j}^{\bar{s}})\right\|\right) \leq \sqrt{\mathbb{E}\left(\max_{j\in[N]}\left\|\Psi_E(Z_{.j}^{\bar{s}})\right\|^2\right)} \tag{3.79}$$

Base on relations (3.77), (3.78) and (3.79)

$$\mathbb{E}\left(\left\|\Psi_E(Z^{\bar{s}})\right\|\right) \leq C_6 d_{\max}^2\varsigma\sqrt{pN} \tag{3.80}$$

By using (3.80) and (3.76) for $pN \gg \log N$ we have

$$\left\|\Psi_E(Z^{\bar{s}})\right\| \leq C_2'' d_{\max}^2\varsigma\sqrt{pN} \tag{3.81}$$

■

4 Spatial Sound Localization via Multipath Euclidean Distance Matrix Recovery

In this chapter a novel localization approach is proposed in order to find the position of an individual source using recordings of a single microphone in a reverberant enclosure. The multipath propagation is modeled by multiple virtual microphones as images of the actual single microphone and a multipath distance matrix is constructed whose components consist of the squared distances between the pairs of microphones (real or virtual) or the squared distances between the microphones and the source. The distances between the actual and virtual microphones are computed from the geometry of the enclosure. The microphone-source distances correspond to the support of the early reflections in the room impulse response associated with the source signal acquisition. The low-rank property of the Euclidean distance matrix is exploited to identify this correspondence. Source localization is achieved through optimizing the location of the source matching those measurements. The recording time of the microphone and generation of the source signal is asynchronous and estimated via the proposed procedure. Furthermore, a theoretically optimal joint localization and synchronization algorithm is derived by formulating the source localization as minimization of a quartic cost function. It is shown that the global minimum of the proposed cost function can be efficiently computed by converting it to a generalized trust region sub-problem. Numerical simulations on synthetic data and real data recordings obtained by practical tests show the effectiveness of the proposed approach. The content of this chapter is under second review for publication at the Journal of Selected Topics in Signal Processing.

4.1 Introduction

Sound source localization is an active area of research with applications in hands-free speech communication, virtual reality, and smart environment technologies. This task is often achieved by collection of spatial observation of multiple acoustic microphones which requires a carefully designed infrastructure. To facilitate distributed processing of ubiquitous sensory data provided by ad hoc microphone arrays, we are motivated to address the problem of single channel source localization in a reverberant enclosure.

The previous approaches to source localization are largely confined to multi-channel processing techniques. In the following, we provide a brief overview of the prior work on reverberant source localization. We investigate the feasibility of single channel localization based on the underlying concepts of multichannel techniques.

The previous studies are directed down two avenues of research: A large body of work is being conducted on variants of multi-channel filtering to estimate the time difference of arrival (TDOA) or steering the directivity pattern of a microphone array. The generalized cross-correlation is typically used where the peak location of the cross-correlation function of the signal of two microphones is mapped to an angular spectrum for direction of arrival estimation [70]. A weighting scheme is often employed to increase the robustness of this approach to noise and multi-path effect. Maximum likelihood estimation of the weights has been considered as an optimal approach in the presence of uncorrelated noise, while the phase transform has been shown to be effective to overcome reverberation ambiguities [87, 17]. In addition, identification of the speaker-microphone acoustic channel has been incorporated for TDOA estimation and reverberant speech localization [95, 84]. Although TDOA-based techniques are practical and robust, they do not offer a high update rate as the short-frame correlations are susceptible to the spurious peaks caused by reverberation [23]. Other strategies have thus been sought for multiple-source localization and tracking [22, 73, 52]. In principle, TDOA-based localization techniques rely on the correlation of multiple spatially distinct measurements of source signal and they can not be applicable for single-channel localization.

An alternative approach to reverberant source localization is to design a beamformer for directional sound acquisition. This procedure enables source localization by scanning the spatial space and computing the steered response power (SRP) of the microphone array for all directions; the source direction corresponds to that of maximum power. Delay-and-sum, minimum variance beamformer, as well as generalized side-lobe canceler have been the most effective techniques for source localization [44, 47, 46, 130]. The SRP-based methods have a higher effective update rate compared to TDOA-based approaches, and they are applicable in multi-party scenarios. In particular, they can be made robust to multipath effect by applying appropriate weighting schemes such as phase-transform. The principle of directional scanning can be implemented through a single channel directional microphone. However, a mechanical engine must be set up for beam steering which requires costly specifications on the microphone design.

From a different perspective, a wide range of research endeavors is dedicated to identifying and exploiting the structure underlying the localization problem; examples include subspace and spatial sparsity methods. The subspace methods exploit the rank structure of the received signals' covariance matrix and impose a stationarity assumption to accurately estimate the source location. The effective techniques applied include minimum variance spectral estimation and multiple signal classification algorithm [45]. The underlying hypotheses are hard to apply in reverberant sound localization and alternative strategies have usually been

considered [5]. Furthermore, the ideas are not applicable on the variance of the measurements of a single microphone.

Sparse methods in the context of reverberant sound localization have been studied for model-based sparse component analysis [80, 9, 96]. It has been shown that incorporating spatial sparsity along with the underlying structure of the sparse coefficients enable super resolution in localization of simultaneous sources using very few microphones [9]. Relying on the image model for characterization of multipath propagation, this approach enables accurate localization of several simultaneous speech sources using recordings of under-determined mixtures; for instance up to eight overlapping speech sources can be localized with only four microphones. Although the principle of spatial sparsity holds for the recordings of a single microphone, it leads to ambiguities in signal reconstruction hence, localization can not be possible unless the original source signal is known. The image model of multipath propagation, however, identifies the relation between the room impulse response and the source/microphone position. This concept is fundamental to enable single-channel localization as we shall see in the subsequent sections.

Furthermore, the data-driven learning and generative modeling of location-dependent spatial characteristics has been shown promising for sound source localization in a reverberant environment; in [43] and [72] room- and microphone location-specific models were trained on white noise signals and incorporated for 2D-localization with two microphones. Nesta and Omologo [86] presented an approach that exploited sparsity of source signals in the cross-power spectral domain and accounted in a statistical manner for deviations of the sources' spatial characteristics from an ideal anechoic propagation model caused by multipath effect.

There is very little work in single-channel sound source localization. Recent studies rely on supervised training of a model of transfer functions for various positions in the room. In [116], the authors estimate the acoustic transfer function from observed reverberant speech using a clean speech model. A maximum likelihood estimation is applied in the cepstral domain assuming a Gaussian mixture model for the source. The estimation involves two stages: in the training stage, the distant speech signal is modeled for the potential locations so the acoustic transfer function is learned. In the testing stage, the location is inferred based on the location dependent speech models.

Another supervised single-channel localization algorithm is proposed in [118]. The problem is cast as recovering the controlling parameters of linear systems using diffusion kernels. The proposed algorithm computes a diffusion kernel with a specially-tailored distance measure. The kernel integrates the local estimates of the covariance matrices of the measurements into a global structure. This structure, referred to as a manifold, enables parameterization of the measurements where the parameters represent the position of the source.

Furthermore, some methods using the (ultra-)wideband radio signals are proposed to enable single-channel localization from the initial (deterministic) support of the impulse response. In [108], the notion of virtual anchors is introduced whose locations are unknown and

exploited via cooperation. Given the floor plan or the enclosure boundaries in [79], a maximum likelihood formulation of the source positioning is derived using the ranges to the virtual anchors. This approach has been shown promising, if the exact mapping between the range measurements and the reflective surfaces is known; However, no effective mechanism is devised to find the range-surface correspondences.

4.1.1 Main Contributions and Outline

In this chapter, we propose a novel approach to single-channel sound source localization exploiting the information carried by spatial sound. In contrast to the previous methods, no supervised training is required. We use the image model to characterize multipath acoustic propagation. According to this model, a single microphone in a reverberant enclosure leads to virtual microphones positioned at the mirrored locations of the microphone with respect to the reflective boundary of the enclosure. A reverberant signal is a collective observation resulting from the superposition of all microphone signals. We assume that the location of the microphone is known a priori and construct a distance matrix consisting of the pairwise distances between the microphone and its images and the source. The distances between the microphone and its images are known from the geometry of the room. The distances between the source and microphones are extracted from the spikes of the room impulse response function. However, extra processing is necessary to match the spikes to their corresponding image microphones. We exploit the low-rank structure of the Euclidean distance matrix and propose a procedure for image identification while compensating for the asynchronous time offset of recording. Furthermore, a joint localization and synchronization algorithm is proposed to find the global optimum of the exploited cost function. The main contributions of this chapter can be summarized as follows

- ◇ A novel approach to single-channel spatial sound localization exploiting the multipath propagation model and properties of Euclidean distance matrices.
- ◇ Algorithms to identify the virtual/real microphones from the early support (location of spikes) of the impulse response while estimating and compensating for the time offset of recording.
- ◇ Proposing a joint localization and synchronization algorithm via the global optimization of the appropriate squared range-based least square cost function.
- ◇ Extending the problem to distributed source localization framework using asynchronous recordings via aggregation of single-channel estimates.

The rest of this chapter is organized as follows. The problem of source localization in a reverberant enclosure is formulated in Section 4.2. In Section 4.3, we explain the proposed spatial sound localization scheme based on multipath distance matrix recovery: The low-rank property of the Euclidean distance matrix (EDM) is established in Section 4.3.1. Relying on the

Table 4.1: Summary of the notation.

Symbol	Meaning	Symbol	Meaning
ϵ	recording time offset	D	microphone-source distance matrix;
c	speed of sound	D_{ij}	element of row i and column j
δ	delay parameter equal to $c\epsilon$	$\widetilde{M}$	microphone-source measured squared distance matrix
d	distance between source microphones	$\hat{M}$	microphone-source estimated squared distance matrix
d_i	element i of distance vector	M	microphone-source squared distance matrix
N	number of microphones and source	Π	actual-virtual microphones Distance matrix
R	number of reflectors	X	positions matrix
z	source location	$\hat{X}$	estimated positions matrix

EDM properties, the algorithms for identifying the microphone-source distances along with localization and synchronization are devised in Section 4.3.2. Given the microphone-source distances, a theoretically optimal method to joint localization and synchronization is proposed in Section 4.3.3. The distributed source localization approach is elaborated in Section 4.4. The experimental results are presented in Section 4.5 and the conclusions are drawn in Section 4.6.

4.2 Statement of the Problem

In this section, we set out the problem formulation and the premises underlying the proposed localization approach.

4.2.1 Signal Model

Consider a scenario in which one microphone records the signal of an omni-directional source in a reverberant enclosure. The single-channel observation in time domain $O(t)$ consists of two components: a filtered version of the original signal $s(t)$ convolved with impulse response of the room and an additive noise term $n(t)$, thus expressed as

$$O(t) = h(t) * s(t) + n(t). \tag{4.1}$$

The time domain impulse response of the enclosure is assumed to be a train of Dirac delta functions corresponding to the direct path propagation and multipath reflections stated as

$$h(t) = \sum_{r=0}^{\mathscr{T}} c_r \delta(t - \tau_r), \tag{4.2}$$

where c_r denotes the attenuation factor of the r^{th} path pertaining to the spherical propagation as well as the absorption of air and reflective surfaces; τ_r designates the delay associated with acquisition of the sound traveling the distance between the source and microphone: τ_0 represents the direct path delay and τ_r, $r > 0$ corresponds to the reflected signal. We denote

the initial support of the room impulse response by

$$\boldsymbol{\Lambda} = \{\boldsymbol{\tau_0}, \ldots, \boldsymbol{\tau}_{\mathscr{R}}\}. \tag{4.3}$$

The goal is to estimate the source location based on the following available prior information:

- ◇ Geometry of the room
- ◇ Location of one microphone
- ◇ Early support of room impulse response, $\boldsymbol{\Lambda}$.

Due to asynchronous recording of signal and blind estimation of the room impulse response[1], there is an indeterminacy in support recovery so that $\boldsymbol{\tau_r} = \boldsymbol{\tau_r^*} + \boldsymbol{\epsilon}$ where $\boldsymbol{\tau_r^*}$ indicates the exact traveling time of the sound signal and $\boldsymbol{\epsilon}$ is the recording time offset.

Table 4.1 summarizes the set of important notation adopted in this chapter.

4.2.2 Image Microphone Model

In this section, we introduce the notion of *virtual microphones* based on the image model of multipath propagation [4]. The image model theory asserts that a reverberant sound field generated by a single source in an enclosure can be characterized as the superposition of multiple anechoic sound fields generated by images of the source with respect to the enclosure boundaries. Thereby, the initial support of impulse response corresponds to the direct-path traveling time of multiple images of the source.

The image model as described above indicates a duality between the image of source and microphones to model the multipath propagation [89]. Indeed, the observation of the source signal in a reverberant environment can be characterized as a collective observation of multiple microphones recording the direct-path propagation of a single source. The *virtual microphone*, $\boldsymbol{m_r}$ is obtained as the image of the actual microphone with respect to the $\boldsymbol{r}^{\text{th}}$ reflective surface. Fig.4.1 illustrates this duality in modeling the multipath effect.

According to the image microphone model, $\boldsymbol{\Lambda}$ is the propagation delay between the source and the set of microphones. We assume a cubic room shape in dimension $\boldsymbol{\kappa}$ consisting of $\boldsymbol{R}$ reflecting walls. The following relation holds between the components of $\boldsymbol{\Lambda}$ and the distances between source and actual/virtual microphones: $\boldsymbol{\tau_r} = \boldsymbol{d_r}/\boldsymbol{c} + \boldsymbol{\epsilon}$ where $\boldsymbol{d_r}$ denotes the microphone-source distance corresponding to time delay $\boldsymbol{\tau_r}$; $\boldsymbol{c}$ is the speed of sound and $\boldsymbol{\epsilon}$ is the recording time offset.

The time delays (support) of the initial echos provide a unique signature of the room geometry [48]. As the impulse response is also a function of the source and microphone

[1] The room impulse response is supposed to be estimated blindly and for this reason it is subject to synchronization (and scaling) ambiguity. A method of blind room impulse response estimation based on cross-relation formula [74] is evaluated in Section 4.5.3.

positions [4], knowing the room geometry and microphone position indicates a unique source position for a specific support structure in $\boldsymbol{\Lambda}$. The source localization thus amounts to addressing the following two problems: (1) finding the correspondence between $\boldsymbol{d_r}$ and $\boldsymbol{r^{th}}$ surface and (2) revealing the synchronization delay. To that end, we construct a multipath distance matrix from the pairs of microphones and source distances. The source localization is achieved exploiting the Euclidean distance matrix properties.

4.3 Spatial Sound Localization

We use the low-rank structure of the Euclidean distance matrices (EDM) to develop novel source localization and synchronization algorithms. To that end, a microphone-source squared distance matrix is constructed. The actual/virtual microphones pairwise distances are assumed to be known from the prior knowledge on the room geometry. The source distances to the microphones can be estimated from the early support of the room impulse response function. The difficulty then arises from the unknown microphone-source correspondence (mapping). Thus different distance matrices can be formed which are considered *incomplete* due to the unknown constellation of the microphone-source distance vector. Section 4.3.1 shows that the squared distance matrix has a low-rank structure. Relying on the results of Section 4.3.1, the low-rank structure of the EDM is exploited in Section 4.3.2 to devise a method to identify the microphone-source distance vector thus referred to as EDM matrix recovery, which in turn enables estimation of the source location and synchronization. Given the microphone-source distances, a joint localization and synchronization algorithm is formulated in Section 4.3.3 as a quartic cost function whose optimal solution can be efficiently computed by converting it to an instance of generalized trust region subproblem.

4.3.1 Multipath Euclidean Distance Matrix Rank Deficiency

The microphone pairwise distance matrix $\boldsymbol{\Pi}$ consists of components $\boldsymbol{\Pi_{ij}}$ where $\boldsymbol{\Pi_{ij}}$ denotes the distance between the actual/virtual microphones $\boldsymbol{i}$ and $\boldsymbol{j}$. These distances are assumed to be known a priori based on the image microphone model as explained in Section 4.2.2. The vector of distances between source and actual/virtual microphones is represented as

$$\boldsymbol{d} = [\boldsymbol{d_0}, \ldots, \boldsymbol{d_R}]^\top \tag{4.4}$$

where $.^\top$ denotes the transpose operator and $\boldsymbol{R}$ is the number of reflectors. The microphone-source multipath distance matrix is constructed as

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Pi} & \boldsymbol{d} \\ \boldsymbol{d}^\top & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{D} \in \mathbb{R}^{\boldsymbol{N}\times\boldsymbol{N}} \tag{4.5}$$

where $\boldsymbol{N} = \boldsymbol{R} + \boldsymbol{2}$.

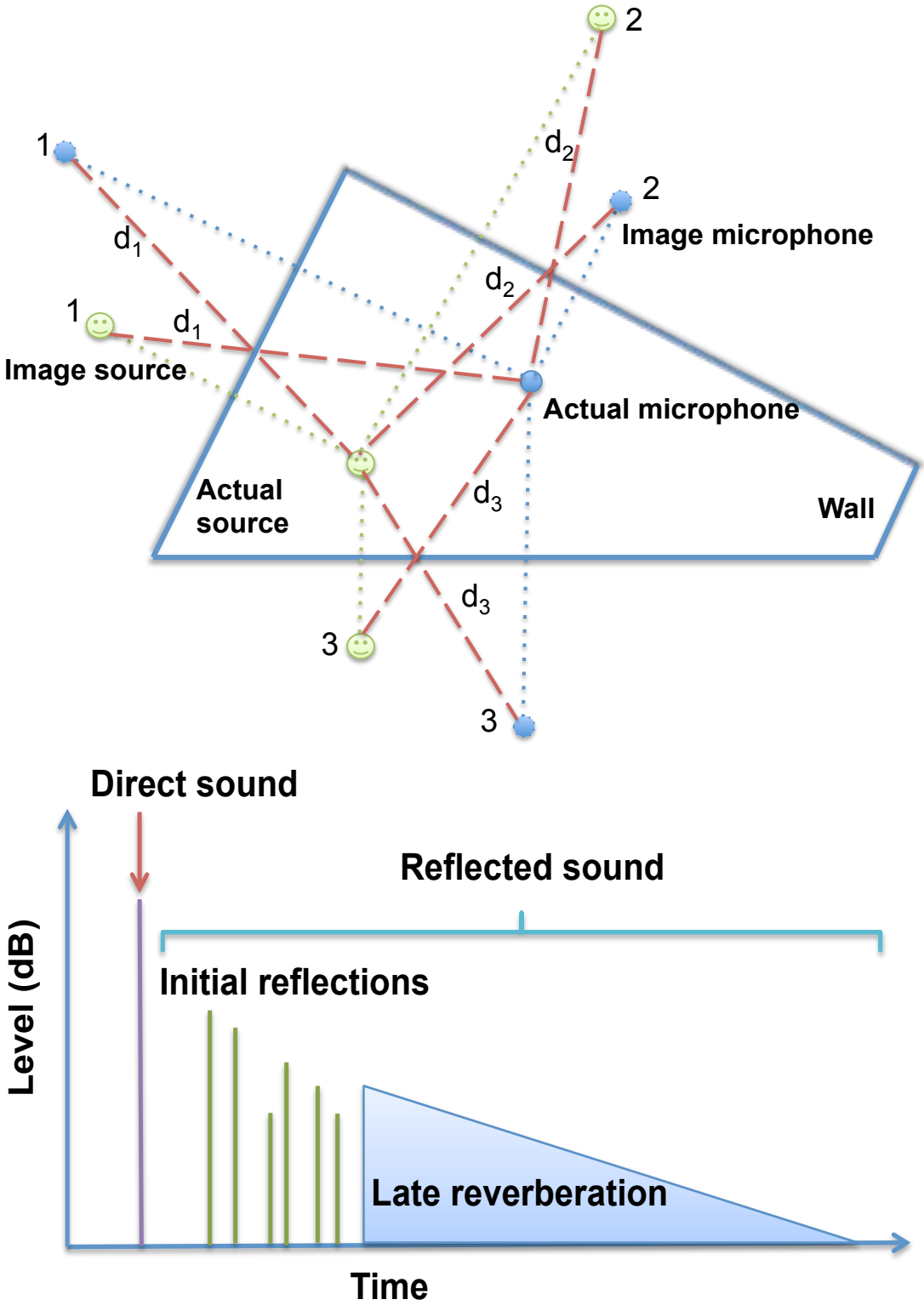

Figure 4.1: Image microphone model of a reverberant enclosure.

The components of $\boldsymbol{d}$ in (4.4) are assumed to be extracted from the identified support of the spikes in the room impulse response function $\boldsymbol{\Lambda}$. Hence, $\boldsymbol{D}$ can be known after estimation of $\boldsymbol{d}$. The matrix $\boldsymbol{D}$ as formed in (4.5) contains zero-diagonal elements and the cross-microphone and microphone-source distances on the off-diagonals. Hence, it also has a symmetric structure.

The Euclidean distance matrix $\boldsymbol{D}$ after applying a simple transformation (Hadamard product) has low rank as stated through the following lemma.

Lemma 1. [50] Consider a matrix $\boldsymbol{M}_{\boldsymbol{N}\times\boldsymbol{N}}$ consisting of the squared pairwise distances between pairs of source and microphones embedded in $\mathbb{R}^{\boldsymbol{\kappa}}$ defined as

$$\boldsymbol{M} = \boldsymbol{D} \circ \boldsymbol{D} = \left[\boldsymbol{D}_{ij}^{\boldsymbol{2}}\right], \quad \boldsymbol{i},\boldsymbol{j} \in \{\boldsymbol{1},\ldots,\boldsymbol{N}\} \tag{4.6}$$

where $\circ$ denotes the Hadamard product. The matrix $\boldsymbol{M}$ has rank at most $\boldsymbol{\eta} = \boldsymbol{\kappa} + \boldsymbol{2} < \boldsymbol{N}$.

Proof. Let $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{\kappa}\times\boldsymbol{N}}$ denote the position matrix consisting of the coordinates of each node (source or microphone) and $\boldsymbol{1}_{\boldsymbol{N}}$ is an all-one vector, we can write $\boldsymbol{M} = \boldsymbol{1}_{\boldsymbol{N}}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Lambda}\boldsymbol{1}_{\boldsymbol{N}}{}^{\top} - \boldsymbol{2}\boldsymbol{X}\boldsymbol{X}^{\top}$; thereby, $\boldsymbol{M}$ is the summation of three matrices where the first two of them are rank-1 and the third is rank-$\boldsymbol{\kappa}$. Hence $\boldsymbol{M}$ has rank at most $\boldsymbol{\eta} = \boldsymbol{\kappa} + \boldsymbol{2}$. □

Based on Lemma 1, there is a strong dependency among the entries of a squared distance matrix. Recent advances in matrix recovery have shown that by exploiting the low-rank structure, $\boldsymbol{N^2}$ components of $\boldsymbol{M}$ can be recovered from a subset of order $\boldsymbol{O(\eta N)}$ of its entries; the mathematical demonstration of this theory is elaborated in [29] and it is not required for the purpose of this chapter.

The squared distance matrix $\boldsymbol{M}$ as defined through (4.5)–(4.6) is indeterminate due to unknown permutation and offset underlying components of $\boldsymbol{d}$. This problem is addressed in the following section and the low-rank structure of $\boldsymbol{M}$ is exploited to recover the correct distances.

4.3.2 Multipath Euclidean Distance Matrix Recovery

Recovery of $\boldsymbol{M}$ can be achieved through the following constrained optimization problem

$$\begin{aligned} \hat{\boldsymbol{M}} &= \underset{\boldsymbol{M}}{\arg\min} \sum_{(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{E}} \left(\boldsymbol{M}_{\boldsymbol{ij}} - \tilde{\boldsymbol{M}}_{\boldsymbol{ij}}\right)^{\boldsymbol{2}} \\ \text{subject to:} \quad & \operatorname{rank}(\hat{\boldsymbol{M}}) = \boldsymbol{\eta} \quad \text{and} \quad \hat{\boldsymbol{M}} \in \mathbb{EDM}^{\boldsymbol{N}} \end{aligned} \tag{4.7}$$

where $\boldsymbol{E}$ denote the set of indices of the measured distances and $\widetilde{\boldsymbol{M}}$ is the corresponding squared distance matrix; by adopting the notation in [42], $\mathbb{EDM}^{\boldsymbol{N}}$ refers to the convex cone of all $\boldsymbol{N}\times\boldsymbol{N}$ Euclidean distance matrices. Furthermore, the Euclidean distance matrix must

satisfy the following properties [42]

$$\hat{\boldsymbol{M}} \in \mathbb{EDM}^{\boldsymbol{N}} \Longleftrightarrow \begin{cases} -\boldsymbol{a}^{\top}\hat{\boldsymbol{M}}\boldsymbol{a} \geq \boldsymbol{0} \\ \boldsymbol{1}^{\top}\boldsymbol{a} = \boldsymbol{0} \\ (\forall \|\boldsymbol{a}\| = \boldsymbol{1}) \\ \hat{\boldsymbol{M}} \in \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}} \end{cases} \tag{4.8}$$

for any vector $\boldsymbol{a} \in \mathbb{R}^{\boldsymbol{N}}$, where $\mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}}$ designates the space of symmetric, positive hollow matrices.

We assume that the components corresponding to the actual-virtual microphones pairwise distances $\boldsymbol{\Pi}$ in (4.6) are known based on prior knowledge on the room geometry and the actual microphone location. However, the following two problems associated with recovering $\boldsymbol{M}$ need to be resolved:

1. The correspondence between the spikes in the impulse response and the boundaries of the room.
2. The time shift for synchronization of the source signal generation and recording.

We refer to the first objective as *image identification* and to the second one as *synchronization*. These two tasks are the goal of the multipath Euclidean distance matrix recovery algorithm and are elaborated in the following sections.

Image Identification

Let $\boldsymbol{\Xi}$ denote the set of all possible permutations of the components of $\boldsymbol{d}$ defined in (4.4). Hence, the cardinality of $\boldsymbol{\Xi}$ is $\boldsymbol{(N-1)!}$. The key idea is that for the correct permutation, the squared distance matrix (4.6) is low-rank (Lemma 1). To formalize this idea, each member $\boldsymbol{\Xi_{\pi}}$ of the set is used to build a distance matrix as

$$\begin{aligned} &\tilde{\boldsymbol{D}}_{\boldsymbol{\pi}} = \begin{bmatrix} \boldsymbol{\Pi} & \boldsymbol{\Xi_{\pi}} \\ \boldsymbol{\Xi_{\pi}}^{\top} & \boldsymbol{0} \end{bmatrix}, \\ &\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}} = \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}} \circ \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}}, \quad \widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}} \in \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}}, \quad \boldsymbol{\pi} \in \boldsymbol{[(N-1)!]}. \end{aligned} \tag{4.9}$$

The goal of image identification is to suitably assign the spikes extracted from the room impulse response to their corresponding microphones. We recall that if the components of vector $\boldsymbol{\Xi_{\pi}}$ are in correct order, augmenting the microphone pairwise distances matrix $\boldsymbol{\Pi}$ with $\boldsymbol{\Xi_{\pi}}$ yields $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}$ of rank $\boldsymbol{\eta}$.

In theory, considering only $\boldsymbol{R}$ initial reflections, i.e. $\boldsymbol{N} = \boldsymbol{R} + \boldsymbol{2}$, seems enough to locate the first order image microphones and their correspondence to the unique location of the source. In practice, however, greater values, i.e., $\mathscr{R} > \boldsymbol{R}$, can be taken into account to distinguish the

echoes of the principal reflectors from the spurious peaks caused by the furniture. We discuss more on this issue in Section 4.5.3.

Synchronization

The time discrepancy between the source signal generation and recording causes a delay of $\boldsymbol{\epsilon}$ in the estimated impulse response function. In this section, we propose a new method to compensate this time shift for synchronization.

To model the effect of time difference in signal generation and recording, we define the vector $\boldsymbol{\Xi}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ whose $\boldsymbol{j}$th element is computed as $\boldsymbol{c}\,(\boldsymbol{\tau}_{\boldsymbol{j}}^{\boldsymbol{\pi}} - \tilde{\boldsymbol{\epsilon}})$ where $\boldsymbol{\tau}_{\boldsymbol{j}}^{\boldsymbol{\pi}}$ is a member of the set $\boldsymbol{\Lambda}$ (defined in (4.3)) at permutation $\boldsymbol{\pi}$ and $\tilde{\boldsymbol{\epsilon}}$ is the current estimate of $\boldsymbol{\epsilon}$. Construction of $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ in (4.9) is then revised using $\boldsymbol{\Xi}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ for augmenting $\boldsymbol{\Pi}$ thus

$$\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} = \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} \circ \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}, \quad \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} \triangleq \tilde{\boldsymbol{D}}_{\boldsymbol{\pi}} - \begin{bmatrix} \boldsymbol{0} & \boldsymbol{c}\,\tilde{\boldsymbol{\epsilon}}\,\boldsymbol{1}_{N-1} \\ \boldsymbol{c}\,\tilde{\boldsymbol{\epsilon}}\,\boldsymbol{1}_{N-1}^{\top} & \boldsymbol{0} \end{bmatrix}. \tag{4.10}$$

Let us denote a vector of desired microphone-source distances with $\bar{\boldsymbol{d}}_{\boldsymbol{\pi}}$ that corresponds to the last row of a desired squared distance matrix $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}$. Similarly, the vector of microphone-source distances extracted from $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}$ is represented by $\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}}$, the synchronization delay parameter $\tilde{\boldsymbol{\delta}} = \boldsymbol{c}\,\tilde{\boldsymbol{\epsilon}}$ is obtained through

$$\begin{aligned} &\mathscr{F}(\boldsymbol{\delta}) = \left\| \bar{\boldsymbol{d}}_{\boldsymbol{\pi}} \circ \bar{\boldsymbol{d}}_{\boldsymbol{\pi}} - \left(\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}} - \boldsymbol{\delta}\,\boldsymbol{1}_N\right) \circ \left(\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}} - \boldsymbol{\delta}\,\boldsymbol{1}_N\right) \right\|_2^2 \\ &\tilde{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\mathbf{arg\,min}}\, \mathscr{F}(\boldsymbol{\delta}) \Rightarrow \tilde{\boldsymbol{\epsilon}} = \tilde{\boldsymbol{\delta}}/\boldsymbol{c} \end{aligned} \tag{4.11}$$

To solve this optimization problem, we take the derivative of the objective function $\mathscr{F}(\boldsymbol{\delta})$ and find the roots as

$$\begin{aligned} \frac{\partial \mathscr{F}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} &= -\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N-1}} \boldsymbol{4}(\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}} - \boldsymbol{\delta}) \left((\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}} - \boldsymbol{\delta})^{\boldsymbol{2}} - \bar{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}^{\boldsymbol{2}} \right) \\ &= (\boldsymbol{N-1})\boldsymbol{\delta}^{\boldsymbol{3}} - \boldsymbol{3}\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N-1}} \tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}\boldsymbol{\delta}^{\boldsymbol{2}} + \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N-1}} (\boldsymbol{3}\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}^{\boldsymbol{2}} - \bar{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}^{\boldsymbol{2}})\boldsymbol{\delta} \\ &+ \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{N-1}} \left(\bar{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}^{\boldsymbol{2}}\tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}} - \tilde{\boldsymbol{d}}_{\boldsymbol{\pi}\boldsymbol{j}}^{\boldsymbol{3}} \right) = \boldsymbol{0}. \end{aligned} \tag{4.12}$$

As the cubic polynomial in (4.12) has at most three roots, one can solve it analytically to find the global minimizer of the cost function defined in (4.11).

Source Localization

The source location is obtained from the recovered multipath distance matrix. The goal of a low-rank matrix recovery algorithm is to estimate a Euclidean distance matrix with elements as close as possible to the known entries. We use an exhaustive search through all possible permutations to solve (4.7) based on iterative augmentation of the distance matrix as expressed in (4.9). Unless $\boldsymbol{\Xi_\pi}$ consists of correct order of images, the $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}$ does not correspond to a Euclidean distance matrix, so we propose to project $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}$ on to the cone of Euclidean distance matrices, $\mathbb{EDM}^{\boldsymbol{N}}$. To this end, we apply a projection, $\mathscr{P}: \mathbb{S}_{\boldsymbol{h}}^{\boldsymbol{N}} \longmapsto \mathbb{EDM}^{\boldsymbol{N}}$ and measure the distance between the estimated matrix and the EDM cone [112, 115].

The simplest way to achieve the objective of (4.7) is via singular value decomposition (SVD). The projection $\mathscr{P}$ is implemented by sorting the singular values and thresholding the smaller ones to achieve the desired rank. This approach is summarized in Algorithm 1. The position matrix is denoted by $\boldsymbol{X}_{\boldsymbol{N}\times\boldsymbol{\kappa}}$ whose $\boldsymbol{i}^{\text{th}}$ row, $\boldsymbol{x}_{\boldsymbol{i}}^{\top} = [\boldsymbol{x}_{\boldsymbol{i}\mathbf{1}}, \ldots, \boldsymbol{x}_{\boldsymbol{i}\boldsymbol{\kappa}}], \forall\, \boldsymbol{i} \in \{\mathbf{1}, \ldots, \boldsymbol{N}-\mathbf{1}\}$, is the position of microphone $\boldsymbol{i}$ in $\boldsymbol{\kappa}$-dimensional space. The order of the positions in $\boldsymbol{X}$ corresponds to the pairwise distances in $\boldsymbol{M}$. Hence, from the definition of (4.9), the last row corresponds to the source position $\boldsymbol{z}^{\top} = [\boldsymbol{z}_{\mathbf{1}}, \ldots, \boldsymbol{z}_{\boldsymbol{\kappa}}]$.

Algorithm 1 is an alternative coordinate descent approach consisting of two steps. Initializing $\tilde{\boldsymbol{\epsilon}}$ to zero and choosing permutation $\boldsymbol{\pi}$, in the first step, the squared distance matrix $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ as defined in (4.10) is double centered [40][2] (steps 4) followed by SVD to obtain a low-rank matrix $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ along with the position matrix $\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$; $\bar{\boldsymbol{d}}_{\boldsymbol{\pi}}$ is equal to the last row of $\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$. In the second step, $\tilde{\boldsymbol{\epsilon}}$ is updated by solving (4.11) and (4.12). Based on the new estimate of $\tilde{\boldsymbol{\epsilon}}$, $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ is updated using (4.10). These steps are repeated until $\tilde{\boldsymbol{\epsilon}}$ converges or the maximum number of iterations is reached. This procedure is executed for all possible permutations and $\boldsymbol{F}_{\boldsymbol{\pi}} = \|\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} - \bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\|_{\mathrm{F}}$ is computed. The permutation with smallest error $\boldsymbol{F}_{\boldsymbol{\pi}}$ denotes the source location, $\hat{\boldsymbol{z}}$, and synchornisation delay $\hat{\boldsymbol{\epsilon}}$.

The SVD-based low-rank projection does not incorporate the full set of EDM properties, thus it is suboptimal. More precisely, to achieve all the EDM properties, the projected matrix must satisfy the properties expressed in (4.8). Hence, we search in the EDM cone using the following cost function [115]

$$\mathscr{H}(\boldsymbol{X}, \widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}) = \left\| \mathbf{1}_{\boldsymbol{N}} \boldsymbol{\Lambda}^{\top} + \boldsymbol{\Lambda} \mathbf{1}_{\boldsymbol{N}}{}^{\top} - \mathbf{2}\boldsymbol{X}\boldsymbol{X}^{\top} - \widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}} \right\|_{\mathrm{F}}^{\mathbf{2}}, \tag{4.13}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{X} \circ \boldsymbol{X})\mathbf{1}_{\boldsymbol{\kappa}}$. The known microphone locations are used as the anchor points and only the source position is updated. The minimum of $\mathscr{H}(\boldsymbol{X}, \widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}})$ with respect to $\boldsymbol{z}_{\boldsymbol{i}}, \boldsymbol{i} = \{\mathbf{1}, \ldots, \boldsymbol{\kappa}\}$ can be computed by equating the partial derivative of equation (4.13) with respect to each individual coordinate $\boldsymbol{z}_{\boldsymbol{i}}$ to zero. Similar to (4.12), a third-order polynomial is obtained with

[2] Torgerson's double centering [40] as implemented in step 4 of Algorithm 1, is subtracting the row and column means of the matrix from its elements, adding the grand mean and multiplying by -1/2. The double centered matrix is scalar products relative to the origin and the coordinates is determined by the singular value decomposition (steps 5-6).

Algorithm 1 SVD-Localization

Input: Matrix $\widetilde{\boldsymbol{M}}$
Output: Estimated positions $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\boldsymbol{\epsilon}}$
1. **For** every $\boldsymbol{\pi} \in [|\Xi|]$ do the following steps
2. Initialize $\tilde{\boldsymbol{\epsilon}} = \mathbf{0}$.
3. **Repeat**
4. Compute $\frac{-1}{2}\boldsymbol{J}\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\boldsymbol{J}$ where $\boldsymbol{J} = \mathbb{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$
5. Take the SVD of $\frac{-1}{2}\boldsymbol{J}\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\boldsymbol{J} = \boldsymbol{U}_{\boldsymbol{\pi}}\boldsymbol{\Sigma}_{\boldsymbol{\pi}}\boldsymbol{U}_{\boldsymbol{\pi}}^\top$
6. $\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} = \boldsymbol{U}_{\boldsymbol{\pi}}^{\boldsymbol{\kappa}}\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\pi}}^{\boldsymbol{\kappa}}}$, based on the largest $\boldsymbol{\kappa}$ eigenvalues
7. $\bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}} = (\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} \circ \bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}})\mathbf{1}_{\boldsymbol{\kappa}}$
8. $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} = \mathbf{1}_N\bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}^\top} + \bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}}\mathbf{1}_N^\top - 2\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}^\top} \longrightarrow \bar{\boldsymbol{d}}_{\boldsymbol{\pi}}$
9. Update $\tilde{\boldsymbol{\epsilon}}$ using (4.11).
10. **Until** $\tilde{\boldsymbol{\epsilon}}$ converges or maximum number of iterations is reached.
11. Compute Frobenius norm of error $\boldsymbol{F}_{\boldsymbol{\pi}} = \|\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} - \bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\|_{\mathrm{F}}$.
12. **End For**
13. **Return** Location and synchronization delay: $\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{z}} \leftarrow \mathbf{arg\,min}_{\boldsymbol{\pi}}\, \boldsymbol{F}_{\boldsymbol{\pi}}$

maximum three roots and the one which globally minimizes the cost function is chosen. Hence, the optimization is done coordinate-wise to obtain the new estimate $\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}$ and the corresponding squared distance matrix $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}$.

Updating $\tilde{\boldsymbol{\epsilon}}$ based on (4.12) causes $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}$ to deviate from the EDM cone. Hence, the optimization of (4.13) is repeated in an iterative fashion to project it back to the EDM cone. The stopping criterion is satisfied when the new estimate of $\tilde{\boldsymbol{\epsilon}}$ differs from the old one by less than a threshold.

Although the coordinate-wise optimization procedure finds the optimal solution for each individual coordinate, reaching the global optimum is not guaranteed. Nevertheless, the experimental evaluation presented in Section 4.5 confirms that indeed the algorithm approximately converges to the optimal point. (cf. Section 4.5).

The procedure of the EDM-Localization is summarized in Algorithm 2.

Algorithm 2 EDM-Localization

Input: Matrix $\widetilde{\boldsymbol{M}}$
Output: Estimated source position $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\boldsymbol{\epsilon}}$
1. **For** every $\boldsymbol{\pi} \in [|\Xi|]$ do the following steps **do**
2. Initialize $\tilde{\boldsymbol{\epsilon}} = \mathbf{0}$.
3. **Repeat**
4. $\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} = \mathbf{arg\,min}_X\, \mathscr{H}(\boldsymbol{X}, \widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}})$
5. $\bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}} = (\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} \circ \bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}})\mathbf{1}_{\boldsymbol{\kappa}}$
6. $\bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} = \mathbf{1}_N\bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}^\top} + \bar{\boldsymbol{\Lambda}}^{\tilde{\boldsymbol{\epsilon}}}\mathbf{1}_N^\top - 2\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\bar{\boldsymbol{X}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}^\top} \longrightarrow \bar{\boldsymbol{d}}_{\boldsymbol{\pi}}$
7. Update $\tilde{\boldsymbol{\epsilon}}$ using (4.11).
8. **Until** $\tilde{\boldsymbol{\epsilon}}$ converges or maximum number of iterations is reached.
9. Compute Frobenius norm of error $\boldsymbol{F}_{\boldsymbol{\pi}} = \|\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}} - \bar{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}\|_{\mathrm{F}}$.
10. **End For**
11. **Return** Location and synchronization delay: $\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{z}} \leftarrow \mathbf{arg\,min}_{\boldsymbol{\pi}}\, \boldsymbol{F}_{\boldsymbol{\pi}}$

4.3.3 Joint Localization and Synchronization via Generalized Trust Region Sub-problem

In Algorithm 1 and Algorithm 2, we used an iterative approach based on low-rank SVD approximation and EDM projection to find the source location and synchronization delay. Although the resulting solutions approximate a stationary point of the cost function, there is a possibility that the resulting stationary point is a local rather than a global minimum of the cost function. Notice that since the cost function is positive and it tends to infinity as the location of the source and the synchronization delay approach infinity, the global minimum is guaranteed to exist.

In this part, we formulate finding the source location $\boldsymbol{z}$ and the delay parameter $\boldsymbol{\delta} = \boldsymbol{c\epsilon}$, as a quartic optimization problem. We assume that the distances between source and microphones are known based on image identification. We theoretically analyze the cost function and show that its global minimum can be efficiently computed under some mild conditions on the position of microphones and their images.

Recall that $\boldsymbol{x_1, x_2, \ldots, x_{N-1}}$ denote the positions of the microphones along with their images, where $\boldsymbol{x_j} \in \mathbb{R}^{\boldsymbol{\kappa}}$ and $\boldsymbol{d_j}$, $\boldsymbol{j = 1, 2, \ldots, N-1}$ are positive numbers corresponding to the last row of the observed square distance matrix $\widetilde{\boldsymbol{M}}$ obtained after image identification; hence, $\boldsymbol{d_j}$ is the measured distance between the position of the $\boldsymbol{j}^{\text{th}}$ real/virtual microphone $\boldsymbol{x_j}$ and the location of the source $\boldsymbol{z}$.

We consider the following cost function for estimating the source location $\boldsymbol{z} \in \mathbb{R}^{\boldsymbol{\kappa}}$ and the synchronization error $\boldsymbol{\delta} \in \mathbb{R}$:

$$\mathscr{G}(\boldsymbol{z}, \boldsymbol{\delta}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N-1}} \left(\|\boldsymbol{z} - \boldsymbol{x_j}\|^2 - (\boldsymbol{d_j} - \boldsymbol{\delta})^2 \right)^2 . \tag{4.14}$$

Let $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\delta}})$ be the optimal estimate globally minimizing the cost function (4.14). Based on [12], we call the resulting estimate $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\delta}})$ the *synchronization-extension of squared-range-based least squares* (SSR-LS) estimate. Notice that because of synchronization error, we obtain a different quartic function than [12] and as a result a completely different instance of the generalized trust region sub-problem (GTRS).

The SSR-LS cost function (4.14) is a non-convex quartic polynomial function of $(\boldsymbol{z}, \boldsymbol{\delta})$. Generally, it is known that global minimization of polynomials in NP-hard. However, some specific instances such as GTRS have efficient polynomial-time algorithms. In the following, we address how the global minimum of (4.14) can be computed efficiently.

We first transform (4.14) into a constrained minimization problem. Notice that

$$\mathscr{G}(\boldsymbol{z}, \boldsymbol{\delta}) = \sum_{\boldsymbol{j=1}}^{\boldsymbol{N-1}} \left(\|\boldsymbol{z}\|^2 - \boldsymbol{\delta}^2 - 2\boldsymbol{x_j}^\top \boldsymbol{z} + 2\boldsymbol{\delta} \boldsymbol{d_j} + \|\boldsymbol{x_j}\|^2 - \boldsymbol{d_j^2} \right)^2 .$$

Therefore, setting $\gamma = \|z\|^2 - \delta^2$, one can write

$$\min_{(z,\delta)} \mathcal{G}(z,\delta) = \min_{(z,\delta,\gamma)} \left\{ \sum_{j=1}^{N-1} (\gamma - 2x_j^\top z + 2\delta d_j + \|x_j\|^2 - d_j^2)^2 : \right.$$
$$\left. \|z\|^2 - \delta^2 = \gamma \right\}.$$

Assuming $y = (z^\top, \delta, \gamma)^\top$, this can be simplified to

$$\min_y \left\{ \|Ay - b\|^2 : y^\top L y + 2f^\top y = 0 \right\}, \tag{4.15}$$

where

$$A = \begin{pmatrix} -2x_1^\top & 2d_1 & 1 \\ -2x_2^\top & 2d_2 & 1 \\ \vdots & \vdots & \vdots \\ -2x_{N-1}^\top & 2d_{N-1} & 1 \end{pmatrix}, b = \begin{pmatrix} d_1^2 - \|x_1\|^2 \\ \vdots \\ d_{N-1}^2 - \|x_{N-1}\|^2 \end{pmatrix}, \tag{4.16}$$

and

$$L = \operatorname{diag}(1_{\kappa\times 1}, -1, 0), f = \begin{pmatrix} 0_{1\times(\kappa+1)} & -0.5 \end{pmatrix}^\top. \tag{4.17}$$

Matrix A has the dimension $(N-1)\times(\kappa+2)$. We assume that $N \geq (\kappa+3)$ and matrix A has full column rank which implies that $A^\top A$ is positive definite and, in particular, nonsingular. Note that (5.3) is a problem of minimizing a quadratic function under a single quadratic constraint. These kinds of problems are called *generalized trust region sub-problem* (GTRS) [83]. Although usually non-convex, GTRS problems have necessary and sufficient optimality conditions which allows them to be efficiently solved. Specially, by [83] and [12], $y \in \mathbb{R}^{\kappa+2}$ is an optimal solution of (5.3) if and only if there is a $\lambda \in \mathbb{R}$ such that

$$(A^\top A + \lambda L)y = A^\top b - \lambda f, \tag{4.18}$$
$$y^\top L y + 2f^\top y = 0, \tag{4.19}$$
$$A^\top A + \lambda L \geq 0. \tag{4.20}$$

Let us define

$$J_{\mathrm{PD}} = \{\lambda \in \mathbb{R} : A^\top A + \lambda L > 0\}. \tag{4.21}$$

Notice that for every $\lambda \in J_{\mathrm{PD}}$, $A^\top A + \lambda L$ is a positive definite thus a nonsingular matrix. We have the following useful proposition which is an application of Theorem 5.1 in [83].

Proposition 1. The set J_{PD} is an open interval.

The proof is stated in Appendix.

Proposition 2. Let J_{PD} be as defined in (4.21). Then, J_{PD} is nonempty and bounded.

Proof. We assumed that A has full column rank, which implies that $A^\top A$ is positive definite thus $\mathbf{0} \in J_{\text{PD}}$. It remains to prove that J_{PD} is bounded from below and above.

For an upper bound, notice that L is an indefinite matrix. Let $w = (\mathbf{0}_{1\times\kappa}, 1, 0)^\top$ be an all zero vector with only one 1 in position $\kappa+1$. One can simply check that $w^\top L w = -1$ and $w^\top A^\top A w = \|Aw\|^2 = 4\sum_{j=1}^{N-1} d_j^2$. This implies that if $A^\top A + \lambda L$ is positive definite then $\lambda < 4\sum_{j=1}^{N-1} d_j^2$. This gives an upper bound $\hat{\lambda}_u = 4\sum_{j=1}^{N-1} d_j^2$ on the interval J_{PD}.

For a lower bound, let v be a unit norm vector with zero in its last two components. It follows that $v^\top(A^\top A + \lambda L)v > 0$ if $\lambda > -v^\top K v$, where $K = 4\sum_{j=1}^{N-1} x_j x_j^\top$. This implies that $\lambda > \hat{\lambda}_l = -\lambda_1(K)$, where λ_1 denotes the smallest eigen-value of the matrix K. Notice that as A is full rank, K is positive definite with $\lambda_1(K) > 0$. Therefore $J_{\text{PD}} \subset (\hat{\lambda}_l, \hat{\lambda}_u)$ and it is bounded. □

We are mostly interested in the feasible set of λ in (4.18). Let us define $J_{\text{PSD}} = \{\lambda \in \mathbb{R} : A^\top A + \lambda L \succeq \mathbf{0}\}$.

Proposition 3. Let $J_{\text{PD}} = (\lambda_l^*, \lambda_u^*)$ be the open interval as characterized by Proposition 2. Then, $J_{\text{PSD}} = \bar{J}_{\text{PD}} = [\lambda_l^*, \lambda_u^*]$ is a closed interval.

Proof. The proof results from Theorem 5.3 in [83]. □

If we assume that the feasible λ in (4.18) belongs to J_{PD}, then $A^\top A + \lambda D$ is positive definite, thus one can obtain the optimal solution by

$$\hat{y}(\lambda) = (A^\top A + \lambda L)^{-1}(A^\top b - \lambda f). \tag{4.22}$$

Moreover, one can find the optimal λ by replacing $\hat{y}(\lambda)$ in (4.20) and solving the equation $\phi(\lambda) = 0, \lambda \in J_{\text{PD}}$, where the function ϕ is defined by

$$\phi(\lambda) = \hat{y}(\lambda)^\top L \hat{y}(\lambda) + 2f^\top \hat{y}(\lambda). \tag{4.23}$$

It is also known from [83] that $\phi(\lambda)$ is strictly decreasing over J_{PD}. Therefore, it has only one solution which can be found by applying the bisection algorithm with the initial interval estimate $(\hat{\lambda}_l, \hat{\lambda}_u)$ obtained in Proposition 2. We assume that the optimal λ^* belongs to J_{PD}, thus $A^\top A + \lambda^* L$ is positive definite and nonsingular. There are rare cases in which λ^* belongs to the boundary. In our case, for example, this occurs when $\lambda^* \in \{\lambda_l^*, \lambda_u^*\}$, where λ_l^*, λ_u^* are as in Proposition 3. This case, as also explained in [54], belongs to the *hard instances* of the trust

Algorithm 3 SSR-LS

Input: Position of microphones and images $\boldsymbol{x_1, x_2, \ldots, x_{N-1}}$
Output: Estimated source position $\hat{\boldsymbol{z}}$ and synchronization delay: $\hat{\boldsymbol{\epsilon}}$
1. Build $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{L}$ and $\boldsymbol{f}$ according to (5.2), (4.17)
2. Define $\hat{\boldsymbol{y}}(\boldsymbol{\lambda})$ from (4.22)
3. Define function $\boldsymbol{\phi}(\boldsymbol{\lambda})$ from (4.23)
4. Set $\hat{\boldsymbol{\lambda}}_{\boldsymbol{u}} = \mathbf{4}\sum_{j=1}^{N-1} \boldsymbol{d_j^2}$
5. Set $\hat{\boldsymbol{\lambda}}_{\boldsymbol{l}}$ to the smallest eigen-value of $\boldsymbol{K} = \mathbf{4}\sum_{j=1}^{N-1} \boldsymbol{x_j x_j^\top}$
6. Solve $\boldsymbol{\phi}(\boldsymbol{\lambda}^*) = \mathbf{0}$ in the interval $(\hat{\boldsymbol{\lambda}}_{\boldsymbol{l}}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{u}})$
7. **Return** $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{\delta}})$ that is found from $\hat{\boldsymbol{y}}(\boldsymbol{\lambda}^*)$

region algorithm that can also be treated with a more refined analysis. In practice, considering the measurement noise, it is very rare to obtain the optimal $\boldsymbol{\lambda}^*$ on the boundary.

The procedure of the proposed SSR-LS joint synchronization-localization algorithm is summarized in Algorithm 3.

An application of the proposed single-channel localization method is to devise a distributed localization framework where each microphone provides an individual estimate of the source location. The microphones may have different recording time offset which is estimated and compensated separately to yield an estimate of the source location. The single-channel estimates are then aggregated to improve the source localization performance. This idea is elaborated in the following Section 4.4.

4.4 Distributed Source Localization

Extension of the algorithms presented in Sections 4.3.2 and 4.3.3 to accommodate more than one microphone data is straightforward and a similar formulation as presented earlier can be applied. However, the exhaustive search required for image identification becomes prohibitive for a large network of microphones. An alternative approach is distributed source localization. That is to aggregate the individual estimates provided by each microphone while the differences in time offsets are compensated locally.

Let the estimated distances between source and microphone $\boldsymbol{l}$ and its images be denoted by $\hat{\boldsymbol{d}}^{\boldsymbol{l}}$. Furthermore, we assume that every $\boldsymbol{R}+\mathbf{1}$ consecutive rows and columns of matrix $\boldsymbol{\Pi}$ correspond to the pairwise distances between each individual microphone and its images. Hence, we can form matrix $\mathscr{D}$ based on (4.5) as

$$\mathscr{D} = \begin{bmatrix} \boldsymbol{\Pi} & [\hat{\boldsymbol{d}}^{\boldsymbol{1}^\top}, \ldots, \hat{\boldsymbol{d}}^{\boldsymbol{m}^\top}]^\top \\ [\hat{\boldsymbol{d}}^{\boldsymbol{1}^\top}, \ldots, \hat{\boldsymbol{d}}^{\boldsymbol{m}^\top}] & \mathbf{0} \end{bmatrix}, \qquad \mathscr{D} \in \mathbb{R}^{\mathscr{N}\times\mathscr{N}} \tag{4.24}$$

where $\mathscr{N} = \boldsymbol{m}(\boldsymbol{R}+\mathbf{1})+\mathbf{1}$ and $\boldsymbol{m}$ is the total number of microphones. Thereby, the squared distance matrix of the microphone array is obtained as $\mathscr{M} = \mathscr{D} \circ \mathscr{D}$. As the last row of $\mathscr{M}$ consists of the separate estimates obtained by low-rank matrix recovery performed for

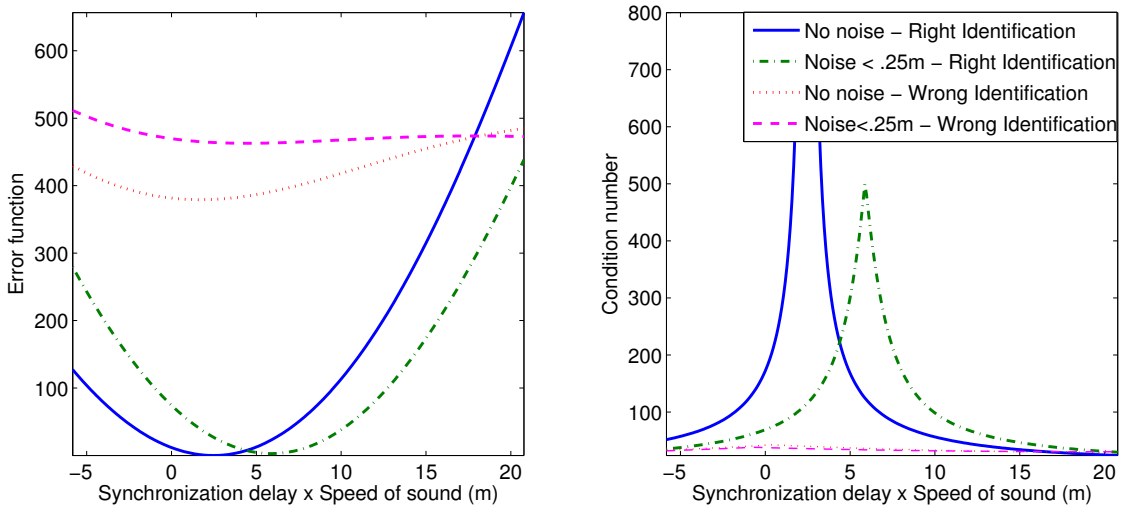

Figure 4.2: (left) Behavior of the error function $\boldsymbol{F_\pi}$ and (right) condition number of $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ in (4.10) for different synchronization delay when the images are identified in a right or wrong order. In this example, $\boldsymbol{\epsilon c} = \mathbf{3.4}\,\mathrm{m}$.

each microphone individually, the resulting matrix after concatenation of the distributed estimations may not fulfill the low-rank property. Thus, Algorithms 1–3 are run to yield the source location while the permutation is remained unchanged.

To summarize, the distributed microphones provide separate estimates of the microphone-source distances and the ultimate localization is achieved by estimating the source location best matching those individual estimates. The distributed localization framework can be particularly useful for ad hoc microphone setups. Further extension to multi-source scenarios is straightforward.

4.5 Experimental Results

In this section, we evaluate the performance of the proposed localization algorithms using synthetic and real data recordings. We assess the robustness of the algorithms with respect to jitter noise in the support of the spikes in the room impulse response as well as synchronization delay.

4.5.1 Single-channel Synchronization-Localization Performance

For simulation, we consider a $\mathbf{8}\times\mathbf{5.5}\times\mathbf{3.5}\,\mathrm{m}^3$ rectangular enclosure. The location of the source and microphone are randomly chosen in 100 trials. The random positions are generated such that the the distances to the boundaries are greater than 0.5 m. The speed of sound is assumed to be $\boldsymbol{c} = \mathbf{342}\,\mathrm{m/s}$. The sampling rate is 16 kHz.

The experiments are carried out on three simulated scenarios considering noise and

Table 4.2: Performance of joint source localization and synchronization using Algorithm 1: SVD-Localization, Algorithm 2: EDM-Localization and Algorithm 3: SSR-LS Algorithm. The left hand side quantifies the level of maximum error in estimation of microphone-source distances, $\boldsymbol{\Delta}$ measured in centimeters. The listed numbers quantifies the error in synchronization ($\boldsymbol{\mu}$s) - finding the correct synchronization parameter $\boldsymbol{\epsilon}$ - and source localization (cm) for different distance-noise levels. The numbers after ± indicate the 95% confidence interval.

Dis-Noise	**Synchronization Error (μs)**			**Localization Error (cm)**		
(cm)	**SVD-Loc.**	**EDM-Loc.**	**SSR-LS**	**SVD-Loc.**	**EDM-Loc.**	**SSR-LS**
0	0	0	0	0	0	0
5	23.52 ± 2.95	19.42 ± 1.38	0.71 ± 1.30	4.15 ± 0.12	3.46 ± 0.11	3.25 ± 0.11
10	44.97 ± 5.77	37.40 ± 2.85	2.49 ± 2.64	8.13 ± 0.25	6.76 ± 0.24	6.35 ± 0.22
15	70.87 ± 9.34	58.29 ± 4.60	2.96 ± 4.39	12.55 ± 0.37	10.19 ± 0.35	9.48 ± 0.32
20	100.1 ± 11.5	85.33 ± 6.00	7.14 ± 5.53	15.97 ± 0.52	13.22 ± 0.47	12.61 ± 0.45
25	123.0 ± 14.9	103.7 ± 8.16	7.61 ± 7.72	20.94 ± 0.66	16.71 ± 0.60	15.47 ± 0.53
30	148.0 ± 17.5	128.4 ± 9.06	10.09 ± 8.40	24.59 ± 0.74	20.07 ± 0.69	19.00 ± 0.63
40	204.2 ± 24.7	178.5 ± 12.9	16.94 ± 12.0	32.61 ± 0.99	27.21 ± 0.90	25.58 ± 0.87
50	242.7 ± 29.7	214.1 ± 16.8	24.84 ± 16.0	40.47 ± 1.27	33.01 ± 1.11	30.97 ± 1.07

synchronization delay in estimation of the room impulse response function. The level of noise indicates the error (cm) in microphone-source distance estimation. Denoting the estimated distance from the source to microphone $\boldsymbol{j}$ by $\tilde{\boldsymbol{d}}_{\boldsymbol{j}}$, we consider in our measurements $\|\tilde{\boldsymbol{d}}_{\boldsymbol{j}} - \boldsymbol{d}_{\boldsymbol{j}}\|_{\mathbf{2}} < \boldsymbol{\Delta}$ and evaluation is conducted for various values of $\boldsymbol{\Delta}$ as listed in the left hand side of Table 4.2. For each scenario, we run 400 random trials and average the results. Table 4.2 summarizes the error of synchronization and source localization using SVD-Localization, EDM-Localization and SSR-LS algorithms. It is important to mention that the SVD-Localization algorithm only extracts a possibly rotated or reflected version of the points in the configuration. Using the known real/virtual microphone positions as anchor points, we use the optimization problem proposed in [102] to find the absolute position of the source whereas EDM-Localization and SSR-LS directly yield the absolute source position.

For SVD-Localization and EDM-Localization, the maximum iterations for $\hat{\boldsymbol{\epsilon}}$ estimation is set to 50 and if the estimates in two successive iterations are less than 10e-5 different, the iterative synchronization is stopped earlier. It may be noted that the iterative synchronization procedure is applied only for the two first algorithms, whereas SSR-LS directly gives the $\boldsymbol{\epsilon}$.

We observe that in all scenarios the image identification is achieved correctly despite the error in estimation of the microphone-source distances. Furthermore, we observe that the results of EDM-Localization are better than SVD-Localization and they are very close to the global optimum solution of SSR-LS cost function, whereas the synchronization performance

of SSR-LS is significantly better. We also observe that the the coordinate-wise minimization in the EDM-Localization almost always converges to the SSR-LS global optimum point, however, in this chapter, we do not theoretically prove its global convergence.

From Table II, we see that the estimated delay parameters for the two approaches are quite different. One justification is that in the SVD method, we are looking for a three dimensional subspace as the embedding dimension for the microphone, images and the source. Now if there is a slight delay in the measurements, intuitively, this delay can be taken into account by keeping four rank-1 terms in the SVD rather than three. More precisely, the SVD method automatically takes this delay into account by adding an extra dimension for the embedding space which is removed in the truncation step in our algorithm. Intuitively that is the reason why the delay is not given exactly as in GTRS method.

Fig. 5.1 (left hand side) illustrates an example of the error curve for EDM-Localization Algorithm. We can see that if the augmented distance vector $\boldsymbol{\Xi}_{\boldsymbol{\pi}}$ in (4.9) for image identification has the correct correspondence, $\boldsymbol{\epsilon}$ can be estimated with reasonable accuracy (cf. Table 4.2). We also observe that for all the permutation except the right one, the error function $\boldsymbol{F}_{\boldsymbol{\pi}}$ has a large value and the rank of the matrix does not change much as depicted in the right hand side of Fig. 5.1; while the condition number of $\widetilde{\boldsymbol{M}}_{\boldsymbol{\pi}}^{\tilde{\boldsymbol{\epsilon}}}$ defined in (4.10) for the right permutation exceeds beyond 500 for noisy measurements, it is less than **25** for a wrong image identification and the measure of error is far less for a correct order. Therefore, the algorithm is able to find the correct order in all scenarios.

4.5.2 Multi-channel Distributed Source Localization

The single-channel estimates can be aggregated to improve the localization performance. To that end, the microphone-source distances are estimated for each microphone and its images individually. The microphones may differ in the synchronization time offset which is estimated and compensated locally. The local distance estimates are used to construct a distance matrix as expressed in (4.24). Consequently, the source location will be updated using either Algorithms 1–3. The performance of the distributed source localization is illustrated in Fig. 4.3 for various noise levels and number of microphones. The results of SSR-LS (Algorithm 3) are very close to the EDM-Localization and they are not further illustrated. The results are repeated for 100 random configurations and averaged over 400 realizations at each noise level. The error bars correspond to 95% confidence interval.

We observe that exploiting additional microphones improves the source localization performance and noise robustness significantly. Furthermore, the performance gap between SVD-Localization and EDM-Localization is reduced as the number of microphones is increased. Indeed, we empirically observe that for more than ten microphones, the algorithms perform very close to each other.

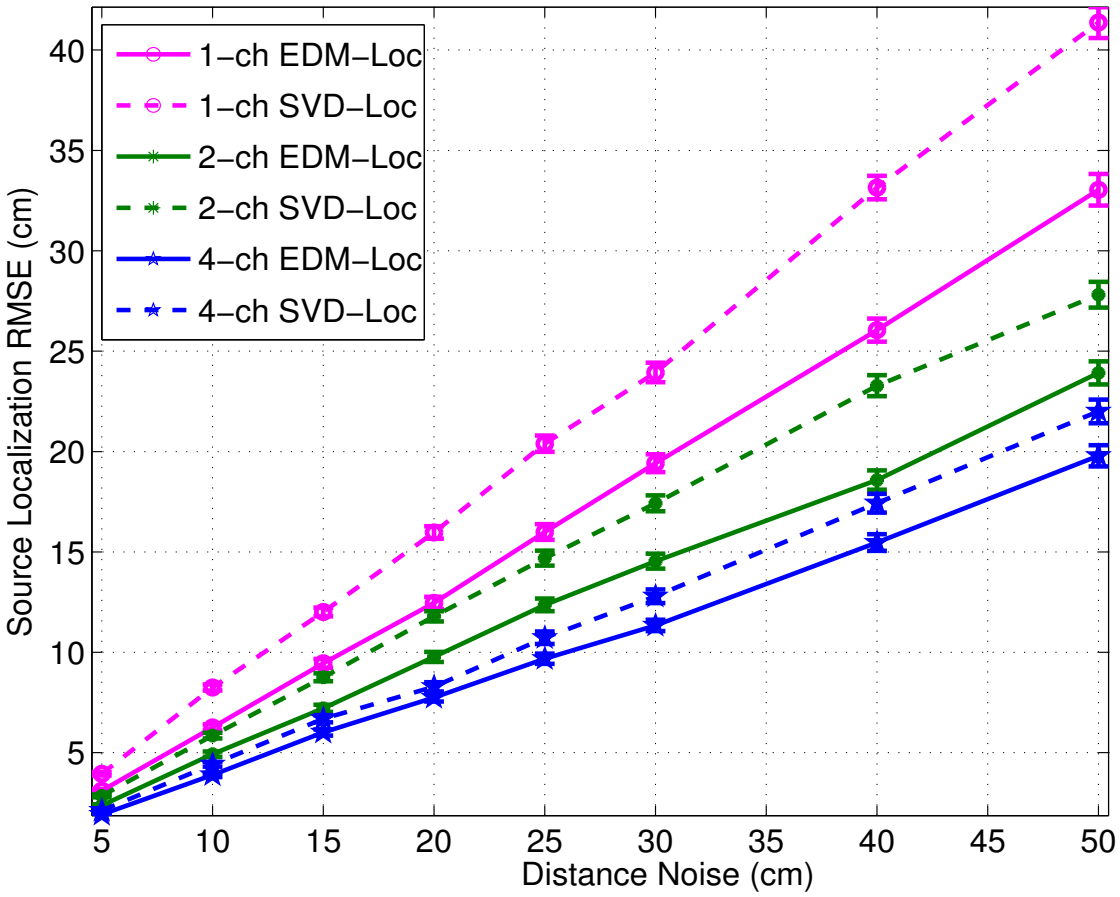

Figure 4.3: Distributed source localization using aggregation of single microphone measurements. The error bars correspond to 95% confidence interval.

4.5.3 Real Data Evaluation

To conduct the real data evaluation, we use the speech recordings performed in the framework of the Multichannel Overlapping Numbers Corpus (MONC) [1] (Section 2.6.1). This database was collected by outputting utterances from Numbers Corpus release 1.0 on a loudspeaker, and recording the resulting sound field using a microphone array [1] at sampling rate of 8 kHz. The recordings were made in a **8.2** m × **3.6** m × **2.4** m rectangular room containing a centrally located **4.8** m × **1.2** m rectangular table. The loudspeaker was positioned at **1.2** m distance from the center of table at an elevation of **35** cm (distance from table surface to center of loudspeaker). An eight-channel, **20** cm diameter, circular microphone array was placed in the center of the table recorded the mixtures. The average signal to noise ratio (SNR) of the recordings was about **10** dB. The room is mildly reverberant with a reverberation time about 250 ms.

We estimate the support of the room impulse response (RIR) function using the blind channel identification approach based on sparse cross-relation formulation [74, 3, 8]. The sparse RIR model is theoretically sound [4], and it has been shown to be useful for estimating real impulse

responses in acoustic environments [11]. This approach can provide an accurate estimation of the acoustic channel up to a time delay and scaling factor. As we only need the support of the early part of the impulse response and the proposed approach can effectively handle the issue of asynchronous recording time offset, the resulting RIR is suitable to evaluate our method.

To employ the sparse cross-relation RIR estimation technique, two microphone recordings are required. Hence, the two microphones in line with the speaker ([1]) are selected. In addition to the sparsity constraint, a positivity constraint is considered to yield more accurate early support estimation [8]. The regularization parameter using the algorithm published in [74] is set to 0.3 and the CVX software package is used for optimization [57]. The length of the impulse response is set to 150.

The results of the RIR estimation for one microphone are depicted in Fig. 4.4. The reflections are extracted by setting a threshold of 0.05 (with respect to the direct path) on the amplitude of the room impulse response. The spikes greater than this threshold define the initial support of the impulse response associated with the principal reflectors; their indices are used for distance calculation in (4.4). As we can see, the support is overestimated, i.e. $\mathscr{R} > \boldsymbol{R}$. A single reflection (corresponding to the wall at distance **8.2** m) can not be captured in this range and thus computed based on the hypothesized correspondence at each permutation. The heuristics as such are helpful to speed-up the support recovery procedure and there is no algorithmic impediment to consider longer filters and drop the duality between the pairs of the spikes (due to parallel walls). The first spike is associated to the direct path, thus assumed to be fixed and all the $\binom{\mathscr{R}-\mathbf{1}}{\boldsymbol{R}-\mathbf{1}}$ combinations of the support are tested. Based on the recovered support, the joint synchronization and source localization procedure estimates the source position with **5** cm error. If the measurements of two microphones are aggregated as described in Section 4.4, the error reduces to **3** cm.

Furthermore, we conducted experiments in more complex acoustics using the data collected at the Laboratory of Electromagnetics and Acoustics (LEMA) at École polytechnique fédérale de Lausanne (EPFL). The psychoacoustic room is considered for data collection. The dimension of the room is $\mathbf{6.6} \times \mathbf{6.86} \times \mathbf{2.69}$ m$^{\mathbf{3}}$ and it is fully equipped with furniture such as shelves, boxes of different textures, distributed tables and chairs. The reverberation time is about 350 ms. The source is located at $\boldsymbol{z} = [\mathbf{2.69\ 1.2\ 0.97}]^\top$ with respect to the origin at the door corner. The source signal is a white Gaussian noise sampled at the rate of 51200 Hz. It is down sampled to 8000 Hz for room impulse response estimation to reduce the computational cost. If we use the channel response at a microphone located at $[\mathbf{2.22\ 4.11\ 0.95}]^\top$, the source localization error is **6** cm. Using an additional microphone located at $[\mathbf{1.43\ 2.71\ 1.47}]^\top$, the error is reduced to **3.5** cm.

The proposed image identification exploiting the EDM properties is robust to noise and channel order estimation and it can further be utilized to enhance the estimation of the impulse response function [8].

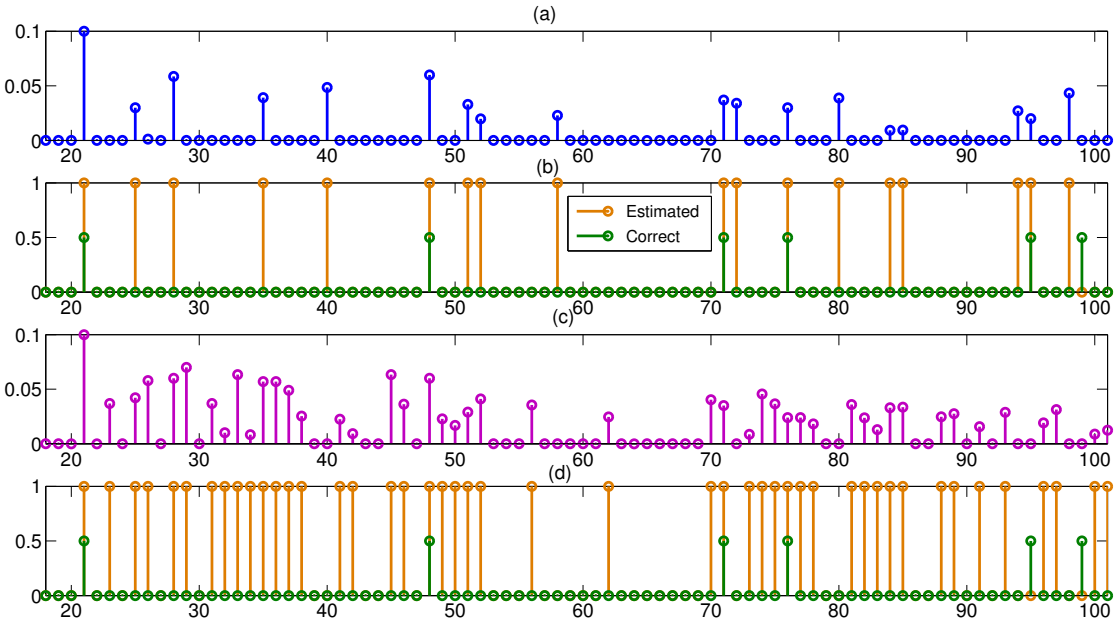

Figure 4.4: (a) Sparse cross-relation based estimation of the early reflections in the room impulse response. (b) Support of early reflections: solid lines depicts the estimated support and the dashed lines illustrates the true support based on the ground truth source location information. (c) Conventional cross-relation estimation of early reflections [129] and (d) Support of estimated and true early reflections: solid lines depicts the estimated support and the dashed lines illustrates the true support based on the ground truth source location information. Based on the estimated support of the early reflections in the room impulse response depicted in (b), the source position is estimated with **5** cm error.

4.6 Conclusions

In this chapter, a novel single-channel source localization approach was proposed applicable to distributed localization scenarios. The image microphone model of multipath propagation was employed to resolve the ambiguities in spatial information recovery. A multipath distance matrix was constructed where the components corresponding to the distance between the actual and virtual microphones were known. The support of the spikes in the room impulse response function indicates the distances between the unknown source location and microphones up to indeterminacies in identifying the correspondence to each image microphone along with a synchronization delay. The properties of the multipath Euclidean distance matrix were exploited to resolve these ambiguities and novel algorithms were proposed to synchronize the recordings and localize the source. In particular, an estimation strategy was derived based on globally optimizing the synchronization extension of squared-range-based least square cost function. The experiments conducted on various simulated and real data recordings of noisy scenarios demonstrated that the proposed approach is robust to jitter noise in the support of spikes in the room impulse response as well as the asynchronous time offsets in recordings. Indeed, it was shown that the synchronization

delay can be estimated with reasonable accuracy and compensated for source localization. Furthermore, aggregation of multi-microphone estimates was elaborated and shown to be effective to improve the source localization performance.

Appendix 1. Proof of Proposition 1.

If $J_{\mathrm{PD}} = \emptyset$, the argument trivially holds. Hence, let us assume that $J_{\mathrm{PD}} \neq \emptyset$. First, we prove that J_{PD} is a convex set which implies that J_{PD} must be an interval. Assume that $\lambda_1, \lambda_2 \in J_{\mathrm{PD}}$ and let $G_i = A^\top A + \lambda_i L$, $i = 1,2$. Notice that for any $u \in \mathbb{R}^{\kappa+2}, u \neq 0$, one has $u^\top G_i u > 0$, which implies that for any $\alpha \in [0,1]$, $u^\top(\alpha G_1 + (1-\alpha)G_2)u > 0$. Thus $\alpha G_1 + (1-\alpha)G_2 \succ 0$. As

$$\alpha G_1 + (1-\alpha)G_2 = A^\top A + (\alpha\lambda_1 + (1-\alpha)\lambda_2)L,$$

it follows that for any $\alpha \in [0,1]$, $\alpha\lambda_1 + (1-\alpha)\lambda_2 \in J_{\mathrm{PD}}$. This proves the convexity of J_{PD}.

To prove the openness of the interval J_{PD}, let $\lambda \in J_{\mathrm{PD}}$ be an arbitrary point and let $G = A^\top A + \lambda L \succ 0$. Consider the function $g : \{u : \|u\| = 1\} \to \mathbb{R}$ defined on the unit ball by $g(u) = u^\top G u$. Notice that g is a strictly positive function since $G \succ 0$. Therefore, it achieves its minimum value g^* on the compact set $\{u : \|u\| = 1\}$ where $g^* > 0$. As all the eigen-values of L consist of $\{\pm 1, 0\}$, one can simply check that $G + \mu L \succ 0$ for all $\mu \in (-\frac{g^*}{2}, \frac{g^*}{2})$. In particular, this implies that for all γ in the open interval $(\lambda - \frac{g^*}{2}, \lambda + \frac{g^*}{2})$ containing λ, $A^\top A + \gamma L \succ 0$. This shows that J_{PD} is an open interval. ■

5 Robust Microphone Placement for Source Localization from Noisy Distance Measurements

In this chapter a novel algorithm to design an optimum array geometry for source localization inside an enclosure is proposed. We assume a square-law decay propagation model for the sound acquisition so that the additive noise on the measured source-microphone distances is proportional to the distances regardless of the noise distribution. We formulate the source localization as an instance of the "Generalized Trust Region Subproblem" (GTRS) the solution of which gives the location of the source. We show that by suitable selection of the microphone locations, one can tremendously decrease the noise-sensitivity of the resulting solution. In particular, by minimizing the noise-sensitivity of the source location in terms of sensor positions, we find the optimal noise-robust array geometry for the enclosure. Simulation results are provided to show the efficiency of the proposed algorithm.

5.1 Introduction

The optimum microphone array placement is a fundamental design problem that seeks the best spatial positions of the microphones such that a certain performance measure in terms of energy or cost efficiency, estimation, detection or identification accuracy is guaranteed. The focus of this chapter is the optimum microphone array geometry for source localization based on noisy observations of the source-microphone distances.

The prior art often formulates the sensor placement problem for linear measurement models and the optimization procedures are derived for a scalar cost related to the mean squared error covariance matrix. It is also referred to as an optimal experimental design problem [20] in which a grid of sensors at all locations is hypothesized and the best subset of $\boldsymbol{M}$ sensors out of $\boldsymbol{G}$ possible locations is selected where $\boldsymbol{M}$ is typically known [90, 59, 66]. This formalism

leads to a non-convex Boolean optimization problem which incurs a combinatorial search over all the $\binom{G}{M}$ possible combinations. In [65] a convex relaxation technique is presented for additive Gaussian linear models and the performance measures are independent of the unknown parameter. Alternative to the convex optimization, the selection is achieved based on the coherence of the sensor measurements [93, 30] or solved using greedy algorithms [105] or heuristics.

Non-linear measurement models are frequently encountered in applications like source localization and tracking. The error covariance matrix for non-linear models is not always available in closed form, and it often depends on the unknown parameter, hence, alternative approaches or performance measures are considered. A sensor selection algorithm for observations related to non-linear models is proposed in [64] within the Bayesian and sequential design. In [76], sensor selection for target tracking based on extended Kalman filtering is developed, in which a selection is performed by designing an appropriate gain matrix for a non-linear measurement model in additive Gaussian noise; the error covariance matrix is computed from the past state estimates so the solution is suboptimal. An alternative sensor selection framework is proposed in [33] where a sparse selection vector is designed such that a certain Cramér-Rao bound optimality on the estimates is guaranteed. This framework enables optimization over the number of microphones as a cardinality minimization problem such that a specified performance bound on localization error is obtained. The optimization procedure relies on the minimum eigen value of the Fisher information matrix and the optimal source localization can not always be achieved.

In this chapter, we consider a minimax approach to design the microphone array without making any assumptions on the source location or statistics of the measurement noise. The main idea is to find a geometry for the array that gives the minimum estimation error for the source location when the source location and noise values are selected adversarily. We use an optimization approach based on Generalized Trust Region Subproblem (GTRS) to design a function whose minimum gives the minimax optimal geometry. We show for the rectangular enclosure, one can find the optimal solution efficiently.

5.2 Problem Statement

5.2.1 Signal Model

We consider a simple scenario for source localization in a rectangular-shaped room with $\boldsymbol{M}$ microphones whose positions are denoted by $\boldsymbol{x_i}$, $\boldsymbol{i} \in [\boldsymbol{M}]$. In our case, we consider a very simple case where $\boldsymbol{M} = \mathbf{4}$. The results can be extended to more general cases. Let $\boldsymbol{s}$ be the location of the source in the room. To find the location of the source, each microphone estimates its distance from the source denoted by $\boldsymbol{d_i}, \boldsymbol{i} \in [\boldsymbol{M}]$. We suppose $\boldsymbol{d_i} = \|\boldsymbol{x_i} - \boldsymbol{s}\|(\mathbf{1} + \boldsymbol{\eta_i})$ where $\boldsymbol{\eta_i}$ is the relative measurement noise. We do not assume any specific distribution for $\boldsymbol{\eta_i}$ except that $\boldsymbol{\eta_i} \in [-\boldsymbol{\delta_i}, \boldsymbol{\delta_i}]$ where $\boldsymbol{\delta_i} \in [\mathbf{0}, \mathbf{1})$ is a fixed given number showing the amount of

noise in measurements of each microphone.

We briefly explain why this is a good model for the measurements. If we assume an square-law propagation model for the sound wave, it immediately results that the received signal power in each microphone is proportional to the inverse-square of its distance from the source. If we consider an algorithm to estimate the distance from the input signal, the estimation variance will be proportional to the squared-distance from the source. Therefore, the standard variation of the noise is proportional to distance. We model this by $\boldsymbol{\eta_i}$, i.e., $\boldsymbol{d_i} = \|\boldsymbol{x_i} - \boldsymbol{s}\| + \boldsymbol{\eta_i}\|\boldsymbol{x_i} - \boldsymbol{s}\|$, where $\boldsymbol{\eta_i}$ denotes the resulting noise.

5.2.2 Algorithm for Source Localization

To recover the position of the source, we consider the following quartic cost function:

$$\boldsymbol{g(z)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{M}} (\|\boldsymbol{z} - \boldsymbol{x_i}\|^2 - \boldsymbol{d_i^2})^2. \tag{5.1}$$

The optimal source location is recovered by finding the global minima of the cost function. Adding an auxiliary variable $\boldsymbol{\gamma}$ and defining $\boldsymbol{y}^T = (\boldsymbol{z}^T, \boldsymbol{\gamma})$ and the following matrices:

$$\boldsymbol{A} = \begin{pmatrix} -2\boldsymbol{x}_1^T & \mathbf{1} \\ -2\boldsymbol{x}_2^T & \mathbf{1} \\ \vdots & \vdots \\ -2\boldsymbol{x}_M^T & \mathbf{1} \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} \boldsymbol{d}_1^2 - \|\boldsymbol{x}_1\|^2 \\ \vdots \\ \boldsymbol{d}_M^2 - \|\boldsymbol{x}_M\|^2 \end{pmatrix}, \tag{5.2}$$

the minimization problem can be written as follows

$$\min_{\boldsymbol{y}} \left\{ \|\boldsymbol{Ay} - \boldsymbol{b}\|^2 : \boldsymbol{y}^T \boldsymbol{L}\, \boldsymbol{y} = \mathbf{0} \right\}, \tag{5.3}$$

where $\boldsymbol{L} = \text{diag}(\mathbf{1}_{\boldsymbol{\kappa}\times\mathbf{1}}, -\mathbf{1})$ is a diagonal matrix and $\boldsymbol{\kappa} = \mathbf{3}$ (in general $\boldsymbol{\kappa}$ is the dimension of the ambient Euclidean space containing the microphones). This an special instant of a quadratic optimization under a single quadratic constraint known as "Generalized Trust Region Subproblem" (GTRS) whose global minimum can be efficiently computed. Specifically, we have the following theorem: Specially, by [83] and [12], $\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{\kappa}+\mathbf{1}}$ is an optimal solution of (5.3) if and only if there is a $\boldsymbol{\lambda} \in \mathbb{R}$ such that

$$(\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\lambda}\boldsymbol{L})\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{b},\ \boldsymbol{y}^T\boldsymbol{L}\,\boldsymbol{y} = \mathbf{0}, \tag{5.4}$$

$$\boldsymbol{A}^T\boldsymbol{A} + \boldsymbol{\lambda}\boldsymbol{L} \succeq \mathbf{0}. \tag{5.5}$$

This system of equations can be efficiently solved for $\boldsymbol{y}$ and $\boldsymbol{\lambda}$ which in particular gives $\boldsymbol{z}$ the optimal position of the source.

5.2.3 Noisy Measurements and Minimax Design

If there are measurement noises, the d_i parameters in the vector b in Equation (5.2) will be the noisy distances, thus the Equations(5.4) and (5.5) give an estimate of the source location. The estimation precision highly depends on on the geometry of the microphone array (matrix A), the real location of the source s and measurement noises η_i. In some applications, one might have good estimates of the statistics of the noise or mobility pattern of the source inside the room specially if the source location is repeatedly estimated during time. In that case, one might design the sensor array based on these prior information.

In this chapter, we consider a minimax approach to design the microphone array. More precisely, without making any assumptions on the initial source location or statistics of the noise, we use an optimization approach based on the Equations (5.1), (5.4) and (5.5), to find the minimax-optimal configuration for the location of the microphones. The main idea is to find a geometry for the array that gives the minimum estimation error for the source location when the source location and noise values are selected in an adversary manner. We design a function whose minimum gives the minimax optimal geometry. Although optimizing this objective function might be difficult (requiring an exhaustive search) for a general room shape, in most cases the symmetry of the problem helps to find the optimal solution efficiently. In this section, we derive the results for a simple rectangular enclosure. Let us define the following function:

$$g(z;X,s,\eta)=\sum_{i=1}^{M}(\|z-x_i\|^2-\|x_i-s\|^2(1+\eta_i)^2)^2, \tag{5.6}$$

where $X=[x_1,\dots,x_M]$ is a matrix consisting of all microphone locations, $\eta=[\eta_1,\dots,\eta_M]$ is the set of all noises and s is the real location of the source. Notice that $g(z;X,s,\eta)$ is the same function as Equation (5.1) which should be minimized with respect to z to find the location of the source where the structure of the function g and thus the estimation quality of the source location depends on (X,s,η). Also, as a design parameter we can choose X but (s,η) might be revealed adversarially as far as s is inside the enclosure and $|\eta_i|\leq\delta_i$. Defining $u=z-s$, one can write g as follows

$$g(u;X,s,\eta)=\sum_{i=1}^{M}(\|u\|^2-2(x_i-s)^Tu-\|x_i-s\|^2t(\eta_i))^2, \tag{5.7}$$

where $t(\eta_i)=2\eta_i+\eta_i^2$. In practice, $|\eta_i|\ll 1$ and for simplicity one can assume that $t(\eta_i)\approx 2\eta_i$ where $\eta_i\in[-\delta_i,\delta_i]$. In order to find the source location, one should find the minimum of the function $g(u;X,s,\eta)$ with respect to u with the only difference that the optimization region is a rectangle centered at $-s$ rather than 0. To simplify the design we assume that the optimization of the function $g(u;X,s,\eta)$ is done over all $\mathbb{R}^2$. Notice that this is still a worst-case assumption because, for example, if in the unrestricted case, the minimum of the function occurs out of the room boundary, one can always find a better estimate of the source location inside the room. Let us denote by $\hat{u}(X,s,\eta)$ the minimum of the function $g(u;X,s,\eta)$

and let us define $e(X,s,\eta) = \|\hat{u}\|^2$ to be the error (variance) of the estimation. It is immediate to check that if there is no noise $\eta = 0$, the minimum point will be $\hat{u} = 0$, i. e., $e(X,s,0) = 0$ and the source location is exactly identified. For the minimax design that we consider, we are interested in $e_{opt} = \min_X \max_{s,\eta} e(X,s,\eta)$ where the maximization over η_i is done over the range $\eta_i \in [-\delta_i, \delta_i]$ and the outer minimization gives the minimax-optimal array geometry that we are interested in.

To further analyze the problem, we convert it into an instance of GTRS problem. Let us define $\gamma = \|u\|^2$, $y = (u^T, \gamma)$ and

$$A = \begin{pmatrix} -2(x_1 - s)^T & 1 \\ -2(x_2 - s)^T & 1 \\ \vdots & \vdots \\ -2(x_M - s)^T & 1 \end{pmatrix}, b = \begin{pmatrix} 2\eta_1 \|x_1 - s\|^2 \\ \vdots \\ 2\eta_M \|x_M - s\|^2 \end{pmatrix}. \tag{5.8}$$

Then, one can formulate finding the optimal $\hat{u}$ as in Equation (5.3) where the global minimum is given by Equations (5.4),(5.5).

Let us define the center of mass and the covariance of the array geometry by

$$\mu = \frac{1}{M} \sum_{i=1}^{M} x_i, \ \Sigma = \frac{1}{M} \sum_{i=1}^{M} (x_i - \mu)(x_i - \mu)^T. \tag{5.9}$$

Choosing a coordinate system with axes parallel to the edges of the room with the origin at the center of the room, from the symmetry of the problem (room shape and symmetry of η), it results that the minimax optimal array geometry must be symmetric with respect to horizontal and vertical axes (in particular, $\mu = 0$). Therefore, there are two types of configurations that we should consider:

- ◇ C1: microphones are located at the vertices of a rectangle with edges parallel to the walls of the room.
- ◇ C2: microphones are on the vertices of a rhombus with diagonals parallel to the walls.

Also, after some simplification, Equations (5.4),(5.5) for this case can be written as follows:

$$\begin{pmatrix} 4M(\Sigma + ss^T) + \lambda I_\kappa & 2Ms \\ 2Ms^T & M - \lambda \end{pmatrix} y = w, \ y^T L y = 0, \tag{5.10}$$

where

$$w = \begin{pmatrix} -4\sum_{i=1}^{M} \eta_i \|x_i - s\|^2 (x_i - s) \\ 2\sum_{i=1}^{M} \eta_i \|x_i - s\|^2 \end{pmatrix}$$

and I_κ is the identity matrix of order κ where κ is the ambient dimension of the microphones

Array Configuration	Relative noise standard deviation					
	0.01	**0.02**	**0.05**	**0.1**	**0.2**	**0.3**
Average-optimum	3.59	7.16	17.9	35.67	73.93	115.4
Robust	4.1	8.56	19.6	37.4	77.2	122.84
Compact	35.33	83.98	170.6	200.2	209.34	215.89
Random	10.44	14.8	35.76	61.57	111.7	173.7
Corner	6.04	12.24	30.1	60.8	98.74	152.33

Table 1: Localization error (cm) using different microphone placements at various Gaussian noise with a relative standard deviation $\boldsymbol{\delta_i}$.

(We take $\boldsymbol{\kappa = 2}$). There are still parameters $\boldsymbol{(s,\eta)}$ in these equation. We prove the following proposition which specifies the worst-case source location in the minimax design.

Proposition 4. Let $\hat{\boldsymbol{X}}$ be the minimax array configuration, i.e.,

$$\max_{s,\eta} e(\hat{X}, s, \eta) = \min_{X} \max_{s,\eta} e(\hat{X}, s, \eta). \tag{5.11}$$

Let $(\hat{s}, \hat{\eta})$ be the worst source location and noise parameter, i.e., $e(\hat{X}, \hat{s}, \hat{\eta}) = \max_{s,\eta} e(\hat{X}, s, \eta)$. Then $\hat{s}$ must be on the vertices of the rectangular enclosure.

Proof. We just provide a sketch of the proof. The main idea is that for both types of configurations C1 and C2, one can increase $\|x_i - s\|^2$ by moving s closer to the vertices of the enclosure. As the noise scales proportional to the distance, this is equivalent to increasing the noise parameter δ_i which can potentially give a larger value in the minimax term $\max_{s,\eta} e(\hat{X}, s, \eta)$. This implies that the worst source location must be on the vertices. □

Conjecture 1. In the minimax design, the worst case for the noise parameter η is when η_i is either $+\delta_i$ or $-\delta_i$.

Proposition 4 and Conjecture 1, completely specify the worst (s,η) parameter. To find the minimax optimal geometry, one only needs to do a simple optimization over all symmetric configurations of type C1 and C2 by simply solving Equation (5.10) for y and λ. Notice that the last component of y is a positive number corresponding to the resulting estimation error $\|u\|^2$ and the minimax configuration is the one minimizing this component.

5.3 Experimental Results

In this section, we conduct some experiments to demonstrate the proposed theories. The goal is to find the minimax optimal microphone array configuration which minimizes the localization error for the worst source location and the worst noise distribution. We consider a rectangular enclosure of dimension $\mathbf{6.6 \times 3.6}$ m^2. The number of microphones is $M = 4$.

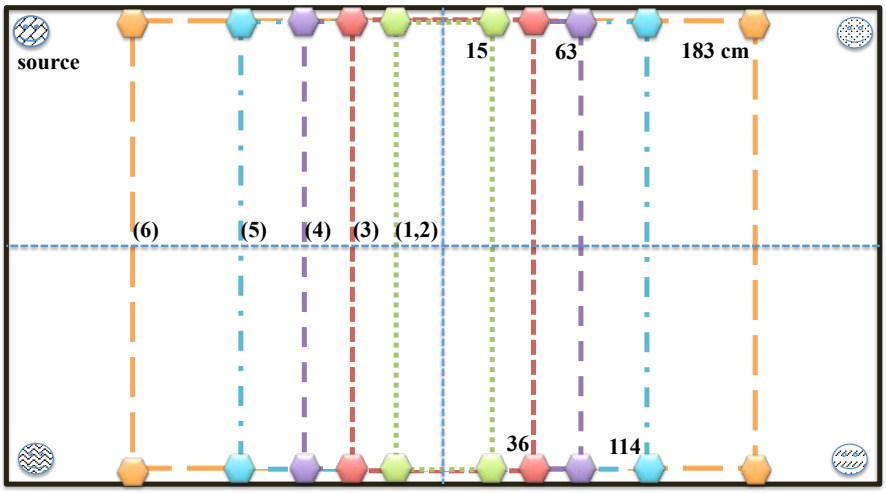

Fig. 1: Robust microphone configurations for source localization using four microphones: The numbers show the placement of the microphones (cm) along the x-axis with respect to the origin located at the room center. The worst-case source location is at the corners of the enclosure depicted by hashed circles. The configurations (1)–(6) correspond to $\boldsymbol{\delta_i} = \{\mathbf{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}\}$; for example if $\boldsymbol{\delta_i} = \mathbf{0.1}$ the purple configuration (4) is obtained by solving Equation (5.10). We can see that larger noise levels lead to the microphone placements closer to the corners to achieve a robust design.

5.3.1 Robust Microphone Array Configuration

As we explained in Section 5.2.3, in the minimax design, we assume a worst case scenario for the source mobility inside the enclosure and the distribution of noise. Based on Proposition 4, the worst-case source position is when it is located in one of the corners of the rectangular enclosure which has been depicted by hashed circles in Fig. 1. To find the minimax-optimal array geometry, we run a simple (one-dimensional) optimization expressed in (5.10) over all symmetric configurations of type C1 and C2.

Fig. 1 illustrates the six configurations (1)–(6) obtained for different noise levels on the source-microphone distances, i.e. $\boldsymbol{\delta_i} = \{\mathbf{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}\}$ respectively. The resulting position of microphones for different noise levels are depicted with hexagonal shape and with different colors. Notice that between two types of array configurations (C1 and C2), the optimal one is always of type C1. The difference between the microphone positions at the green configurations (1) and (2) (corresponding to $\boldsymbol{\delta_i} = \mathbf{0.01, 0.02}$) is less than 5 cm so they are not distinguishable in the picture. One can observe that as the level of noise increases, the microphones positioned on rectangles move away from the y-axis towards the corners. The exact positions are at 15, 36, 63, 114 and 183 cm distance from the y-axis corresponding to the different values of $\boldsymbol{\delta_i}$ as stated above.

The robust configurations obtained in this section do not exploit any prior knowledge on the source position or the noise distribution. In the next section, we assume that both the source

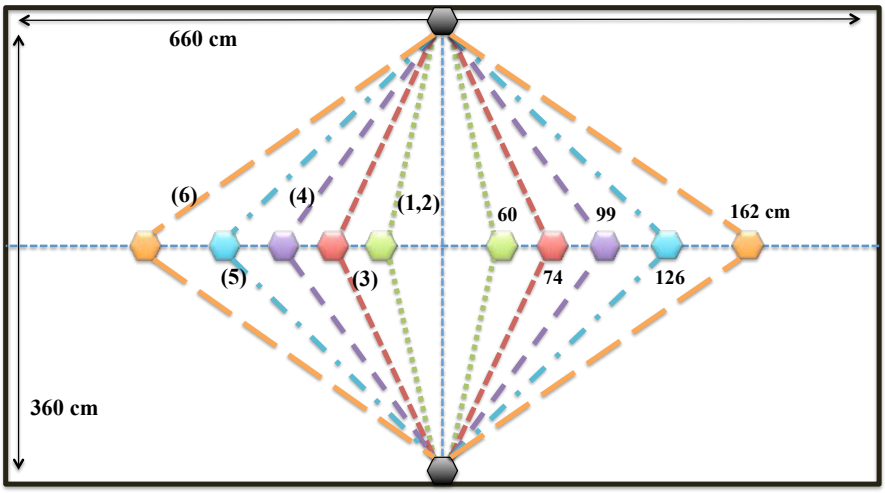

Fig. 2: Average-optimum microphone configurations for source localization using four microphones. The numbers show the placement of the microphones (cm) with respect to the origin located at the room center. Number (1)-(6) corresponding to $\boldsymbol{\delta_i} = \{\mathbf{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}\}$ accordingly. One can see that larger noise levels lead to larger apertures.

mobility and the noise distribution are known and we find the average-optimal configuration using an exhaustive search over all possible microphones placements to find the configuration corresponding to the minimum average source localization error where the average is taken over the source location and the noise statistics.

5.3.2 Comparison with an Average-optimal Array Geometry

In the minimax-optimal design, the philosophy is to guard against the worst source location and measurement noise. This is a reasonable assumption if one does not have any prior knowledge about the mobility pattern of the source or the statistics of the noise. In some cases, it might be possible to know both the mobility of the source and the statistics of the noise, thus it would be possible to find an average-optimal array geometry where the average is taken over the distribution of the source and measurement noise. It will be interesting to know how the minimax-optimal design compares with this average-optimal design. In this section, we assume that the source is uniformly distributed inside the enclosure and the measurement noise is Gaussian. To find the average-optimum array configuration, the area of the room is discretized into a grid of 600 (uniform) cells. All of the $\binom{600}{4}$ array configurations are considered while the source is uniformly randomly sampled inside the room to quantify the average localization errors. The simulated noise on the source-microphone distances is Gaussian with the relative standard deviation $\boldsymbol{\delta_i} = \{\mathbf{0.01, 0.02, 0.05, 0.1, 0.2, 0.3}\}$. For each source and array configuration, 50 realizations of the noise are considered and the average localization error for each configuration is quantified. The resulting average-optimum microphone placement is depicted in Fig. 2 where for each value of $\boldsymbol{\delta_i}$ the configurations (1)–(6) are obtained accordingly.

We can see that as the level of noise increases, two of the microphones (black ones) remain at fixed positions on the walls while the other two microphones positioned at the middle move towards the walls to minimize the mean (expected) source localization error. The exact placement of the middle microphones with respect to the origin of the room coordinates (room center) is indicated at the pictures; we can see that the positions move from 60 cm to 162 cm as the noise level is increased.

The first two rows of Table 1 compare the performance of the average-optimal and minimax design for this scenario. It is seen that minimax-optimal design performs very well (comparable with the optimal one).

5.3.3 Source Localization Performance

In this section, we evaluate the performance of source localization using different microphone array configurations. Table 5.1 lists the localization error for different microphone array design at different Gaussian noise levels. Each number is obtained by averaging over 600 arbitrary positions of the source where the noisy distances are given from 50 realizations. The first row corresponds to the average-optimum placement. The second row indicates the error for the robust configuration obtained through the proposed algorithm. We can see that although the robust configuration is different than the average-optimum one, the expected localization performance is very close to the optimal value at all noise levels. Hence, the proposed algorithm enables an efficient microphone array design to achieve robust source localization. Further empirical observations show that the localization error using the robust configuration is 40% less than the average-optimum configuration if the source is located at the enclosure corners (the worst-case scenario). It may be noted the the average-optimum configuration is obtained under the assumptions that the noise distribution is known (Gaussian) and the source mobility is uniform. If these assumptions are violated, it leads to the degradation of the performance obtained from the average-optimum configuration. On the other hand, the robust configuration is achieved without any assumption on the noise distribution and source mobility. Hence, the performance can be generalized to other setups.

The third row of Table 1 presents the error if a compact circular microphone array of diameter 20 cm is used at the center of the room. We can see that using a compact microphone for localization leads to huge error which is up to 8 times bigger than the localization accuracy achieved using the robust design. This error increases quickly by increasing the noise on the distances. In addition, we evaluate the localization error if the microphones are positioned randomly. For this experiment, we choose 20 random setups and compute the average localization error for 50 realizations of the noisy distances. We can see that the localization error is about two times more than the robust configuration. Finally, the last row shows the localization error when the microphones are positioned at the corner of the room.

5.4 Conclusions

In this chapter, we proposed a minimax design for a microphone array consisting of four microphones in a rectangular enclosure. We assumed a square-law decay propagation model for the sound and designed the array for the worst source location and statistics of the measurement noise. We proposed an efficient algorithm to identify the robust microphone array configuration to minimize the worst-case source localization error. We showed that this robust configuration yields the performance very close to the average-optimum design. The robust placement was also shown to achieve substantial improvement over the compact, ad hoc and heuristic microphone array configurations. As an extension, one can consider a more complicated signal model for the source consisting of reflections from the boundaries which can be characterized using the image-source model of multipath propagation.

6 An Integrated Framework for Multi-Channel Multi-Source Localization and Voice Activity Detection

Two of the major challenges in microphone array based adaptive beamforming, speech enhancement and distant speech recognition, are robust and accurate source localization and voice activity detection. This chapter introduces a spatial gradient steered response power using the phase transform (SRP-PHAT) method which is capable of localization of competing speakers in overlapping conditions. We further investigate the behavior of the SRP function and characterize theoretically a fixed point in its search space for the diffuse noise field. We call this fixed point the *null* position in the SRP search space. Building on this evidence, we propose a technique for multi-channel voice activity detection (MVAD) based on detection of a maximum power corresponding to the *null* position. The gradient SRP-PHAT in tandem with the MVAD form an integrated framework of multi-source localization and voice activity detection. The experiments carried out on real data recordings show that this framework is very effective in practical applications of hands-free communication[1].

6.1 Introduction

Speaker localization is a demanding area of research in hands-free speech communication, virtual reality, and smart environment technologies. The previous approaches to deal with this problem are largely confined to multi-channel processing techniques, using microphone arrays. In such applications, accurate knowledge of the speaker location is essential for an effective beampattern steering and interference suppression. This task gets even more challenging in meeting acquisition and conference recordings due to the presence of competing speakers [124]. We will briefly review the main approaches to address this issue as

[1] This chapter has bee published in IEEE workshop on Hands-free Speech Communication and Microphone Arrays [109]

follows:

I. High Resolution Spectral Estimation: Several algorithms have been proposed based on high resolution spectral estimation, such as minimum variance spectral estimation, auto-regressive modeling and various techniques based on eigen-analysis such as Multiple Signal Classification (MUSIC). These approaches are based on analysis of the received signals' covariance matrix, hence need an accurate estimation of the source signals, and impose a stationarity assumption. The underlying hypotheses are hardly realistic in case of speech signals as well as the room acoustics and the results are not very promising [45].

II. Time Difference Of Arrival (TDOA) Estimation: A large body of work exists on variants of multi-channel filtering to estimate the time difference of arrival (TDOA). A common localization approach is based on TDOA estimation of the sources with respect to a pair of sensors. This approach is very practical if the placement of the microphones provides an accurate 3D estimation of the delays.The generalized cross-correlation is typically used where the peak location of the cross-correlation function of the signal of two microphones is mapped to an angular spectrum for direction of arrival estimation [70]. A weighting scheme is often employed to increase the robustness of this approach to noise and multi-path effect. Some commercial applications such as automatic steering of cameras for video-conferences have been developed based on this idea [125]. In such applications, an updating rate of 300ms for location information is possible even in unfavorable acoustic conditions. However, in the scenario of multiple-target tracking and adaptive beam-steering, higher update rate is usually beneficial [22]. The generalized cross correlation (GCC) is the most celebrated technique for TDOA estimation. The basic idea is to find the peak of the cross-correlation function of the signal of two microphones. A weighting scheme is usually applied to increase the robustness of this approach to noise and multi-path effects. The maximum likelihood (ML) weighting is theoretically optimal when there is an uncorrelated noise source and there is no reverberation effect. In practice however, the performance of GCC-ML is highly degraded due to reverberation, and the Phase Transform (PHAT) yields better results [87, 17].
Alternative TDOA estimation approaches are based on room impulse response identification. The basic idea behind this approach is that the acoustic channel defined for each speaker-microphone pair is a function of the speaker location. Hence, identifying the room impulse response enables us to compute TDOAs and localize the speakers. When there is no prior knowledge about the microphone array geometry, this scenario could be formulated as a blind Multiple-Input Multiple-Output (MIMO) channel identification problem. The solution usually incorporates blind source separation at the pre-processing step and resolves the ambiguity of the acoustic mixing process by localization along with the separation of the individual sources [31, 25].
In addition, identification of the speaker-microphone acoustic channel has been incorporated for TDOA estimation and reverberant speech localization [95, 84]. Although TDOA-based techniques are practical and robust, they do not offer a high update rate. Other strategies have thus been sought for multiple-source localization and tracking [22, 73, 52].
Some other alternatives for TDOA estimation are based on singular value decomposition for

estimation of the room impulse response which is very practical for the speech signal but requires at least 250ms of data to converge [63].

III. Beamformer Steered Response Power (SRP): Finally, it is possible to localize the speaker directly based on the beamformer output power. In this approach, the space is scanned by steering the beam-pattern and finding the maximum power. The delay-and-sum beamformer, minimum variance beamformer and generalized side-lobe canceler have been the most effective methods for speaker localization [58, 44, 47, 46, 130]. Unlike TDOA-based approaches, SRP-based localization approaches have a higher effective update rate, i.e., they can work with much shorter frames even in adverse acoustic conditions; hence, they are practically appropriate for realistic applications, especially in multi-party scenarios [6]. In particular, they can be made robust to multipath effect by applying appropriate weighting schemes such as phase-transform [109]. Different filtering proposals have been used in SRP techniques, among which the phase-transform filter (PHAT) has been shown to provide a robust localization framework [44, 109].Considering a set-up of ad hoc microphone array, it is possible to make these techniques robust to some level of asynchronous recording by devising energy-based localization approaches [75].

From a different perspective, a wide range of research endeavours are dedicated to identifying and exploiting the structure underlying the localization problem; examples include subspace and spatial sparsity methods. The subspace methods exploit the rank structure of the received signals' covariance matrix and impose a stationarity assumption to accurately estimate the source location. The important techniques applied include minimum variance spectral estimation and multiple signal classification algorithm [45]. The underlying hypotheses do not generally apply to reverberant sound localization and alternative strategies have usually been considered [5, 71].

Sparse methods in the context of reverberant sound localization have been studied [80, 9, 96]. It has been shown that incorporating spatial sparsity along with the underlying structure of the sparse coefficients enables super resolution in localization of simultaneous sources using very few microphones which is particularly useful when the number of microphone observations is very limited [9].

This chapter is organized as follows: The general concepts of SRP localization approaches are introduced in 6.2.2. We then provide theoretical as well as empirical evidence that the SRP output power for the silent frames exhibits a peak corresponding to a fixed point in its search space. Relying on this observation, we formulate a multi-channel voice activity detection (MVAD) in Section 6.2.3. In Section 6.2.4 a multi-speaker modification of SRP-PHAT for localization of competing sources is proposed by applying a spatial gradient function on the beamformer output. We further carry out some experiments on the real data recordings to evaluate the proposed framework in Section 6.3. Conclusions are drawn in Section 6.4.

6.2 Multi-Source Localization and Voice Activity Detection

6.2.1 Signal model

We consider a scenario in which M microphones record the signal of L sources; the single-channel received signal, $x_m(t)$, is composed of two components: (1) a filtered version of the original signal, s_l, which has been convolved with the source-microphone room impulse response, $h_{m,l}$ and (2) an uncorrelated independent additive noise $n_m(t)$

$$x_m(t) = \sum_{l=1}^{L} s_l(t) * h_{m,l} + n_m(t) \tag{6.1}$$

6.2.2 SRP-PHAT source localization

The general procedure of the beamforming applies filter-and-sum on the input microphone-channels. The filters are usually adapted in order to enhance the source signal whilst suppressing the interference; hence the beamformer output is maximized when the beampattern is focused accurately towards the speaker. In the SRP localization, the output power is used for a 3D scanning of the space where the maximum power corresponds to the location of the active speaker. To state it concisely, the Generalized Cross Correlation (GCC) is defined as

$$R_{m,n}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (G_m(\omega) X_m(\omega)) (G_n(\omega) X_n(\omega))^* e^{j\omega\tau} d\omega \tag{6.2}$$

where X and G are the Fourier transform of the signal and filter, respectively. Defining the weighting function $\Psi_{m,n}(\omega) = G_m(\omega) G_n^*(\omega)$, the GCC function would be

$$R_{m,n}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{m,n}(\omega) X_m(\omega) X_n^*(\omega) e^{j\omega\tau} d\omega \tag{6.3}$$

The PHAT weighting function is defined as

$$\Psi_{m,n}(\omega) = |X_m(\omega) X_n^*(\omega)|^{-1} \tag{6.4}$$

Substituting 6.4 into 6.3 and taking the summation of all possible microphone pairs, the SRP-PHAT is obtained

$$P(\rho, \theta, \varphi) = 2\pi \sum_{m,n} R_{m,n}(\tau_{m,n}) \quad m, n \in \{1, 2, ..., M\} \tag{6.5}$$

where $\tau_{m,n}$ is the time difference of arrival of the source signal located at $\kappa(\rho, \theta, \varphi)$ to the two microphones m and n. Note that the source location is represented in spherical coordinates where ρ denotes the range and θ and φ correspond to the azimuth and elevation, respectively.

The largest peak corresponds to the dominant speaker located at

$$\boldsymbol{\kappa}(\hat{\boldsymbol{\rho}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\varphi}}) = \underset{\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\varphi}}{\mathbf{arg\,max}}\, \boldsymbol{P}(\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\varphi}) \tag{6.6}$$

6.2.3 Multi-Channel VAD

In this section, we will investigate the SRP-PHAT formulation when the input is a diffuse noise, which is often the case in realistic environments without presence of any active speaker. We characterize theoretically the existence of a predefined point for the SRP function for the diffuse noise; hence, there is no speech activity. Suppose $\boldsymbol{\Gamma_{ik}(\omega)}$ is real part of coherence between signals of microphones $\boldsymbol{i}$ and $\boldsymbol{k}$. For the diffuse noise, we have [15]

$$\boldsymbol{\Gamma_{ik}(\omega)} = \mathrm{sinc}\left(\frac{\boldsymbol{\omega d_{ik}}}{\boldsymbol{c}}\right), \tag{6.7}$$

where $\boldsymbol{d_{ik}}$ is the distance between the two microphones and $\boldsymbol{c}$ is the speed of sound. Substituting equation 6.7 into equation 6.3, we obtain

$$\begin{aligned}\boldsymbol{R_{i,k}(\tau)} &= \frac{\mathbf{1}}{\boldsymbol{2\pi}}\int_{-\infty}^{\infty}\frac{\boldsymbol{\Gamma_{ik}(\omega)}}{\boldsymbol{|\Gamma_{ik}(\omega)|}}\boldsymbol{e^{j\omega\tau}}\,\boldsymbol{d\omega} \\ &= \frac{\mathbf{1}}{\boldsymbol{2\pi}}\int_{-\infty}^{\infty}\frac{\mathrm{sinc}(\frac{\boldsymbol{\omega d_{ik}}}{\boldsymbol{c}})}{|\mathrm{sinc}(\frac{\boldsymbol{\omega d_{ik}}}{\boldsymbol{c}})|}\boldsymbol{e^{j\omega\tau}}\,\boldsymbol{d\omega}.\end{aligned} \tag{6.8}$$

In order to find the maximum of the SRP-PHAT function, we compute the derivative w.r.t. $\boldsymbol{\tau}$; hence

$$\frac{\boldsymbol{\partial R_{i,k}(\tau)}}{\boldsymbol{\partial\tau}} = \frac{\boldsymbol{j}}{\boldsymbol{2\pi}}\int_{-\infty}^{\infty}\frac{\boldsymbol{\omega}\mathrm{sinc}(\frac{\boldsymbol{\omega d_{ik}}}{\boldsymbol{c}})}{|\mathrm{sinc}(\frac{\boldsymbol{\omega d_{ik}}}{\boldsymbol{c}})|}\boldsymbol{e^{j\omega\tau}}\,\boldsymbol{d\omega} = \mathbf{0}. \tag{6.9}$$

The above equality holds for $\boldsymbol{\tau} = \mathbf{0}$; hence the maximum of $\boldsymbol{R_{i,k}(\tau)}$ is obtained for a point with equal distance to the two microphones. The same argument is true for all microphone pairs; therefore on the direction perpendicular to the microphone array the closest point to the center of the microphone array is where the output power of the SRP is maximized. Obviously, this has a strong dependence on the elevation and less sensitivity to the azimuth ($\boldsymbol{\theta}$). Since this is obtained only when there is no speech activity, we call it the *null* point of the SRP search space. We can exploit this fact to detect voice activity in the acquired multi-channel speech frames.

The integrated framework of SRP localization and MVAD reduces the complexity of speech analysis in microphone-array applications such as hands-free speech recognition. The previous proposals on MVAD which takes advantage of the extra information provided by additional sensors [49, 131] increases the computational load. A few others have been also

published recently based on Gaussianity assumption of the frequency components [91] or non-uniform phase assumption [69]. In practice however, these hypotheses are not realistic. On the other hand, the technique that we propose here is based on a realistic model of the acoustic conditions as a diffuse noise field and imposes no computational load on the source localization and beamforming designed for data acquisition. Moreover we don't need any training or threshold optimization which is a common computational load in any VAD structure.

6.2.4 Spatial Gradient SRP-PHAT

In this section, we further exploit the integrated frame work of SRP localization and MVAD, and extend it for multi-party scenarios. The PHAT transform whitens the microphone signals; hence yields sharper peaks at the output power corresponding to the actual location of the L sources. In multi-speaker scenarios, the localization of L competing sources amounts to the detection of the largest L peaks of the beamformer output power. In practice however, the SRP-PHAT output has many local maxima due to the multi-path effect which make the extraction of the largest L peaks very difficult. Considering the fact that the SRP has a discrete search space, we first apply a three dimentional box filtering (averaging) defined as follows:

$$\bar{P}(\rho_i,\theta_i,\varphi_i) = \frac{\sum_{c=-1}^{1}\sum_{b=-1}^{1}\sum_{a=-1}^{1} P(\rho_{i-c},\theta_{i-b},\varphi_{i-a})}{27} \tag{6.10}$$

To find the second source location, we have to remove the data corresponding to the dominant speaker from the search space. Therefore, the data points from all directions of ρ,θ,φ which correspond to the negative spatial gradient of the SRP output power ($\bar{P}$) are discarded. The directional derivative of $\bar{P}$ at point κ in direction u is obtained by

$$\nabla_u \bar{P}(\kappa) = \lim_{h\to 0^+} \frac{\bar{P}(\kappa+hu)-\bar{P}(\kappa)}{h} = \nabla\bar{P}(\kappa)\,u, \tag{6.11}$$

where u is the unit vector and ∇ on the right denotes the gradient and

$$\nabla\bar{P}(\rho,\theta,\varphi) = \frac{\partial\bar{P}}{\partial\rho}e_\rho + \frac{1}{\rho}\frac{\partial\bar{P}}{\partial\varphi}e_\varphi + \frac{1}{\rho\sin\varphi}\frac{\partial\bar{P}}{\partial\theta}e_\theta, \tag{6.12}$$

where $e_\rho, e_\theta, e_\varphi$ are the canonical basis vectors of the coordinate system. Then, the directional derivative defined in equation 6.11 is computed at the location of the largest peak denoted by κ in 26 u directions. Hence,

$$u \in \left\{ \frac{i e_\rho + j e_\theta + k e_\varphi}{\sqrt{i^2+j^2+k^2}}; i,j,k \in \{-1,1,0\}, i^2+j^2+k^2 \neq 0 \right\} \tag{6.13}$$

Then in all directions as long as the gradient function has a negative value, we take a small step $\Delta d = \rho\,\varepsilon\sqrt{i^2+j^2+k^2}$ with $0 < \varepsilon \ll 1$ to the next data point and this procedure is continued

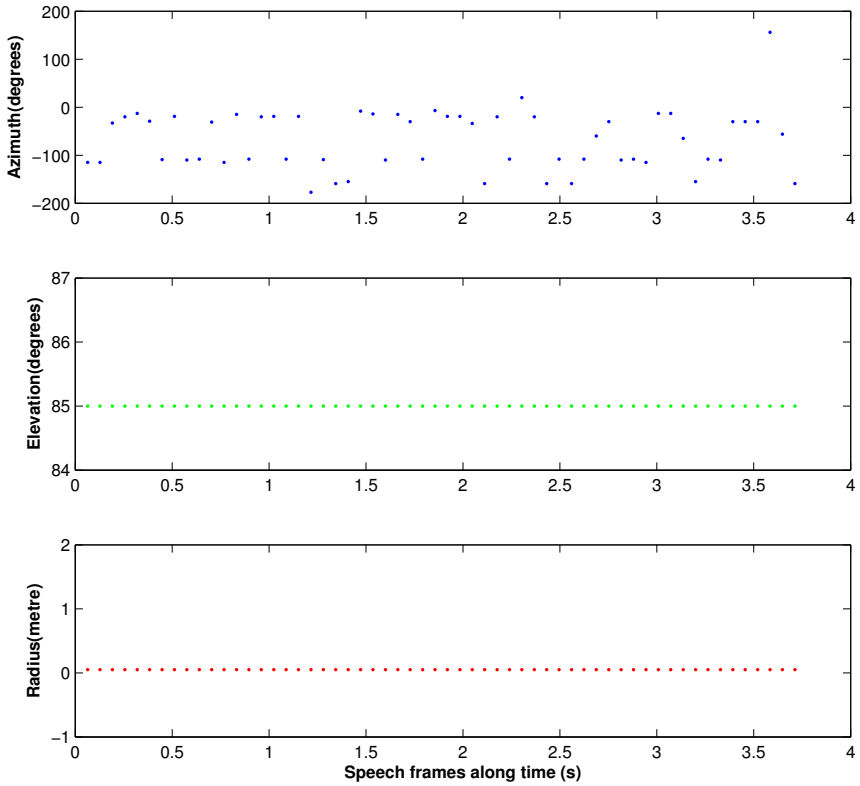

Fig. 2: SRP-PHAT localization in diffuse noise field

until all the data points with negative gradient are discarded from the search space. The residual is then searched to find the maximum power corresponding to the second dominant speaker. This procedure is continued until the SRP maximum corresponds to the *null* point in the search space. The number of active speakers at each frame is determined by detecting this *null* point in the SRP residual.

6.3 Experiments

In this section, we present experimental results on the proposed integrated framework of SRP-PHAT localization and MVAD based on (1) simulated data with the diffuse noise field and (2) real recording using the MONC as well as RT09 databases.

6.3.1 Diffuse Noise Field Simulation and Results

We consider a scenario in which three white noise sources are located at random positions in the room. The room impulse responses are generated with the image model technique [4] using intra-sample interpolation, up to $\mathbf{15}^{th}$ order reflections and omni-directional microphones. The corresponding reflection ratio, $\boldsymbol{\beta}$ used by the image model was calculated via Eyring's formula:

$$\boldsymbol{\beta} = \mathbf{exp}(-\mathbf{13.82}/[\boldsymbol{c} \times (\boldsymbol{L_x^{-1}} + \boldsymbol{L_y^{-1}} + \boldsymbol{L_z^{-1}}) \times \boldsymbol{T}]) \tag{6.14}$$

where $\boldsymbol{L_x}, \boldsymbol{L_y}$ and $\boldsymbol{L_z}$ are the room dimensions, $\boldsymbol{c}$ is the sound velocity in the air (≈342m/s) and $\boldsymbol{T}$ is the room reverberation time. In our experiments $\boldsymbol{T}$=300ms and the room direct-paths are discarded from the impulse responses for generation of the semi-diffuse noise signals [60]. Three noise sources are randomly positioned in the room and a circular microphone array with 8-channels and diameter of 20cm located at the center of the room records the diffuse noise.

The SRP-PHAT run on the multi-channel diffuse noise signal exhibits a consistent peak corresponding to the nearest point to the maximum elevation to the center of the array. The experiments are carried out for 128ms frames with 50% overlap. The maximum elevation in search space of our simulation as well as real data tests are $\mathbf{85}^\circ$ and $\mathbf{75}^\circ$ respectively. As illustrated in Fig.6.1 the *null* point exists at the $\boldsymbol{\rho}$=5cm and $\boldsymbol{\varphi} = \mathbf{85}^\circ$ in the search space. As expected, the azimuth value is almost random. These results provide empirical evidence of the formulation derived in section 6.2.3. This joint framework of gradient SRP-PHAT localization and voice activity detection is shown to work well under the assumption of the diffuse noise field for the real environments. In the following section, we conduct some experiment on the real data recordings.

6.3.2 Speech Database

We have evaluated our framework using two databases recorded in real environments: (1) The Multichannel Overlapping Numbers Corpus (MONC) [1]. We have used the following two recording scenarios; S1: one speaker located at $\boldsymbol{L}$1 (78cm,135°,23°) without overlapping, S12: one competing speaker located at $\boldsymbol{L2}$ (78cm,45°,23°), (2) Rich transcription (RT09) is structured for mixing metadata extraction and speech-to-text (STT) technologies. We use this database of a precise evaluation of MVAD using 8-channel microphone recordings. The details are explained in [2]. We use file EDI_20071128_1000_ci01_NONE.sph that was recorded in the IMR meeting room by array1 at Edinburgh. We use the ICSI ground truth, that is, a hand transcription automatically aligned with the data. In this sense, the ground truth can contain errors; however, the results are still informative. File EDI_20071128-1000.rttm was used as the ground truth.

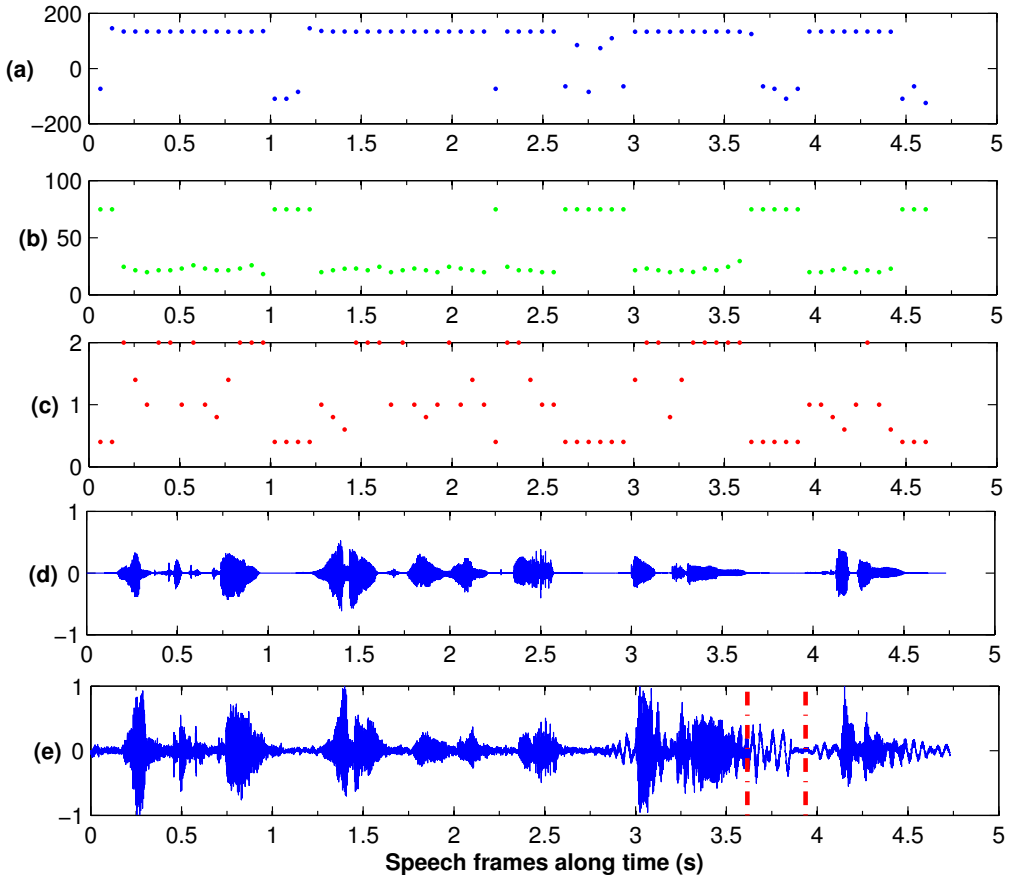

Fig. 2: Speaker localization using SRP-PHAT in non-overlapping conditions. (a) estimated azimuth (degrees), (b) estimated elevation (degrees), (c) estimated range (metre), (d) clean speech waveform, (e) distant speech recorded by microphone array

6.3.3 Single Speaker Localization and MVAD

In the first scenario, we run our algorithm on S1. The Signal-to-noise Ratio (SNR) is estimated about **9**dB. The results are depicted in Fig. 6.2. The 3D search space of SRP-PHAT consists of 150,000 points. The nearest vertical point to the center of the array (SRP *null*-point) is N ($\boldsymbol{\rho}$=0.45m,$\boldsymbol{\varphi}$=75°). The *null*-point is detected when there is no speech activity in the frame, e.g. the first two frames in the Fig. 6.2. The high accuracy of the proposed MVAD can be seen in Fig. 6.2(e). For instance, there exists a high energy noisy region between **3.62** and **3.94** seconds which has been correctly identified as a non-speech part of the signal. By removing the silent frames using MVAD, Fig. 6.3 is obtained. Note that the joint localization-VAD framework exhibits highly accurate results as the standard deviation (SD) of azimuth estimation is 0.5° and the SD of elevation estimation is 2°. Accurate estimation of the range however, is not possible.

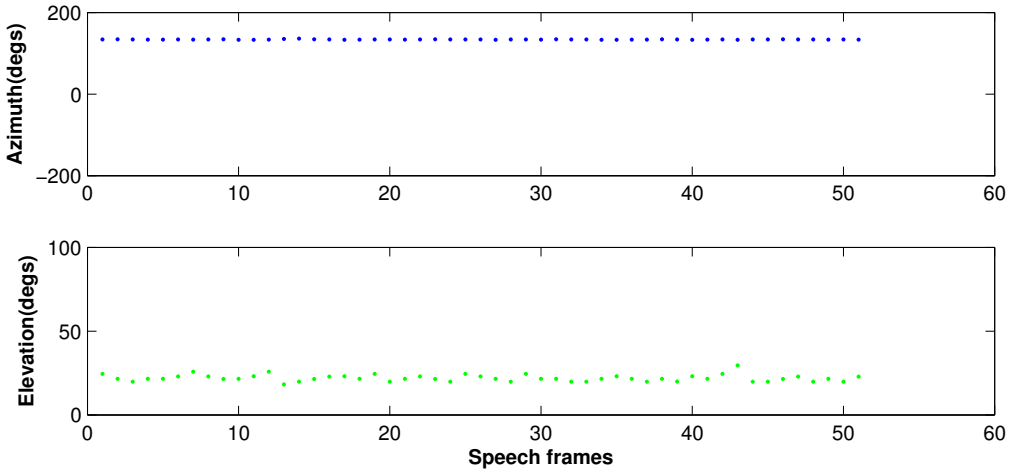

Fig. 2: Improvement of joint source localization and voice activity detection framework in non-overlapping condition

6.3.4 Multi-Speakers Localization and MVAD

The second scenario considers overlapping speech segments. In Fig. 6.4, generic SRP-PHAT has been used along with MVAD. As we can observe, only the dominant speaker is detected at each frame. The silent frames detected by MVAD are shown at azimuth = 0. As the figure illustrates, the dominant speaker is localized very accurately, the SD of azimuth estimation is 1°. If we use the spatial gradient modification of SRP-PHAT, we can localize both the dominant as well as the inferior speakers precisely at each frame. The results of this experiment are depicted in Fig. 6.5. The results are shown for the azimuth estimation. Upon the detection of a peak at elevation = 75°, the MVAD has detected a noisy region where no speech activity is present. The dominant speaker is indicated by circles; the inferior speaker is extracted by the gradient method and it is denoted by dots. The number of active speakers is determined when the *null*-point is detected in the gradient SRP-PHAT residual.

In our final experiment, we evaluate the proposed MVAD on part of the RT09 database. A total 315s of speech signal is processed in frames of length 256ms with 50% overlap. The speech material is taken from an **83**s and another **232**s segment of a file. This is in order to avoid physical noise such as door slams in the background and such that more than **18%** of frames are silent. This enables us to have a sound evaluation of MVAD. It is not an exhaustive test; we only aim to have an evaluation on a modern corpus.

The total error rate for MVAD is 6.4%, which consists of 2.7% missed speech and 3.7% false alarms. The proposed MVAD is practical in meeting recordings, which are usually moderate SNR speech but highly reverberant situations. The sample spectrogram and the speech waveform are illustrated in Fig. 6.6. The recognized silent parts (output of MVAD) are indexed with boxes in a yellow strip. As the figure shows, the signal is highly noisy. The majority of the errors in MVAD happen at the transitions of silent and speech.

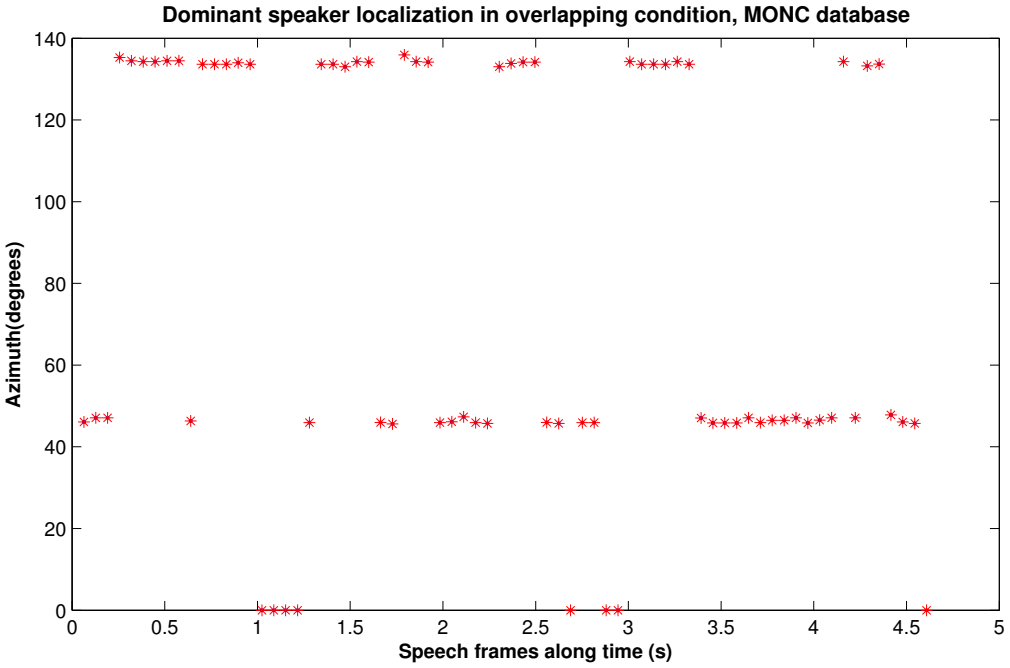

Fig. 2: Dominant speaker localization in overlapping condition using SRP-PHAT on MONC.

6.4 Conclusions

We proposed an integrated framework for multi-channel multi-source localization and voice activity detection which is very effective in real acoustic conditions and practical hands-free speech scenarios. Our method exploits the SRP localization technique. We introduced a spatial gradient modification to SRP-PHAT for localization of competing sources. We further worked out the SRP search space for the diffuse noise field and characterized a fixed point corresponding to the SRP peak for non-speech frames. This formulation led to introducing another application of the gradient SRP-PHAT as an MVAD. Experiments conducted on real data recordings showed that the framework could exhibit highly accurate results for multi-source localization and voice activity detection in microphone array applications, in particular in highly reverberant environments, such as aircraft cockpits and automobile interiors, where the noise fields are usually diffuse.

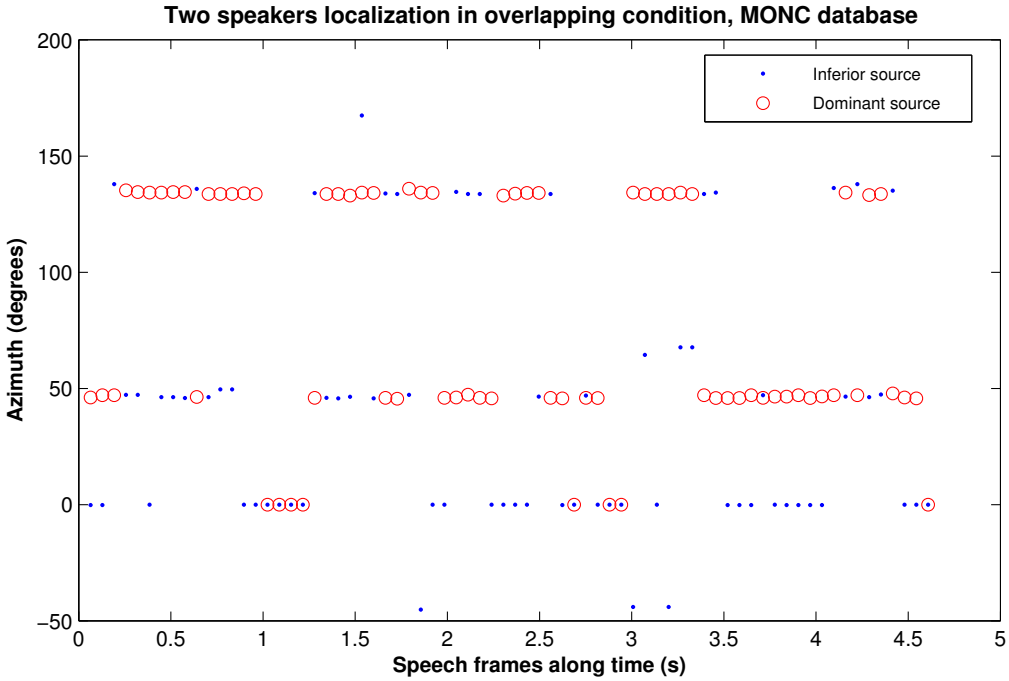

Fig. 2: Two competing speakers localization using spatial gradient extension of SRP-PHAT on MONC.

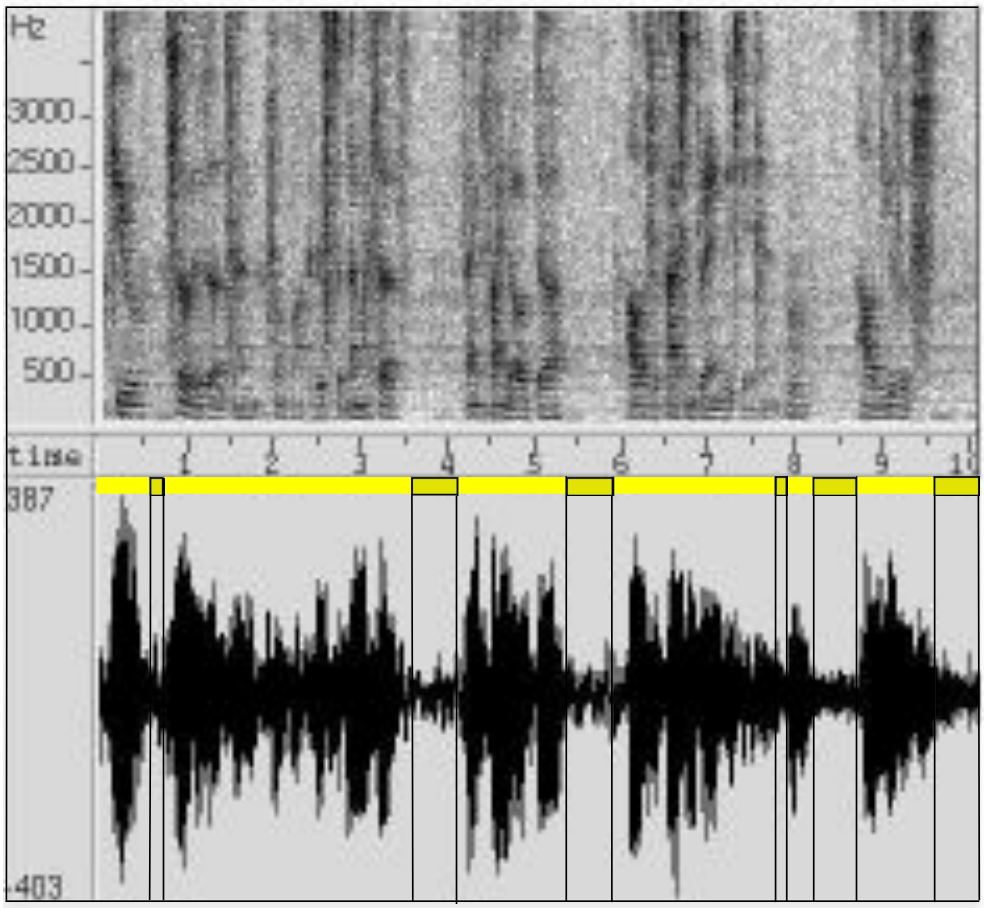

Fig. 2: Top: spectrogram and Bottom: speech waveform. Silent parts that recognized by MVAD have been showed with boxes in yellow strip.

7 Conclusion

This dissertation has addressed some of the key challenges to enable speech applications using ad hoc microphone arrays for recording distant speech signals. We proposed new methods for ad hoc microphone array calibration where only partial information on noisy pairwise distances is available. Furthermore, we worked out novel approaches to distributed source localization from asynchronous recordings. We also studied robust design of the array configuration as well as multi-source localization and voice activity detection.

7.1 Summary of achievements

The unknown geometry of the ad hoc microphone array is the first bottleneck to process multichannel recordings. To address this problem, we exploited the properties of a diffuse sound field for estimating the microphones' pairwise distances. We proposed an enhanced diffuse field coherence model that benefits from averaging the frame-based estimates to diminish the noise prior to model fitting. In addition, histogramming was found quite effective in elimination of the outliers and outperformed the alternative parametric approach used for clustering. The averaging and histogramming led to robustness to noisy estimates due to lack of diffuseness. In this context, we derived the relation between dimension of the room and the diffuseness level and the error in estimated pairwise distances. Furthermore, we proposed a measure to quantify the level of sufficient diffuseness for pairwise distance estimation.

If all the pairwise distances are available, a simple method such as multidimensional scaling can extract the geometry of the microphone array. However, in the scenario of ad hoc microphones, typically only a subset of distances are available. The missing pairwise distances correspond to the large ones or the unseen nodes due to environmental barriers. To address the problem of calibration using partial and noisy pairwise distances, we proposed a Euclidean distance matrix completion algorithm to recover the missing distances, and derived its theoretical upper bound of distance estimation error. These theoretical results suggested that the calibration error is decreased by increasing the number of microphones and further experimental studies demonstrated the theoretical insights. Furthermore, it is confirmed that

incorporating the properties of Euclidean distance matrices in a matrix completion procedure improves the calibration performance compared to the alternative generic matrix completion methods.

Building on our work on exploiting the properties of Euclidean distance matrices, we devised a novel technique for source localization and synchronization using single channel recordings. An image microphone model of multipath propagation is used to estimate the source distances to the virtual microphones and a EDM matrix recovery is developed for joint localization and synchronization. It is demonstrated that this approach achieves close to the global solution of this problem. Furthermore, this method is extended for distributed source localization using ad hoc sensors where synchronization is achieved locally and the location estimates are aggregated for globally optimized source localization.

In addition, we studied the design of a microphone array for source localization robust to noise in distance estimation. We formulated a minmax approach to find the best position of the microphones for the worst conceivable setting. This problem can be solved efficiently using the optimization procedure relying on generalized trust region subproblems. Although, the microphone positions are found by minimizing the worst-case source localization error, we found that the result of the proposed method was very close to the average optimal solution.

Finally, we proposed an integrated framework for multi-source localization and voice activity detection in a diffuse sound field. The theoretical findings are demonstrated on real data experiments to verify the applicability of underlying hypotheses and demonstrate the performance of the proposed methods.

7.2 Future Directions

Numerous research ideas are interesting to explore along the lines of the achievements of this thesis:

The diffuse sound field model and the assessment criteria can be integrated into the sound field reproduction and rendering systems for a better design that avoids noise amplification. Furthermore, it can be used as a new feature for robust higher level speech applications such as speech recognition.

The idea of matrix completion of EDMs for missing pairwise distance estimation and localization can be implemented within the procedure of different matrix completion algorithms, which results in new theoretical performance bounds and robustness analysis. One recent development based on the alternating direction of multipliers (ADMM) has been shown to improve the performance of this approach for Euclidean distance matrix completion. Hence, it remains to further study alternative matrix completion schemes to improve calibration and localization results.

In addition, enabling the use of sensors embedded in smart devices faces the problem of

synchronization. This problem is addressed in the framework of single channel and distributed localization given some prior knowledge on the pairwise distances. More research is required to tackle this problem in more general settings where we have limited knowledge on the geometry of the microphone array.

Along the lines of optimal design of the microphone array for robust localization, we can further consider the multipath propagation model in a reverberant enclosure. Moreover, we can formulate the design problem to achieve the best performance in terms of sound recording and reproduction for 3D audio technologies.

Finally, several higher level applications can be foreseen using the ad hoc microphone array recordings. Some preliminary results are provided in the appendix in terms of speech recognition. Indeed, we should study more realistic scenarios and other distant audio applications.

7.3 Concluding Remarks

Enabling speech applications using ad hoc microphone arrays can potentially have a significant impact on future sound technologies. The ubiquitous use of sensor embedded devices provides an abundance of acoustic information that can be exploited to develop sophisticated systems that earlier could only be devised for complex array infrastructures. This thesis provided some initial answers to some of the challenging problems that we need to tackle for the future sound design and distant technologies.

A Microphone Array Beampattern Characterization for Hands-free Speech Applications

Spatial filtering is the fundamental characteristic of microphone array based signal acquisition, which plays an important role in applications such as speech enhancement and distant speech recognition. In the array processing literature, this property is formulated upon beampattern steering and it is characterized for narrowband signals.

This appendix proposes to characterize the microphone array broadband beampattern based on the average output of a steered beamformer for a broadband spectrum. Relying on this characterization, we derive the directivity beampattern of delay-and-sum and superdirective beamformers for a linear as well as a circular microphone array. We further investigate how the broadband beampattern is linked to speech recognition feature extraction; hence, it can be used to evaluate distant speech recognition performance. The proposed theory is demonstrated with experiments on real data recordings. The content of this appendix has been published in IEEE Sensor Array and Multi-channel Signal Processing Workshop (SAM 2011) [110].

A.1 Introduction

Multi-channel signal acquisition relies on beamforming or spatial filtering for directional discrimination, and space-time filtering of the signals in the acoustic scene [123, 39]. An important issue then is to design an optimal microphone array to achieve a desired look-angle directivity and suppression of interference. This task is usually entangled with constraints on Signal-to-Noise Ratio (SNR), side-lobe level and beamwidth. Whilst these parameters have been characterized in the array processing literature, the theory usually revolves around narrowband assumptions. To address the issues in acquisition of wideband signals, Coleman

et al. [37] propose to characterize the wideband beampattern based on random-process autocorrelations and cross correlations. Their formulation provides mean-square-error and average-gain measures for far-field beamforming obtained through convex optimization. In the microphone array signal processing literature, the broadband characterization of beampattern has not been addressed and, to the extent of our knowledge, the microphone array acquisition has been understood through the narrowband properties [128].
In this chapter, we formulate the broadband beampattern and directivity for acquisition of speech signals. Our formulation exploits the concept of the narrowband beampattern while the beamformer's output power for a broadband spectrum is used to characterize the power-pattern. We then show how this formulation is linked to the speech recognition front-end processing, hence the power-pattern enables us to evaluate the performance of the distant speech recognition system.

The rest of the chapter is organized as follows. We give a quick overview of the narrowband formulation of the beampattern in Section A.2.1 The extension of this formulation for speech signal is explained in Section A.2.2 Section A.2.3 is dedicated to the simulations of the proposed method. Section A.3.1 present the experimental demonstration of the theory. The link between the broadband beampattern and the speech recognition front-end processing is shown in Sections A.3.2 and A.3.3. The conclusions are drawn in Section A.4.

A.2 Broadband Beampattern

A.2.1 Microphone Array Pattern

We consider a general array of isotropic elements in a homogenous medium. A plane wave of the external signal field impinges on the array from the direction of $\bar{a}(\theta,\phi)$ where θ and ϕ denote the elevation and the azimuth angles in spherical coordinates. Wavenumber k is defined as

$$k = \frac{2\pi}{\lambda}\bar{a}(\theta,\phi), \tag{A.1}$$

where λ is the wavelength in radians with frequency ω. The narrowband beampattern is defined as the spatial and temporal frequency response of the array evaluated against the direction [21, 122] and stated as

$$B(\omega,\bar{a}(\theta,\phi)) = \sum_{m\in M} H_m(\omega)e^{-jk.m} \tag{A.2}$$

where $H_m(\omega)$ is frequency response filter of the microphone located at m; M is the set of microphone locations. The power-pattern is then defined as

$$P(\omega,\bar{a}(\theta,\phi)) = |B(\omega,\bar{a}(\theta,\phi))|^2. \tag{A.3}$$

The array directivity denoted by $\boldsymbol{D}$ is defined as the ratio of maximum power-pattern to the average of power-pattern in all directions, stated concisely as

$$\boldsymbol{D(\omega,\bar{a}(\theta_0,\phi_0))} = \frac{\boldsymbol{P(\omega,\bar{a}(\theta_0,\phi_0))}}{\frac{1}{4\pi}\int_{-\pi/2}^{\pi/2} \boldsymbol{d\theta} \int_0^{2\pi} \boldsymbol{d\phi sin(\theta) P(\omega,\bar{a}(\theta,\phi))}}, \tag{A.4}$$

where $\boldsymbol{\bar{a}(\theta_0,\phi_0)}$ is the steering direction, which is constant along frequency bands.

A.2.2 Broadband Beampattern for Speech Acquisition

In this section, we exploit the concept of narrowband beampattern and formulate the beampattern for acquisition of broadband signals such as speech. Suppose that $\boldsymbol{S(\omega)}$ is the spectral representation of the clean speech signal in Fourier domain (estimated from a database). The spectrum of speech has a complex structure. Most of the energy is generated during the voiced parts and concentrated in three to four formants up to 2 KHz in frequency. So we considered this structure for extracting the beampattern. The spectrum of the speech signal can be extracted by the Welch's method, with non-overlapping block processing of size **128**ms. Hence, the response of the array or the beamformer (e.g., superdirective, delay-and-sum) denoted by $\boldsymbol{F(\omega,\bar{a}(\theta,\phi))}$ to the plane wave $\boldsymbol{S(\omega)}$ would be

$$\boldsymbol{Y(\omega,\bar{a}(\theta,\phi))} = \boldsymbol{F(\omega,\bar{a}(\theta,\phi))S(\omega).} \tag{A.5}$$

In other words, $\boldsymbol{Y(\omega)}$ is the beamformer output for the broadband spectrum of the signal over the sphere of look directions. Given $\boldsymbol{Y(\omega)}$, we define the broadband beampattern as

$$\boldsymbol{B_{sp}(\bar{a}(\theta,\phi))} = \frac{\sqrt{\int_0^{\omega_N} \boldsymbol{Y^2(\omega,\bar{a}(\theta,\phi))d\omega}}}{\sqrt{\int_0^{\omega_N} \boldsymbol{Y^2(\omega,\bar{a}(\theta_0,\phi_0))d\omega}}}, \tag{A.6}$$

where $\boldsymbol{\omega_N}$ is the Nyquist frequency. The proposed beampattern can be interpreted as a weighted average of the beamformer's output over the speech signal. Thereby, it is mostly influenced by the dominant frequencies of speech spectrum. Accordingly, the power-pattern for the broadband spectrum would be

$$\boldsymbol{P_{sp}(\bar{a}(\theta,\phi))} = \boldsymbol{|B_{sp}(\bar{a}(\theta,\phi))|^2}, \tag{A.7}$$

and the directivity for speech acquisition will be defined as

$$\boldsymbol{D_{sp}(\bar{a}(\theta_0,\phi_0))} = \frac{\boldsymbol{P_{sp}(\bar{a}(\theta_0,\phi_0))}}{\frac{1}{4\pi}\int_{-\pi/2}^{\pi/2} \boldsymbol{d\theta} \int_0^{2\pi} \boldsymbol{d\phi sin(\theta) P_{sp}(\bar{a}(\theta,\phi))}}. \tag{A.8}$$

The directivity as defined above, can be interpreted as the array gain for speech acquisition in the presence of isotropic noise. The $3\boldsymbol{dB}$ beamwidth is a measure of the width of the main lobe of the beampattern. It is defined as the maximum angle in normalized power-pattern for which power is above **0.5**. Applying the normalization in weights so that $\boldsymbol{P_{sp}(\bar{a}(\theta_0,\phi_0))} = \boldsymbol{1}$,

the normalized directivity can be written as

$$\hat{D}_{sp}(\bar{a}(\theta_0,\phi_0)) = \left(\frac{1}{4\pi}\int_{-\pi/2}^{\pi/2} d\theta \int_0^{2\pi} d\phi sin(\theta) P_{sp}(\bar{a}(\theta,\phi)\right)^{-1}. \tag{A.9}$$

The broadband directivity provides an objective for the acquisition of the speech signal. Assuming that the background noise is diffuse, SNR maximization becomes equivalent to the maximization of the broadband directivity.

A.2.3 Simulations

We present some empirical validation of the proposed theory. These studies aim to provide a broad view of the beampattern for different acquisition set-up of microphone array and to compare and contrast the speech vs. narrowband signal's beampattern.

Linear Microphone Array

We consider a linear uniform array of 5 omnidirectional microphones. The aperture size is 38.25 cm and the sampling frequency is 8 kHz. The spacing between the microphones is set to half of the wavelength of a (2246 Hz) frequency in the speech spectrum below which most of the power exists to suppress the majority of the grating lobes. The speech broadband beampattern as expressed through equation A.6 is plotted in figure A.1 vs. the narrowband beampattern of a delay-and-sum beamformer for frequencies equal to 250, 500, 1000 and 2246 Hz. As we can observe, there is no null in the speech beampattern and the microphone array captures the signal in all directions, whereas there are clearly 8 nulls in the 2246 Hz beampattern. The 3 dB beamwidth for speech beampattern is about 80° while for the for a $\mathbf{2246}\boldsymbol{Hz}$ beampattern, it is 20°. The difference quantifies the reduction in directivity of the microphone array for speech signal. The broadband as well as the narrowband beampatterns for a superdirective beamformer are illustrated in figure A.2. As we can see, the same argument holds as for the delay-and-sum beamformer; i.e., the speech beampattern does not have any null and thus null steering by a uniform linear microphone array is impractical. Side-lobe levels at 0° and 180° are -20 dB.

Circular Microphone Array

The circular microphone array is a common structure used for meeting acquisition and robotics. We consider here a scenario of 8 uniformly placed omnidirectional microphones with diameter of 20 cm. So the spacing between the microphones is equal to our previous study with linear microphone array to minimize the majority of the sidelobes. Figure A.3 shows the speech beampattern vs. the narrowband beampattern of the delay-and-sum beamformer at 250, 500, 1000 and 2246 Hz. We observe that the 3 dB beamwidth at 2246 Hz is about 20°. There are 6 nulls and the opposite sidelobe level is -9 dB. However, the speech beampattern

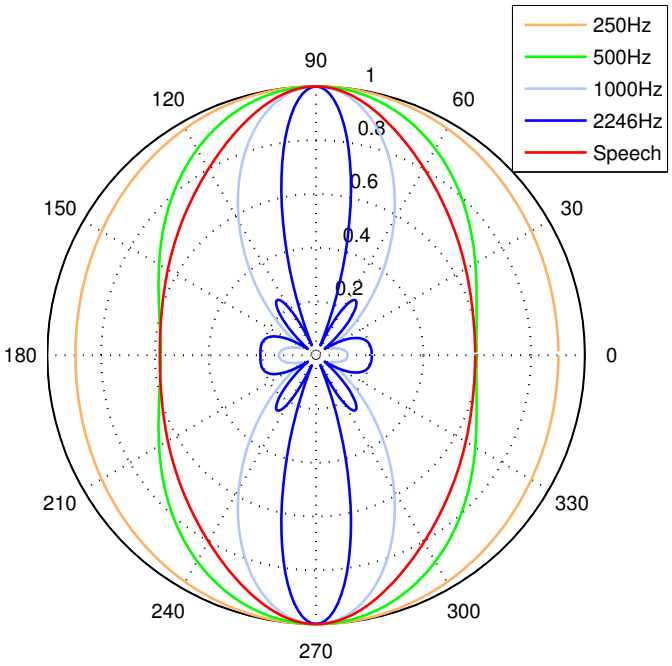

Figure A.1: Speech vs. narrowband beampattern for delay-and-sum beamformer with linear microphone array

does not have any null and the 3 dB beamwidth is about 80°; it is evident that the directivity is much smaller. In figure A.4, the broadband vs. narrowband beampatterns are contrasted for a superdirective beamformer. The 3 dB beamwidth at 2246 Hz is about 25° and the opposite sidelobe level is about -8 dB whereas the 3 dB beamwidth for the speech beampattern is 40° while the sidelobe level is decreased to -28 dB. The 8 nulls exhibited in the 2246 Hz beampattern do not exist in the speech beampattern.

A.3 Experiments

We now present some evaluations on real data recordings to see how the theory formulated in the previous section matches multi-channel speech acquisition. We also demonstrate the relationship between the beampattern and distant speech recognition performance.

A.3.1 Speech Acquisition

The experiments are performed in the framework of Multi-channel Overlapping Numbers Corpus [1]. We used the single speaker recordings captured by an 8-channel circular microphone array of diameter 20 cm. The speaker is located at azimuth and elevation 135° and 25° respectively, related to center of microphone array. The hypothesized room set-up for the theory described in Section A.2 is a 6 sided enclosure with reflection coefficients of

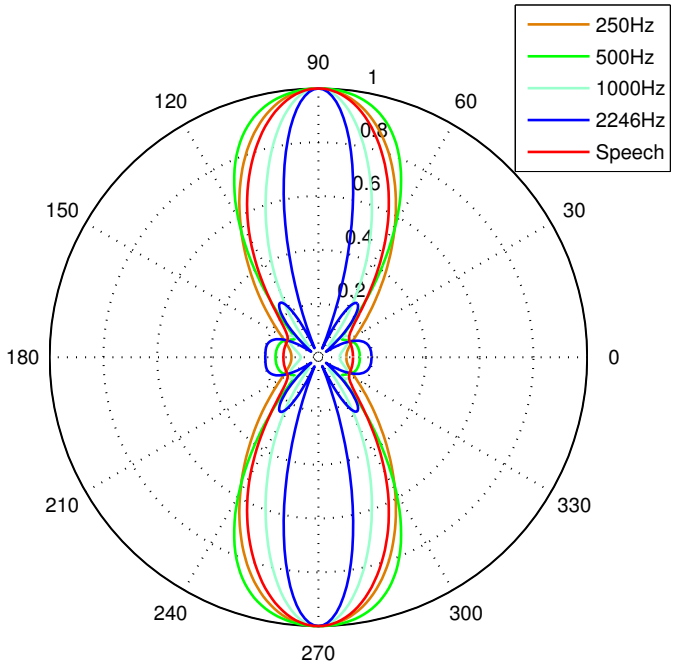

Figure A.2: Speech beampattern vs. narrowband beampattern for superdirective beamformer with linear microphone array

zero. We also assumed that the mutual coupling between the microphones is almost zero. However, our empirical evaluations are performed in a meeting room fully furnished and the microphone array is installed on wood. Hence, the hypothesized ideal condition does not hold in practice and this could justify the difference between the theoretical expected results and our experimental evaluation.

The measured beampattern for a superdirective beamformer is illustrated in figure 5(d) which has a similarity to what we obtain in figure 4. It is wider however due to the wall reflections. Because we use a planar array, the beampattern has a fan style that captures more signals from the walls. Moreover, the recordings were made in a moderately reverberant 8.2 m × 3.6 m × 2.4 m rectangular room. The reverberation time is estimated about 200 ms. The corresponding reflection ratio, $\boldsymbol{\beta}$ is about 0.8, calculated via Eyring's formula [51]. This can justify the -8 dB opposite sidelobe level exhibited in figure 5(d). The measured beampattern of the delay-and-sum beamformer as depicted in 5(e) also looks similar to the speech beampattern that we obtain in figure 3. Similar to the previous argument, we obtain a -8 dB increase in the opposite sidelobe level. In summary, our experiments demonstrate the proposed theory of broadband beampattern; however, it does not take reverberation and mutual coupling into account.

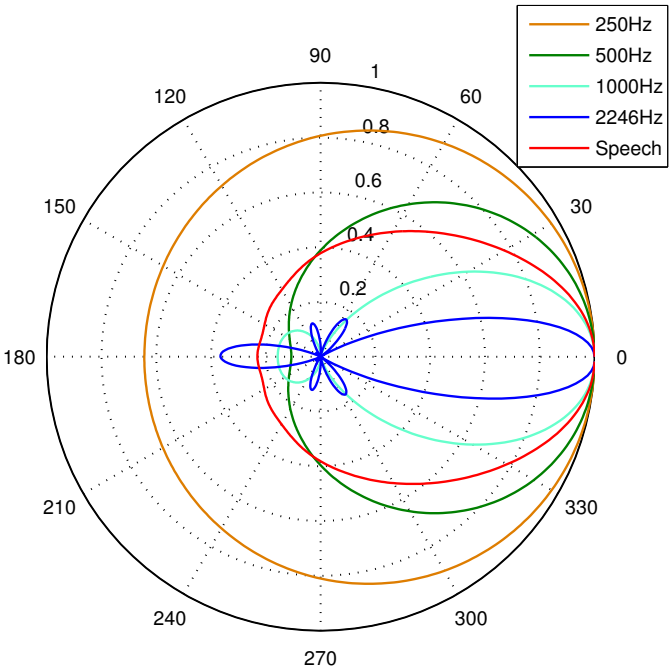

Figure A.3: Speech beampattern vs. narrowband beampattern for delay-and-sum beamformer with circular array

A.3.2 Speech Recognition

The automatic speech recognition (ASR) scenario was designed to broadly mirror that of Moore and McCowan [81]. A typical front-end was constructed using the HTK toolkit with 25 ms frames at a rate of 10 ms. This produced 12 mel-cepstra plus the zeroth coefficient and the first and second time derivatives; 39 features in total. Cepstral Mean Normalization (CMN) is applied to the feature vectors which improves the speech recognition performance about 15%. The average SNR of the recordings is 9 dB. The dominated noise has diffuse characteristics [109] so we use McCowan-Bourlard post-filter [77] to achieve a higher accuracy using superdirective beamformer. Whereas Moore and McCowan [81] performed MAP adaptation, our results were obtained by training directly on beam-formed data. The maximum ASR accuracy of the system is about 95%. We extract the ASR pattern by scanning all directions using a superdirective beamformer. Figure A.5(a) shows the ASR results after normalization with respect to the maximum word recognition rate.

A.3.3 Discussion

Neither the predicted beampattern nor measured power pattern match the ASR pattern; we would not expect an exact match as they are different measures. However, notice that the logarithm of the power-pattern plus one fits the ASR pattern reasonably well. It is illustrated

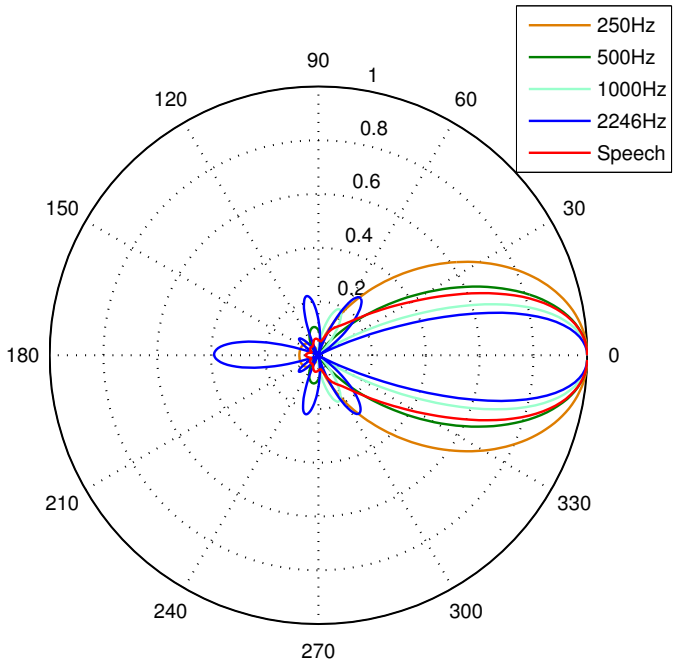

Figure A.4: Speech beampattern vs. narrowband beampattern for superdirective beamformer with circular microphone array.

in figure A.5 (b).

To investigate the link between the broadband power-pattern and speech recognition performance, we consider a simplistic, but informative view of an ASR front-end. The acoustic signal is treated in Fourier domain and a non-linear frequency warping is applied through a filterbank. To obtain the cepstrum features, a logarithm is applied followed by another linear transform to achieve the decorrelation and dimensionality reduction. The logarithm is motivated to approximate the sensitivity of the ear. The CMN is a common practice to reduce the channel effect and improve the performance. Garner [56] showed that in the presence of CMN, the feature presented to the ASR decoder is (a linear transform of)

$$\boldsymbol{c} = \boldsymbol{log}(\mathbf{1} + \boldsymbol{s/n}), \tag{A.10}$$

where $\boldsymbol{s/n}$ is the signal to noise ratio in the spectral domain, and $\boldsymbol{c}$ is the normalized cepstrum. We tentatively conclude that this logarithm of the speech power pattern plus one is a reasonable predictor of ASR performance. This measure is also the Shannon channel capacity for a Gaussian channel, suggesting an information theoretic relationship too. A limitation of this work is that we don't consider perceptual warping in characterization of the speech beampattern.

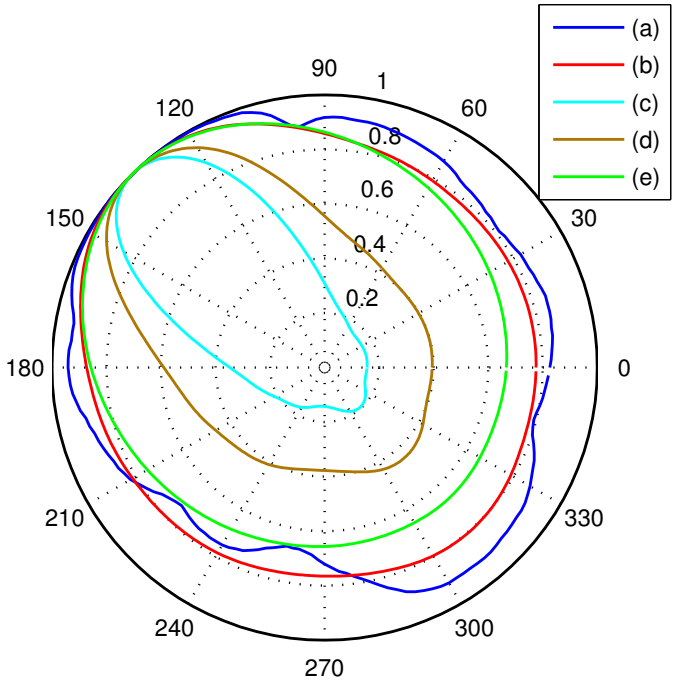

Figure A.5: beampattern, power-pattern and distant speech recognition performance for circular microphone array used in MONC recordings: (a) normalized ASR word accuracy, (b) logarithm of measured speech power-pattern plus one, (c) measured speech power-pattern, (d) measured speech beampattern, (a)-(d) are plotted for superdirective beamformer and (e) measured speech beampattern of delay-and-sum beamformer

A.4 Conclusion

We described a new method for characterizing a microphone array beampattern for the broadband spectrum of a speech signal. We demonstrated the theoretical implications on a variety of microphone array designs, suggesting a generally wider beampattern with small sidelobes that can show the response of the microphone array for broadband signals. We also observed that the broadband beampattern provides a good estimation of the observable beam in all directions. A high similarity is observed between the logarithm of power-pattern plus one and the speech recognition performance. We hence conclude that the proposed method is a valuable approach for analysis of the microphone array structure in terms of speech quality and recognition in hand-free acquisition and it could be further exploited in the design of an optimal geometry for speech applications.

Bibliography

[1] The multichannel overlapping numbers corpus (MONC). Idiap resources available online:. http://www.cslu.ogi.edu/corpora/monc.pdf.

[2] The 2009 (rt-09) rich transcription meeting recognition evaluation plan. RT-06S Transcription Evaluation Plan:. http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf.

[3] Abdeldjalil Aissa-El-Bey and Karim Abed-Meraim. Blind simo channel identification using a sparsity criterion. *Proc. 9th IEEE Workshop on Signal Processing Advances for Wireless Communications SPAWC, Recife, Brazil*, pages 271–275, 2008.

[4] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of Acoustical Society of America*, 60(s1), 1979.

[5] S. Araki, H. Sawada, R. Mukai, and S. Makino. DOA estimation for multiple sparse sources with arbitrarily arranged multiple sensors. *Journal of Signal Processing Systems*, 63(2), 2011.

[6] A. Asaei, M. J. Taghizadeh, M. Bahrololum, and M. Ghanbari. Verified speaker localization utilizing voicing level in split-bands. *Signal Processing*, 89(6), 2009.

[7] A. Asaei, P. N. Garner, and H. Bourlard. Sparse component analysis for speech recognition in mullti-speaker environment. In *Proceedings of INTERSPEECH*, 2010.

[8] A. Asaei, M. J. Taghizadeh, H. Bourlard, and V. Cevher. Multi-party speech recovery exploiting structured sparsity models. In *The 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[9] A. Asaei, H. Bourlard, M. J. Taghizadeh, and V. Cevher. Model-based sparse component analysis for reverberant speech localization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2014.

[10] Afsaneh Asaei. *Model-based Sparse Component Analysis for Multiparty Distant Speech Recognition*. PhD thesis, École Polytechnique Fédéral de Lausanne (EPFL), 2013.

[11] Afsaneh Asaei, Mohammad Golbabaee, Hervé Bourlard, and Volkan Cevher. Structured sparsity models for reverberant speech separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(3):620–633, 2014.

[12] Amir Beck, Petre Stoica, and Jian Li. Exact and approximate solutions of source localization problems. *Signal Processing, IEEE Transactions on*, 56(5):1770–1778, 2008.

[13] P. Biswas, T. chen Liang, K. chuan Toh, T. chung Wang, , and Y. Ye. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *IEEE Transactions on Automation Science and Engineering*, 3, 2006.

[14] P. Biswas, T. C. Liang, K. C. Toh, T. C. Wang, and Y. Ye. Semidefinite programming approaches for sensor network localization with noisy distnce measurments. *IEEE Transactions on Automation Science and Engineering*, 3, 2006.

[15] J. Bitzer, K. Kammeyer, and K. U. Simmer. An alternative implementation of the superdirective beamformer. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1999.

[16] J. Bitzer, K. U. Simmer, and K. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[17] C. Blandin, A. Ozerov, and E. Vincent. Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92, 2012.

[18] I. Borg and P. J. F. Groenen. Modern multidimensional scaling theory and applications. Springer, 2005.

[19] S. Boyd and L.Vandenberghe. Convex optimization. In *Cambridge*, 2012.

[20] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[21] M. Brandstein and D. Ward. Superdirective microphone arrays. In *Microphone Arrays*, chapter 2, pages 19–37. Springer, 2001.

[22] M. S. Brandstein and H. F. Silverman. A practical methodology for speech source localization with microphone arrays. In *Computer Speech and Language*, 1997.

[23] Michael Brandstein and Darren Ward, editors. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[24] T. Bravo and C. Maury. Enhancing low frequency sound transmission measurements using a synthesis method. *Journal of the Acoustical Society of America*, 122(2), 2007.

[25] H. Buchner, R. Aichner, and W. Kellerman. Trinicon: A versatile framework for multichannel blind signal processing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[26] A. Buja and D. F. Swayne. Visualization methodology for multidimensional scaling. *Journal of Classification*, 19, 2002.

[27] J.A. Cadzow. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36, 1988.

[28] E. J. Candes and Y. Plan. Matrix completion with noise. *IEEE Signal Processing Magazine*, 98(6), 2010.

[29] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Magazine Communications of the ACM (CACM)*, 55, 2012.

[30] Avishy Carmi and Pini Gurfil. Sensor selection via compressed sensing. *Automatica*, 49 (11):3304–3314, 2013.

[31] J. Chen, J. Benesty, and A. Huang. Mimo acoustic signal processing. In *HSCMA Workshop, Invited Talk*, 2005.

[32] M. Chen, Z. Liu, L. He, P. Chou, and Z. Zhang. Energy-based position estimation of microphones and speakers for ad-hoc microphone arrays. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[33] Sundeep Prabhakar Chepuri and Geert Leus. Sparsity-promoting sensor selection for non-linear measurement models. *arXiv preprint arXiv:1310.5251*, 2013.

[34] C. F. Chien and W. W. Soroka. Spatial cross-correlation of acoustic pressures in steady and decaying reverberant sound fields. *Journal of Sound and Vibration*, 48(2), 1976.

[35] W. T. Chu. Eigenmode analysis of the interference patterns in reverberant sound fields. *Journal of the Acoustical Society of America*, 68(1), 1980, .

[36] W.T. Chu. Spatial cross-correlation of reverberant sound fields. *Journal of Sound and Vibration*, 62(2), 1979, .

[37] J. O. Coleman and R. J. Vanderbei. Random-process formulation of computationally efficient performance measures for wideband arrays in the far field. In *Proc. Midwest Symp. on Circuits and Systems (MWSCAS)*, 1999.

[38] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson. Measurement of correlations coefficients in reverberan sound fields. *Journal of the Acoustical Society of America*, 27, 1955.

[39] H. Cox, R. M. Zeskind, and T. Kooij. Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34, 1986.

[40] T. F. Cox and M. A. A. Cox. Multidimensional scaling. *Chapman-Hall, 2001.*

[41] Marco Crocco, Alessio Del Bue, and Vittorio Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Transactions on Signal Processing*, 60(2):660–673, 2012.

[42] J. Dattorro. *Convex Optimization and Euclidean Distance Geometry*. Meboo Publishing, USA, 2012.

[43] Antoine Deleforge, Florence Forbes, and Radu Horaud. Variational em for binaural sound-source separation and localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 76–80, 2013.

[44] J. H. DiBiase. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. In *PhD Thesis, Brown University in Providence, Rhode Island, United States*, 1993.

[45] J. Dmochowski, S. Benesty, and S. Affes. Broadband music: opportunities and challenges for multiple source localization. In *IEEE WASPAA*, 2007.

[46] J. P. Dmochowski and J. Benesty. Steered beamforming approaches for acoustic source localization. *I. Cohen, J. Benesty, and S. Gannot (Eds.), Speech Processing in Modern Communication, Springer*, 24 (4):307 – 337, 2010.

[47] H. T. Do. *Robust cross-correlation-based methods for sound-source localization and separation using a large-aperture microphone array*. PhD thesis, Brown University, 2011.

[48] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30): 12186–12191, 2013.

[49] N. Doukas, P. Naylor, and T. Stathaki. Voice activity detection using source separation techniques. In *Eurospeech*, 1997.

[50] P. Drineas, M. Javed, M. Magdon-Ismail, G. Pandurangant, R. Virrankoski, and A. Savvides. Distance matrix reconstruction from incomplete distnce information for sensor network localization. *Sensor and Ad Hoc Communications and Networks*, 2, 2006.

[51] C. F. Eyring. Reverberation time in 'dead' rooms. *Journal of the Acoustical Society of America*, 1:217–241, 1930.

[52] M. Omologo F. Nesta, P. Svaizer. Cumulative state coherence transform for a robust two-channel multiple source localization. In *Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009.

[53] B. P. Flanagan and K. L. Bell. Array self-calibration with large sensor position errors. *Signal Processing*, 81, 2001.

[54] Charles Fortin and Henry Wolkowicz. The trust region subproblem and semidefinite programming. *Optimization methods and software*, 19(1):41–67, 2004.

[55] B. Gapinski, M. Grezelka, and M. Ruck. The roundness deviation measurment with coordinate measuring machines. *Engineering Review*, 26(2), 2006.

[56] Philip N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.

[57] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, 2011. http://cvxr.com/cvx.

[58] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. In *IEEE Trans. Ant. Prop.*, 1982.

[59] Fredrik Gustafsson and Fredrik Gunnarsson. Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements. *Signal Processing Magazine, IEEE*, 22(4):41–53, 2005.

[60] E. Habets. Generating sensor signals in isotropic noise fields. In *Journal of the Acoustical Society of America (JASA) Volume 122, Issue 6,*, 2007.

[61] Marius Hennecke, Thomas Plotz, Gernot A. Fink, Joerg Schmalenstroeer, and Reinhold Haeb-Umbach. A hierarchical approach to unsupervised shape calibration of microphone array networks. In *IEEE/SP 15th Workshop on Statistical Signal Processing (SSP)*, pages 257–260, 2009.

[62] Timo Hiekkanen, Tero Lempiäinen, Martti Mattila, Ville Pulkki, and Ville Veijanen. Reproduction of virtual reality with multichannel microphone techniques. In *Proceeding of 122nd AES Convention*, 2007.

[63] Y. Huang, J. Benesty, and G. W. Elko. Adaptive eigenvalue decomposition algorithm for real-time acoustic source localization. In *Proceedings of ICASSP*, 1999.

[64] D. M. Titterington I. Ford and Christos P. Kitsos. Recent advances in nonlinear experimental design. *Technometrics*, 31:49–60, 1989.

[65] Siddharth Joshi and Stephen Boyd. Sensor selection via convex optimization. *Signal Processing, IEEE Transactions on*, 57(2):451–462, 2009.

[66] Vassilis Kekatos, Georgios B Giannakis, and Bruce Wollenberg. Optimal placement of phasor measurement units via convex relaxation. *IEEE Trans. on Power Systems*, 27(3): 1521–1530, 2012.

[67] R. H. Keshavan and S. Oh. OptSpace: A gradient descent algorithm on the Grassman manifold for matrix completion. 2009. `arXiv:0910.5260`.

[68] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11, 2010.

[69] G. Kim and N.I. Cho. Voice activity detection using phase vector in microphone array. In *Electronics Letters*, 2007.

[70] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, Signal Processing*, 24 (4):320 – 327, 1976.

[71] K. Kumatani, J. McDonough, and B. Raj. Microphone array processing for distant speech recognition:from close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine, Special Issue on Fundamental Technologies in Modern Speech Recognition*, 2012.

[72] Bracha Laufer, Ronen Talmon, and Sharon Gannot. Relative transfer function modeling for supervised source localization. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4, 2013.

[73] Avinoam Levy, Sharon Gannot, and Emanuël AP Habets. Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1540–1555, 2011.

[74] Yuanqing Lin, Jingdong Chen, Youngmoo Kim, and Daniel D Lee. Blind channel identification for speech dereverberation using l1-norm sparse learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2007.

[75] Z. Liu, Z. Zhang, L. He, and P. Chou. Energy-based sound source localization and gain normalization for ad-hoc microphone arrays. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[76] Engin Masazade, Makan Fardad, and Pramod K Varshney. Sparsity-promoting extended kalman filtering for target tracking in wireless sensor networks. *Signal Processing Letters, IEEE*, 19(12):845–848, 2012.

[77] I. McCowan and H. Bourlard. Microphone array post-filter based on noise field coherence. *IEEE Transactions on Speech and Audio Processing*, 11, 2003.

[78] I. McCowan, M. Lincoln, and I. Himawan. Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio,Speech and Language Processing*, 16(3), 2008.

[79] Paul Meissner, Christoph Steiner, and Klaus Witrisal. UWB positioning with virtual anchors and floor plan information. In *IEEE 7th Workshop on Positioning Navigation and Communication (WPNC)*, 2010.

[80] Dmitri Model and Michael Zibulevsky. Signal reconstruction in sensor arrays using sparse representations. *Signal Processing*, 86 (3):624 – 638, 2006.

[81] D. C. Moore and I. A. McCowan. Microphone array speech recognition : Experiments on overlapping speech in meeting. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[82] C.T. Morrow. Point-to-point correlation of sound pressures in reverberation chambers. *Journal of Sound and Vibration*, 16(1), 1971.

[83] Jorge J. Moré. Generalizations of the trust region problem. *Optimization Methods and Software*, 2:189–209, 1993.

[84] S. Nam and R. Gribonval. Physics-driven structured cosparse modeling for source localization. 2012.

[85] H. Nelisse and J. Nicolas. Characterization of a diffuse field in a reverberant room. *Journal of the Acoustical Society of America*, 101(6), 1997.

[86] Francesco Nesta and Maurizio Omologo. Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 213–216, 2012.

[87] M. Omologo and P. Svaizer. Acoustic source localization in noisy and reverberant environments using csp analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996.

[88] R. Parhizkar. *Euclidean Distance Matrices: Properties, Algorithms and Applications*. PhD thesis, École École polytechnique fédérale de Lausanne EPFL, 2014.

[89] Reza Parhizkar, Ivan Dokmanic, and Martin Vetterli. Single-channel indoor microphone localization. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1434–1438, 2014.

[90] Neal Patwari, Joshua N Ash, Spyros Kyperountas, Alfred O Hero, Randolph L Moses, and Neiyer S Correal. Locating the nodes: cooperative localization in wireless sensor networks. *Signal Processing Magazine, IEEE*, 22(4):54–69, 2005.

[91] I. Potamitis. Estimation of speech presence probability in the field of microphone array. In *IEEE Signal Processing Letters*, 2004.

[92] B. Rafaely. Spatial-temporal correlation of a diffuse sound field. *Journal of the Acoustical Society of America*, 107, 2000.

[93] Juri Ranieri, Amina Chebira, and Martin Vetterli. Near-optimal sensor placement for linear inverse problems. *Signal Processing, IEEE Transactions on*, 62(5), 2014.

[94] V. C. Raykar, I. V. Kozintsev, and R. Lienhart. Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE Transactions on Speech and Audio Processing*, 13(1), 2005.

[95] F. Ribeir, C. Zhang, D. Florencio, and D. Ba. Using reverberation to improve range and elevation discrimination in sound source localization. 18(7), 2010.

[96] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey. Source localization in reverberant environments using sparse optimization. 2013.

[97] J. M. Sachar, H. F. Silverman, and W. R. Patterson. Microphone position and gain calibration for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, 13(1), 2005.

[98] M. R. Schroder and K. H. Kuttruff. On frequency response curves in rooms. comparison of experimental, theoretical and monte carlo results for the average frequency spacing between maxima. *Journal of the Acoustical Society of America*, 76–80, vol. 34, 1962.

[99] Manfred R. Schroeder. Measurement of sound diffusion in reverberation chambers. *Journal of the Acoustical Society of America*, 31(11), 1959, .

[100] Manfred R. Schroeder. The schroeder frequency revisited. *Journal of the Acoustical Society of America*, 99(5), 1997, .

[101] T.J. Schultz. Diffusion in reverberation rooms. *Journal of Sound and Vibration*, 16(1), 1971.

[102] G. A. F. Seber. *Multivariate Observations*. Wiley & Sons, Inc, 2004.

[103] Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9, 2000.

[104] M. L. Seltzer. Microphone array processing for robust speech recognition. In *PhD Thesis, Carnegie Mellon University*, 2001.

[105] Manohar Shamaiah, Siddhartha Banerjee, and Haris Vikalo. Greedy sensor selection: Leveraging submodularity. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 2572–2577. IEEE, 2010.

[106] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz. Localization from mere connectivity. In *MobiHoc*, pages 201–212, 2003.

[107] Y. Shang, W. Ruml, Y. Zhang, and M. P. J. Fromherz. Localization from connectivity in sensor networks. *IEEE Transactions on Parallel Distribute Systems*, 15(11), 2004.

[108] Yuan Shen and Moe Z Win. On the use of multipath geometry for wideband cooperative localization. In *IEEE Global Telecommunications Conference (GLOBECOM)*, 2009.

[109] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei. An integrated framework for multi-channel multi-source localization and voice activity detection. In *IEEE workshop on Hands-free Speech Communication and Microphone Arrays*, 2011.

[110] M. J. Taghizadeh, P. N. Garner, and H. Bourlard. Microphone array beampattern characterization for hands-free speech applications. In *IEEE 7th Sensor Array and Multichannel Signal Processing Workshop*, 2012.

[111] M. J. Taghizadeh, R. Parhizkar, P. N. Garner, and H. Bourlard. Euclidean distance matrix completion for ad-hoc microphone array calibration. In *IEEE 18th International Conference in Digital Signal Processing*, 2013.

[112] M. J. Taghizadeh, A. Asaei, P. N. Garner, and H. Bourlard. Ad hoc microphone array calibration from partial distance measurements. In *Proceedings of Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

[113] Mohammad J. Taghizadeh, Philip N. Garner, and Hervé Bourlard. Enhanced diffuse field model for ad hoc microphone array calibration. *Signal Processing*, 101(0):242 – 255, 2014.

[114] Mohammad J. Taghizadeh, Philip N. Garner, Hervé Bourlard, and Afsaneh Asaei. Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees. *Signal Processing*, 2014.

[115] Mohammadjavad Taghizadeh, Reza Parhizkar, Philip Garner, Hervé Bourlard, and Afsaneh Asaei. Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees. *Signal Processing*, 107:123–140, 2014.

[116] R. Takashima, T. Takiguchi, and Y. Ariki. *Single-Channel Sound Source Localization Based on Discrimination of Acoustic Transfer Functions*. InTech, In book: Advances in Sound Localization, 2011.

[117] Michel Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996.

[118] R. Talmon, I. Cohen, and S. Gannot. Supervised source localization using diffusion kernels.

[119] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(3), 2012.

[120] Giuseppe Valenzise, Luigi Gerosa, Marco Tagliasacchi, E Antonacci, and Augusto Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 21–26, 2007.

[121] J. M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *International Conference on Intelligent Robots and Systems,(IROS)*, 2004.

[122] H. L. Van Trees. *Optimum Array Processing*, chapter 2, pages 17–89. John Wiley & Sons Ltd., 2002.

[123] B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE Signal Processing Magazing*, 5, 1988.

[124] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *Proceedings of ICASSP*, 2001.

[125] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997.

[126] M. Wankling and B. Fazenda. Studies in modal density-its effect at low frequency. In *Proceedings of the Institute of Acoustics*, 2009.

[127] R. V. Waterhouse. Interference patterns in reverberant sound fields. *Journal of the Acoustical Society of America*, 27(2), 1955.

[128] M. Wolfel and J. McDonough. *Distant Speech Recognition*, chapter 13, pages 409–492. John Wiley & Sons Ltd., 2009.

[129] Guanghan Xu, Hui Liu, Lung Tong, and Thomas Kailath. A least-squares approach to blind channel identification. *IEEE Transactions on Signal Processing*, 43:2982–2993, 1995.

[130] Dmitry N Zotkin and Ramani Duraiswami. Accelerated speech source localization via a hierarchical search of steered response power. *IEEE Transactions on Audio, Speech, and Language Processing*, 12(5):499–508, 2004.

[131] Q. Zou, X. Zou, M. Zhang, and Z. Lin. A robust speech detection algorithm in a microphone array teleconferencing system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.