Generalized Sound Field Interpolation in Rotation-robust Microphone Array Signal Processing

Department of Intelligent Systems
Graduate School of Informatics
Nagoya University

Shuming Luan

Contents

Abstract

Auditory information is integral to daily communication for humans and humanoid robots alike. While a range of signal processing techniques has been developed to enhance auditory data acquisition, most state-of-the-art methods rely on a time-invariant acoustic transfer system (ATS) to maintain their performance. In dynamic acoustic environments, however, ATS variability necessitates frequent re-estimation of spatial filters, which imposes significant computational demands and hinders real-time processing. Addressing time-variant ATS challenges is thus a critical step toward enabling robust and practical applications in real-world scenarios.

This thesis investigates array signal processing under dynamic conditions, with a focus on auditory systems equipped with circular microphone arrays (CMAs) mounted on an interactive robot's head for detecting surrounding acoustic signals. In such scenarios, the rotational motion of the CMA introduces ATS variability, necessitating innovative approaches to achieve rotation-robust processing. A beamforming framework named sound field interpolation (SFI) has been proposed to address this challenge. However, SFI requires the CMA to be an equally spaced CMA (es-CMA), where microphones are uniformly distributed with equal angular distances between them. This strict requirement limits its applicability in real-world scenarios, where achieving an es-CMA is impractical. In most cases, the microphones are non-uniformly distributed, resulting in an unequally spaced CMA (unes-CMA). Additionally, for irregularly distributed

microphones, the exact microphone distribution on a unes-CMA is often unknown. In this thesis, the term 'microphone distribution' refers to the spatial configuration of the microphones within the array. Furthermore, in more practical conditions, the interactive robot's head is not a perfect sphere, meaning that even achieving a perfectly circular array is unrealistic, and only a nearly-circular microphone array (NCMA) can be obtained.

Building on the SFI framework, this thesis systematically addresses three major challenges—unes-CMAs, unknown microphone distributions, and NCMAs—step by step, overcoming the inherent limitations of SFI to achieve robust signal interpolation and beamforming.

First, the thesis introduces the unequally spaced sound field interpolation (unes-SFI) method to address the positional deviations of microphones in unes-CMAs. By estimating virtual signals at equally spaced positions, unes-SFI reconstructs before-rotation signals and enables robust beamforming despite array rotation. Detailed analyses and simulations confirm that unes-SFI significantly improves signal reconstruction and beamforming performance under various conditions.

Second, to overcome the limitation of requiring prior knowledge of microphone positions, the thesis presents a novel Generalized Sound Field Interpolation (GSFI) framework. GSFI integrates unes-SFI with unsupervised calibration, employing an iterative optimization technique to estimate microphone positional errors without pre-existing knowledge. This approach enables effective interpolation and beamforming for unes-CMAs with unknown configurations. Simulation results validate GSFI's robustness, demonstrating substantial performance improvements over prior SFI and unes-SFI methods.

Finally, the thesis extends GSFI to address NCMAs, which are prevalent in human-

computer interaction due to the non-spherical shape of the robot's head. By constructing a virtual pseudo-CMA (pCMA) through unsupervised calibration, GSFI reduces spatial complexity and facilitates interpolation for NCMAs. A two-stage strategy is further introduced to handle the positional variability of NCMAs during rotation, ensuring accurate before-rotation signal estimation and robust beamforming. Comprehensive simulations demonstrate that the proposed two-stage GSFI method consistently outperforms previous approaches, achieving satisfactory results across diverse scenarios.

In summary, this thesis provides a stepwise progression of advancements in sound field interpolation for rotation-robust signal processing. Starting from unes-CMAs, extending to unknown microphone distributions, and culminating in NCMAs, it offers a comprehensive framework for addressing the challenges posed by dynamic acoustic environments. The methods presented lay a solid foundation for practical applications in wearable auditory systems, advancing the field of microphone array signal processing.

1 Introduction

1.1 General background

Auditory communication is a fundamental aspect of human interaction that enables the exchange of information, emotions, and intentions through sound. Humans rely heavily on auditory information to navigate complex environments, engage with their surroundings, and communicate effectively. Similarly, in robotics, acoustic data plays a vital role in enabling robots to perceive their environments and interact seamlessly with humans. This has driven significant advancements in array signal processing techniques aimed at optimizing auditory data acquisition, particularly in challenging acoustic scenarios. Among these techniques, source separation and sound enhancement are central areas of focus.

In real-world environments, such as bustling urban settings or crowded rooms, microphones often capture a mixture of sounds, including speech, ambient noise, and other non-target signals. Source separation addresses this challenge by disentangling overlapping sound sources, isolating the desired signals (*e.g.*, a specific speaker's voice), while suppressing irrelevant noise. Sound enhancement further improves the clarity and quality of these isolated signals, reducing residual noise and improving intelligibility. Recent advancements have integrated diverse approaches, such as the alteration of spatial models (*e.g.*, beamforming [1–3]), the use of source models (*e.g.*, independent vector analysis [4–7], nonnegative matrix factorization [8, 9]), the use of a variational

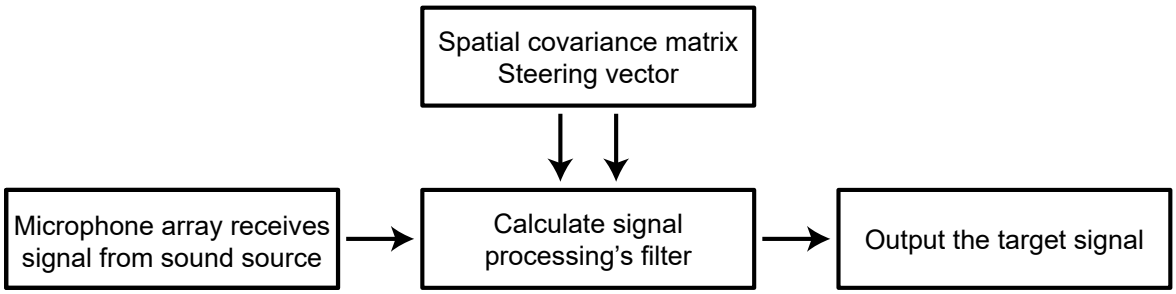

Figure 1.1: *A typical array signal processing framework.*

autoencoder [10, 11] and hybrid techniques (*e.g.*, independent low-rank matrix analysis [12, 13], multichannel variational autoencoders [14–16]). These methods effectively suppress artifacts and amplify key sound features, making them indispensable in applications such as speech recognition, hearing aids, and telecommunications, where high audio fidelity is crucial.

Despite these advancements, the performance of contemporary array signal processing techniques is heavily based on the assumption of a time-invariant acoustic transfer system (ATS). As depicted in Figure 1.1, a typical array signal processing framework involves computing two critical components: the spatial covariance matrix and the steering vector. The spatial covariance matrix captures the correlation between signals across microphone array elements, facilitating accurate estimation of incoming signal directions and enhancing discrimination between sound sources. The steering vector, on the other hand, aligns spatial filters to the target signal direction, ensuring precise selectivity. These calculations, however, demand significant computational resources due to the high dimensionality and complexity of the matrix operations involved.

The ATS—comprising the sound source, microphones, and acoustic transfer functions (ATFs) that define the sound propagation paths—must remain stationary during processing to ensure the stability and effectiveness of spatial filtering algorithms. Any variation in the ATS, such as changes in source positions or microphone configurations,

results in a time-variant ATS, disrupting the spatial information within the array and between the sound source and microphones. This necessitates the re-estimation of spatial filters to maintain optimal performance. However, re-estimation is computationally demanding, particularly in real-time scenarios and high-resolution applications. Such challenges limit the practicality of array signal processing in dynamic environments. Consequently, addressing the obstacles posed by ATS variations has emerged as a critical area of research, with the goal of enabling robust, real-time performance in evolving acoustic settings.

1.2 Thesis Scope

1.2.1 Research Scenario and Previous Work

ATS variation can occur in two primary scenarios: moving sound sources or moving sensors. Addressing challenges associated with moving sources has been the focus of numerous research efforts [17], typically involving several methods, such as adaptive beamforming [18–21], fusion of audio-visual information [22, 23], source localization and tracking [24, 25], deep learning and machine learning-based approach [26, 27], and the combination of these methods [28–32]. When the ATS varies due to source movement, performance degradation can be mitigated through time-block processing. This approach integrates direction-of-arrival (DOA) information estimated by an auxiliary module, tracks multiple sources using these estimates, and subsequently separates them. However, this method introduces certain limitations. If the block length exceeds the time frame of the short-time Fourier transform (STFT), it results in a delay proportional to the block length, which hinders its suitability for real-time processing applications. Additionally, this approach requires the spatial filter to be re-estimated for each

block where DOA changes occur, adding further computational overhead. To address these issues, alternative methods have been proposed. For example, Taseska and Habets [33] introduced an online source separation technique that sequentially estimates the covariance matrix while incorporating DOA information. This method demonstrates potential for reducing delays and improving adaptability in dynamic acoustic environments. Additionally, wavelet-based beamforming [34] is widely applied in scenarios involving moving sources, as it effectively mitigates the adverse effects associated with the time block length in STFT-based beamforming, thereby enhancing real-time processing performance. Chen *et al.* [35] proposed a novel wavelet-based beamforming approach specifically designed to address challenges posed by high-speed rotating sources. Compared to conventional STFT-based beamforming, their method achieves significant improvements in real-time processing. However, in contrast to STFT-based beamforming, which does not require prior information about the source, Chen *et al.*'s approach relies on such prior knowledge to compensate for the Doppler effect, limiting its generalizability.

This thesis primarily focuses on scenarios involving moving sensors, an area that has received limited research attention. Specifically, we examine an auditory system featuring a circular microphone array (CMA) worn on an interactive robot's head. For human-computer interaction in noisy environments, such as social gatherings, robots are often required to receive sound signals from users while moving. The movement of a robot can generally be categorized into translational movement and rotational movement. In this thesis, we focus specifically on rotational movement. Consequently, the CMA rotates along with the robot's head, enabling the system to capture audio signals from the target source while suppressing ambient noise. Our investigation focuses on this typical scenario of CMA rotation, where the array moves but the sound

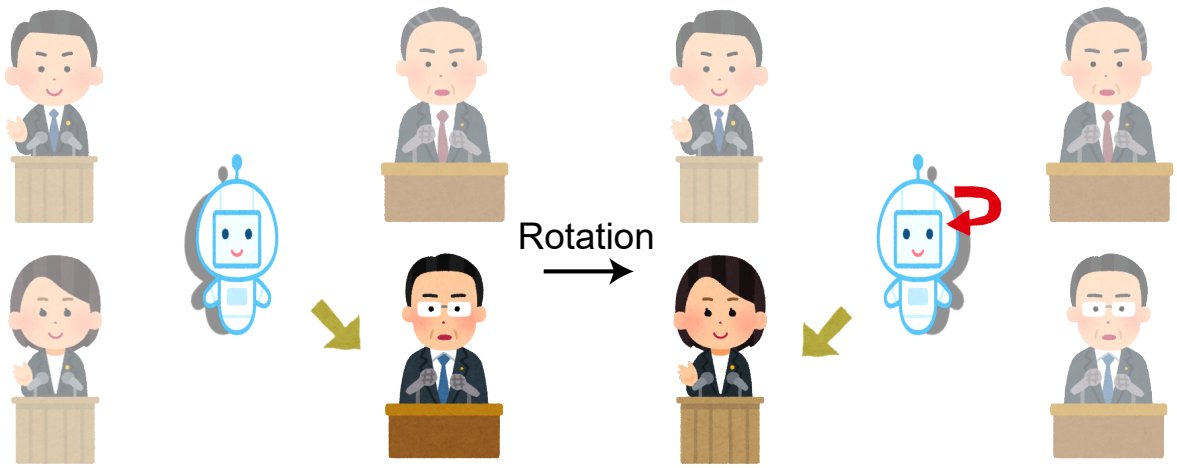

Figure 1.2: *A situation where the CMA rotates but the sound sources remain stationary.*

sources remain stationary. An example of this scenario is illustrated in Figure 1.2. The rotation of the CMA introduces variability in the ATS, shifting it from a time-invariant to a time-variant state. This variation necessitates the re-estimation of spatial filters, a process that, as discussed in Section 1.1, involves computationally demanding operations. These challenges significantly hinder the feasibility of real-time processing in dynamic environments, making CMA rotation a critical problem to address for practical applications.

To address the challenges posed by ATS variability, one proposed strategy is the "stop–perceive–act" principle [36], which involves halting movement temporarily to stabilize the ATS and facilitate accurate signal processing. This approach aims to approximate a time-invariant ATS, simplifying the application of array signal processing techniques. However, this method is impractical in real-world scenarios, as robot movement is typically continuous. Such constraints may disrupt natural interactions with the environment and prevent the microphone array from capturing new signals during motion. If movement is relatively slow, the time-variant ATS can be approximated as time-invariant, allowing the "stop–perceive–act" principle to be fulfilled.

While this approximation can yield some effectiveness, it remains vulnerable to performance degradation, similar to the limitations of time-block processing. Continuous movement, even at a slow pace, perpetuates ATS variability, requiring spatial filters to be updated in real time to maintain optimal performance. Thus, slow movement still poses significant challenges to real-time processing in dynamic environments.

Tourbabin and Rafaely [37] proposed a novel technique for DOA estimation tailored to microphone arrays mounted on moving humanoid robots. This approach compensates for the robot's motion using a motion compensation matrix in the spherical harmonic domain and constructs the rotation matrix with the Wigner-D matrix [38, 39]. In a related study, Ma *et al.* [40] explored two methods for estimating sound signals at virtual rotating array (VRA) microphones: the cross-spectral matrix based on modal decomposition (CSM-MD) and the cross-spectral matrix based on linear interpolation (CSM-LI). The CSM-MD method employs Fourier interpolation to process sound pressures recorded by all real microphones, whereas the CSM-LI method uses linear interpolation between two neighboring microphones in the time domain. Additionally, Casebeer *et al.* [41] introduced a learning-based approach using a recurrent neural network (RNN) to estimate time-varying spatial covariance matrices. This method specifically addresses the challenges posed by rapid pose changes in wearable devices, demonstrating its potential for handling dynamic scenarios effectively.

To address the challenges of online processing with moving sensors, Wakabayashi *et al.* proposed an innovative beamforming framework [42, 43] to mitigate the effects of CMA rotation. This technique employs sound field interpolation (SFI), based on the theorem of non-integer sample shifts, leveraging the periodicity of the sound field around the CMA's circumference and the relationship between array perception and sound field discretization. When the CMA rotates to a new position, the newly

recorded sound signals at that position are used to perform SFI to estimate what the signals would have been at the original position before rotation. This process allows the rotated CMA to be treated as fixed, eliminating the need for filter updates. The previously computed beamformer filters can then be directly applied regardless of the array's rotation, significantly simplifying processing. The SFI method has shown potential for wide-ranging applications and serves as a foundation for developing derivative algorithms. For example, Nakashima *et al.* [44] applied SFI to propose an online-independent vector analysis method that remains robust against CMA self-rotation. Similarly, Lian *et al.* [45] utilized SFI for precise self-rotation angle estimation of CMAs. While SFI shares similarities with the CSM-MD method in [40], as both employ linear interpolation in the Fourier domain for all microphones, Wakabayashi *et al.* introduced a clearer matrix-form analytical expression for interpolation, enhancing its usability and analytical rigor.

1.2.2 Research Question

It is crucial to recognize that both SFI and CSM-MD methods rely on the assumption of an equally spaced CMA (es-CMA) to effectively capture and interpret the sound field, owing to utilizing the periodicity of sound pressure around the array's circumference. Any angular deviation from these equally spaced positions can significantly degrade the performance of these methods, as they rely on the uniformity of the array for reliable sound field estimation. In practical applications, achieving a perfectly uniform distribution of microphones on a CMA is often impractical and unnecessary due to the limitations of design, user behavior, and physical constraints. Here, this thesis uses the term 'distribution' to refer to the spatial arrangement of microphones in an array. Consequently, in many real-world scenarios, an unequally spaced CMA (unes-CMA)—

characterized by irregular angular intervals between adjacent microphones—is more commonly used, which introduces notable implementation challenges. The CSM-LI method offers an advantage in that it can be applied to unes-CMAs by using linear interpolation between only two adjacent microphones in the time domain. However, this method has lower spectral reconstruction capabilities [40] compared to SFI and CSM-MD, resulting in suboptimal source enhancement performance. This gap in performance underscores the need to develop more robust techniques for estimating sound signals on unes-CMAs. To that end, the first research question we seek to address is: **How can the signal prior to rotation be accurately estimated on an unes-CMA to eliminate the effects of rotation?**

To answer this question, a critical consideration is whether the distribution of microphones on the unes-CMA is known in advance. If prior knowledge of the microphone distribution is available, it would simplify the task of compensating for the rotation. However, to avoid imposing additional constraints and to enhance the robustness, it is desirable to eliminate the need for any prior information. Therefore, we consider the scenario in which the specific distribution of the microphones is unknown—–a situation that aligns well with practical applications, where irregular microphone placement often results in uncertainty about their exact positions. This leads us to the second research question: **How can the array signal prior to rotation be estimated when the microphone distribution on the unes-CMA is unknown?**

Moreover, while we have discussed the unes-CMA in the context of a circular array, it is important to consider that the robot head is not a perfect sphere. As a result, a microphone array worn on the head may not maintain a strictly circular shape, but instead form a NCMA, where the microphones are positioned in irregular patterns, just as Figure 1.3 shows. These variations in shape and microphone placement introduce

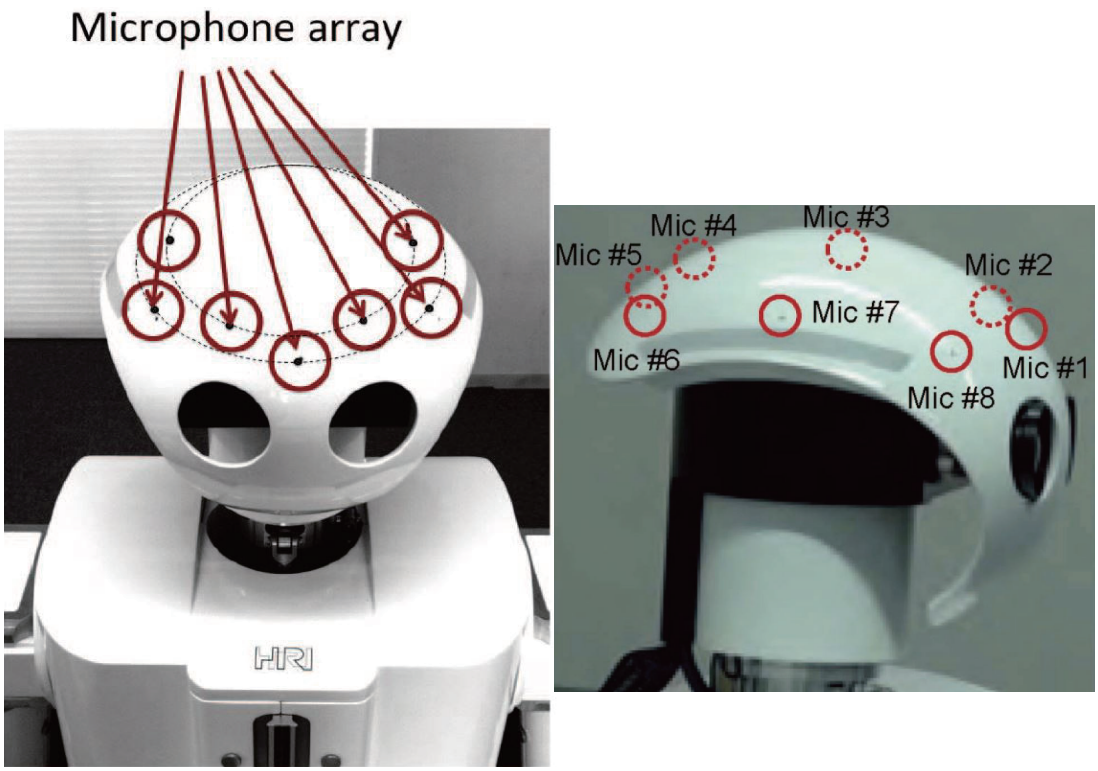

Figure 1.3: *Examples of an NCMA on an interactive robot's head.*

further complexity in signal estimation and beamforming processes. This leads to our third research question: **How can the signal before rotation be estimated on an NCMA, where the array shape deviates from a perfect circle?**

These research questions address the core challenges of designing robust array signal processing techniques for dynamic, wearable systems where CMA rotation, irregular microphone distribution, and nearly-circular array shapes must be accounted for. By investigating these problems, we aim to develop methods that can provide accurate, real-time signal processing in a range of practical applications, from hearing aids to advanced robotics.

1.3 Thesis Overview

This thesis aims to address the three key challenges outlined in Section 1.2. These challenges encompass the accurate estimation of signals on an unes-CMA, the development of methods to handle unknown microphone distributions, and the adaptation of techniques to accommodate NCMAs. Each of these problems presents unique obstacles to achieving robust and efficient sound field processing on a wearable hearing augmentation system. Figure 1.4 provides an overview of the thesis's scope, illustrating the interconnections between these challenges and the proposed solutions.

Chapter 2 lays the foundation for this thesis by presenting the fundamentals of beamforming and SFI. The primary objective of this thesis is to propose a novel preprocessing method for array signal processing that effectively mitigates the effects of CMA rotation, eliminating the need for re-estimating the spatial filter, a process that is computationally intensive and challenging in dynamic environments. To achieve this goal, a comprehensive understanding of mainstream array signal processing is crucial, as it serves as the baseline for the proposed enhancements. In this thesis, beamforming is taken as a representative example of the mainstream array signal processing techniques. Additionally, because the methods introduced in this thesis build upon the principles of SFI, a detailed and systematic overview of SFI will also be provided in this chapter. This foundational knowledge is critical for understanding the proposed advancements.

To address the first problem outlined in Section 1.2, Chapter 1.2 introduces an enhanced method called unequally spaced SFI (unes-SFI) to effectively manage the rotation challenges associated with an unes-CMA. In this framework, it is assumed that the microphone distribution on the unes-CMA is known beforehand. The proposed method begins by compensating for the non-uniform microphone distribution to standardize the array configuration before addressing variations in the ATS. Using the signals cap-

tured by the unes-CMA, the method estimates the hypothetical sound signals that would have been recorded if the microphones were positioned at their corresponding equally spaced locations. By doing so, the unes-CMA is conceptually transformed into a virtual es-CMA. This transformation simplifies the complex rotation issue of the unes-CMA into the more manageable rotation problem of an es-CMA, for which solutions have been previously established in [42, 43]. By leveraging the sound signals from the virtual es-CMA prior to rotation, unes-SFI generates the sound signals that would have been recorded by the actual unes-CMA before rotation. As a result, the time-variant ATS of the unes-CMA is successfully converted into a time-invariant ATS of an es-CMA, laying a strong foundation for robust signal processing while preserving computational efficiency.

To address the second problem outlined in Section 1.2, Chapter 4 presents a further enhanced rotation-robust beamforming method called generalized sound field interpolation (GSFI). Unlike the unes-SFI method introduced in Chapter 3, which requires prior knowledge of each microphone's angular displacement from its ideal equally spaced position to estimate the hypothetical sound signals on a virtual es-CMA and compensate for the non-uniform distribution, GSFI is designed to work with freely distributed unes-CMAs even when such prior information is unavailable. GSFI incorporates an unsupervised calibration method to estimate the positions of microphones on the unes-CMA through an iterative optimization process without any pre-existing location data. Crucially, this calibration process relies exclusively on the multichannel signals captured by the array and needs to be performed only once. Once the microphone positions are calibrated, the problem reduces to a form that can be addressed using the unes-SFI framework presented in Chapter 3. By providing calibrated positions, the unsupervised calibration ensures that unes-SFI can effectively compensate

for the non-uniform microphone distribution. Thus, GSFI is realized by combining unsupervised calibration with unes-SFI. This combination enables GSFI to function as a robust preprocessing step before applying beamforming, offering a robust solution to the challenges posed by unknown microphone distributions.

To address the third problem identified in Section 1.2, Chapter 5 extends the applicability of GSFI to handle NCMAs, which deviate from the standard circular geometry, and presents a reliable framework for signal estimation in such scenarios. The method begins by radially repositioning the microphones on the NCMA onto a common circle to form a pCMA. Through unsupervised calibration using the NCMA's signal, the distribution of microphones on the pCMA is determined to ensure that its received signal closely approximate that of the original NCMA. The pCMA then serves as a proxy for the NCMA, reducing the complexity of processing signals from the NCMA by leveraging the simpler circular geometry of the pCMA. The goal is to estimate the NCMA signal before rotation using the signal recorded after rotation. However, since rotation alters the correspondence between the pCMA and NCMA, a single pCMA cannot reliably represent the NCMA for both states. Each rotation angle of the NCMA produces a unique distribution of microphones on the pCMA. To address this, the framework employs a two-stage approach. The before-rotation and after-rotation states are treated as distinct stages, with separate pCMAs representing the NCMA in each stage. The method establishes correspondence between the two pCMAs and leverages the after-rotation NCMA signal to estimate the before-rotation signal. This extended framework enhances GSFI by making it adaptable to NCMAs, thereby broadening its utility to scenarios involving head-wearable arrays that deviate from strict circular geometries.

Table 1.1 provides a comparative summary of the SFI method introduced in Chapter 2 alongside the novel methods proposed in Chapters 3 to 5, elucidating the relation-

Table 1.1: *Comparative summary of the different interpolation methods.*

Requirements	Interpolation Methods			
	Sound Field Interpolation	Unequally Spaced Sound Field Interpolation	Generalized Sound Field Interpolation	Two-stage Generalized Sound Field Interpolation
Time-invariant Acoustic Transfer System	Unnecessary	Unnecessary	Unnecessary	Unnecessary
Uniformly Distributed Microphones along the Circle	Necessary	Unnecessary	Unnecessary	Unnecessary
Known Distribution of Microphones	Necessary	Necessary	Unnecessary	Unnecessary
All Microphones Positioned on the Same Circle	Necessary	Necessary	Necessary	Unnecessary

ships and advancements made throughout this thesis. This structured comparison highlights the progressive refinement and expansion of the methodologies, showcasing how each method systematically addresses and removes key assumptions and constraints. This progression reflects a deliberate strategy to enhance the robustness and versatility of the proposed approaches, ultimately enabling their application to increasingly complex and realistic scenarios. By incrementally building upon the foundational concepts introduced in Chapter 2, the thesis demonstrates a cohesive and systematic evolution of techniques to address challenges associated with CMA's rotation, unes-CMAs, and NCMAs.

Finally, Chapter 6 concludes the thesis by summarizing the key findings, contributions, and implications of the thesis. This chapter also discusses the limitations of the proposed methods and suggests potential directions for future work, highlighting opportunities for further improving the robustness and applicability of the techniques developed in this thesis.

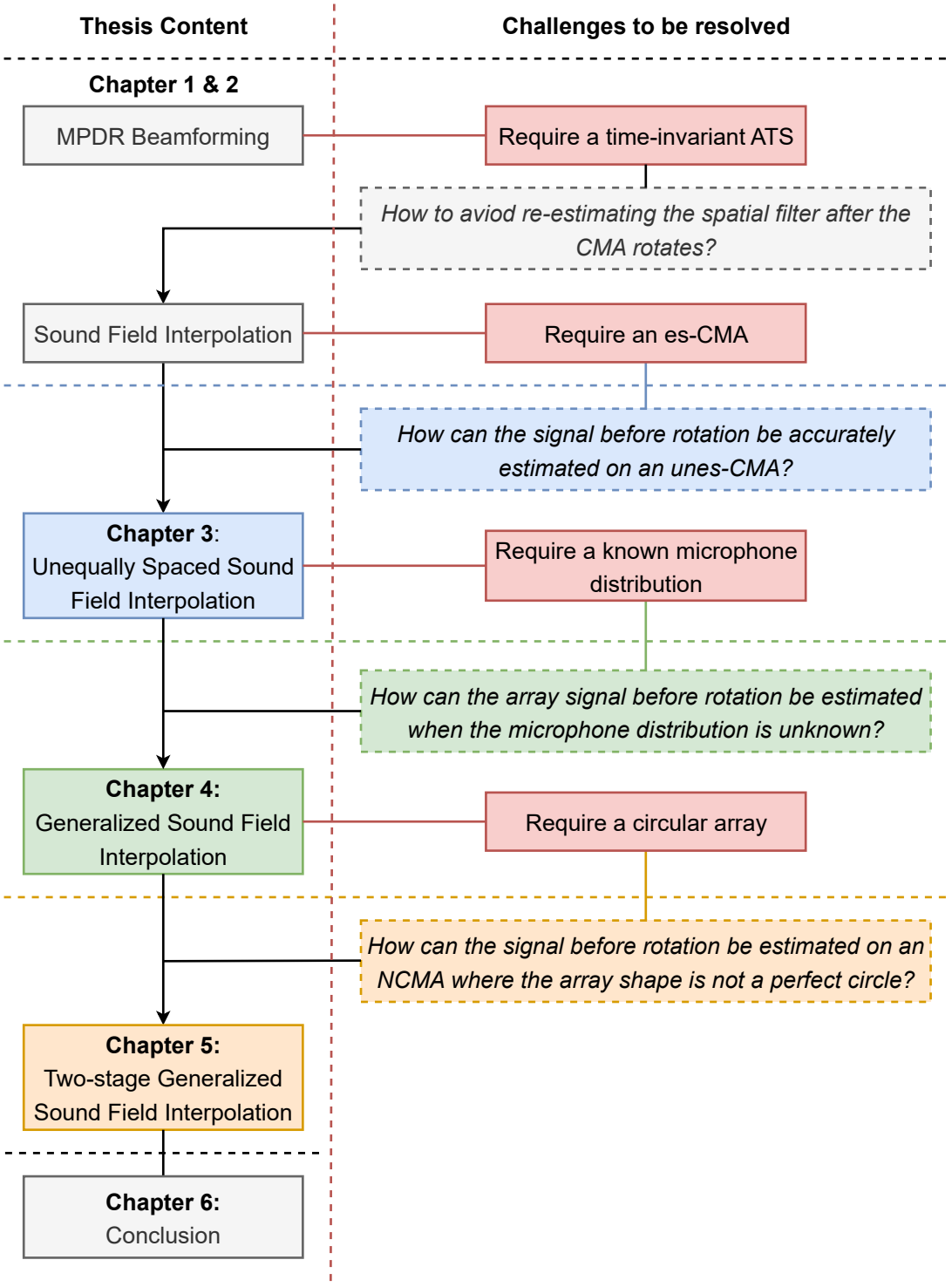

Figure 1.4: *Thesis's scope.*

2 Background and related Work

This chapter introduces the foundational principles of beamforming and sound field interpolation (SFI), which are closely related to the research presented in this thesis.

2.1 Introduction

As highlighted in Chapter 1, the primary objective of this thesis is to develop a method that enhances the robustness of microphone array signal processing against rotational movements, thereby eliminating the need to re-estimate spatial filters caused by such motion. Achieving this objective is critical for enabling real-time processing in practical applications. Although there are various types of microphone array signal processing, such as sound source localization and acoustic echo cancellation, this thesis focuses exclusively on beamforming. Beamforming is used as a representative example among microphone array signal processing methods to explore how rotation-robust beamforming can be achieved. Thus, an understanding of beamforming fundamentals is essential, as this forms the foundation of the research presented in this thesis.

Wakabayashi *et al.* previously introduced a novel framework leveraging SFI for microphone array signal processing robust to the rotation of a circular microphone array (CMA). This framework was evaluated in beamforming applications. Based on the sampling theorem on the circle, SFI enables the estimation of microphone signals at their original positions prior to rotation, allowing conventional array signal processing

methods to operate without re-estimation. However, this framework assumes the use of an equally spaced CMA (es-CMA) to ensure the discretized sound field exhibits periodicity. The methods developed in this thesis build upon the SFI concept, introducing unequally spaced SFI (unes-SFI) and generalized SFI (GSFI). These extensions generalize interpolation techniques to accommodate unequally spaced circular microphone arrays (unes-CMAs) and nearly-circular microphone arrays (NCMAs), addressing more complex and practical scenarios. Consequently, a thorough understanding of SFI principles is vital for appreciating the contributions of this thesis.

This chapter is structured as follows: Section 2.2 introduces the widely used Minimum Power Distortionless Response (MPDR) beamforming framework [46–54], providing essential context for understanding the integration of beamforming into rotation-robust processing. Section 2.3 delves into prior work on SFI, laying the foundation for the advanced methods proposed in subsequent chapters.

2.2 Minimum power distortionless response beamforming

Beamforming is a signal processing technique used to enhance the desired signal while suppressing interference by steering the mainlobe of the array's response toward the target direction. Among various beamforming methods, the MPDR algorithm is a widely adopted approach. MPDR minimizes interference and noise by adjusting the spatial filter to achieve the minimum output power, ensuring an undistorted response in the desired direction. This section provides an introduction to the signal model used in beamforming, which serves as the basis for spatial filtering. It then details the formulation of the MPDR beamforming algorithm.

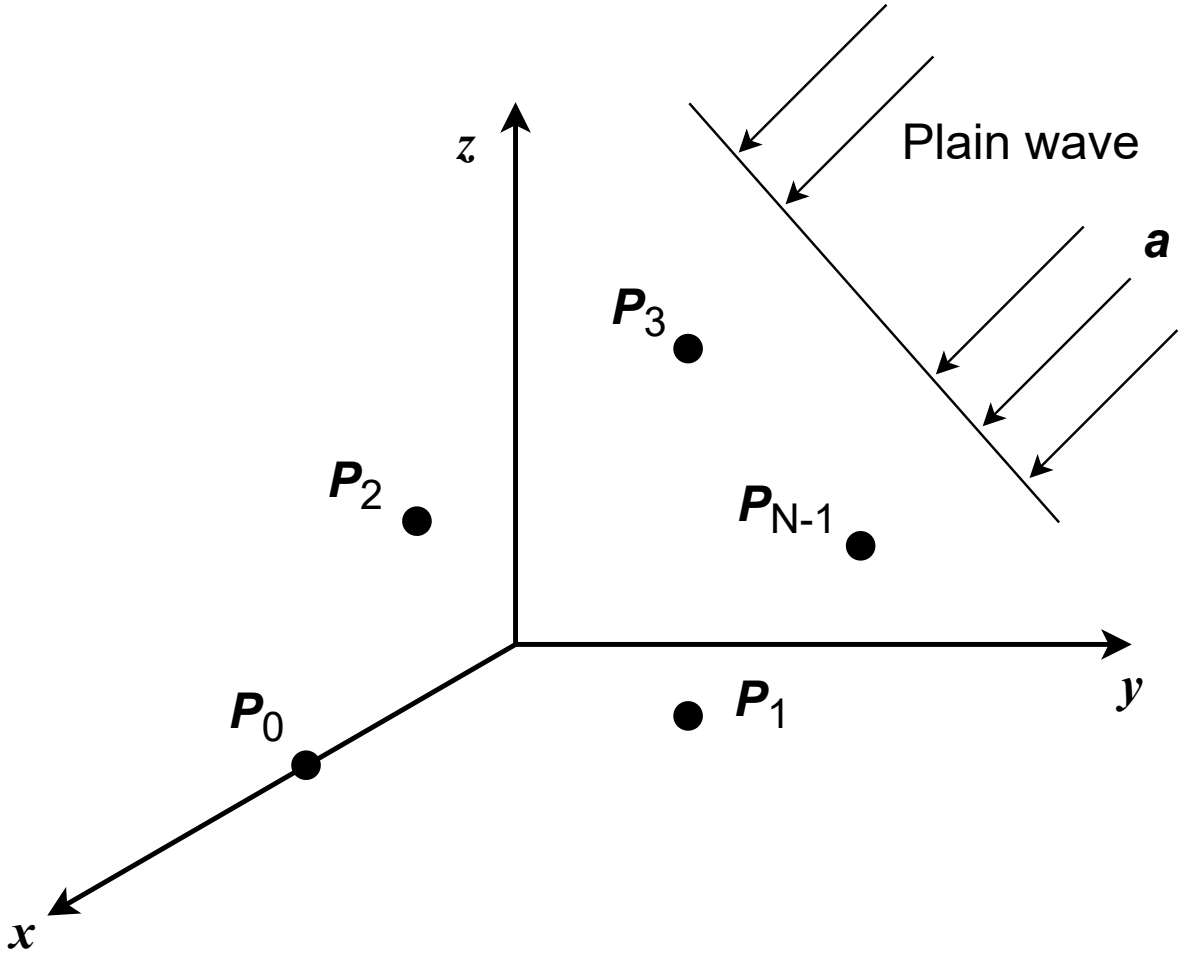

Figure 2.1: *Array with plane-wave input.*

2.2.1 Beamforming signal model

We consider the response of a microphone array to a plane wave propagating in the direction of the unit vector $\boldsymbol{a} \in \mathbb{R}^{3\times 1}$, characterized by a temporal (radian) frequency ω, in a noise-free environment. The array comprises N isotropic sensors positioned at positions $\boldsymbol{P}_n \in \mathbb{R}^{3\times 1}$, where $n = 0, 1, ..., N-1$, as depicted in Figure 2.1. These sensors act as spatial samplers of the signal field, capturing the wavefront at their respective locations $\boldsymbol{P}_n$. The resulting set of sampled signals is represented collectively by the

vector

$$\boldsymbol{f}(t, \boldsymbol{P}) = \begin{bmatrix} f(t, \boldsymbol{P}_0) \\ f(t, \boldsymbol{P}_1) \\ \vdots \\ f(t, \boldsymbol{P}_{N-1}) \end{bmatrix}. \tag{2.1}$$

Alternatively, the time-domain representation in (2.1) can be transformed into the frequency domain as

$$\boldsymbol{F}(\omega, \boldsymbol{P}) = \int_{-\infty}^{\infty} \boldsymbol{f}(t, \boldsymbol{P}) e^{-j\omega t}\, dt, \tag{2.2}$$

In most cases, the explicit dependence on $\boldsymbol{P}$ on the left-hand side of (2.2) is omitted for simplicity, and the frequency domain representation is denoted as $\boldsymbol{F}(\omega) \in \mathbb{C}^{N\times 1}$.

Considering the case shown in Figure 2.1, we illustrate a simple beamforming operation by expressing the time-domain signals received by the sensors in a manner that highlights the time delays associated with the signal's time of arrival at each sensor. Let the signal that would be received at the origin of the coordinate system in Figure 2.1 be $f(t)$, then (2.1) can be written as

$$\boldsymbol{f}(t, \boldsymbol{P}) = \begin{bmatrix} f(t - \tau_0) \\ f(t - \tau_1) \\ \vdots \\ f(t - \tau_{N-1}) \end{bmatrix}, \tag{2.3}$$

where

$$\tau_n = \frac{\boldsymbol{a}^\top \boldsymbol{P}_n}{c}, \tag{2.4}$$

and c is the velocity of propagation in the medium. From (2.3), the nth component of $\boldsymbol{F}(\omega)$ can be calculated using the time-shifting property of Fourier transform

$$F_n(\omega) = \int_{-\infty}^{\infty} f(t - \tau_n) e^{-j\omega t}\, dt = F(\omega) e^{-j\omega\tau_n}. \tag{2.5}$$

$F(\omega)$ represents the Fourier transform of $f(t)$, which is the signal expressed in the frequency domain, and

$$\omega\tau_n = \frac{\omega}{c}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}_n = \frac{2\pi}{\lambda}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}_n, \tag{2.6}$$

where λ is the wavelength associated with the frequency ω. Defining

$$\boldsymbol{v}(\boldsymbol{a}) = \begin{bmatrix} e^{-j2\pi\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}_0/\lambda} \\ e^{-j2\pi\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}_1/\lambda} \\ \vdots \\ e^{-j2\pi\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}_{N-1}/\lambda} \end{bmatrix}, \tag{2.7}$$

according to (2.5), we can write $\boldsymbol{F}(\omega)$ as

$$\boldsymbol{F}(\omega) = F(\omega)\boldsymbol{v}(\boldsymbol{a}). \tag{2.8}$$

The vector $\boldsymbol{v}(\boldsymbol{a})$ encapsulates the spatial characteristics of the microphone array, including the positions of the sensors and their relative orientation with respect to the incoming wavefront. Its role is central to both formulating and implementing beamforming algorithms.

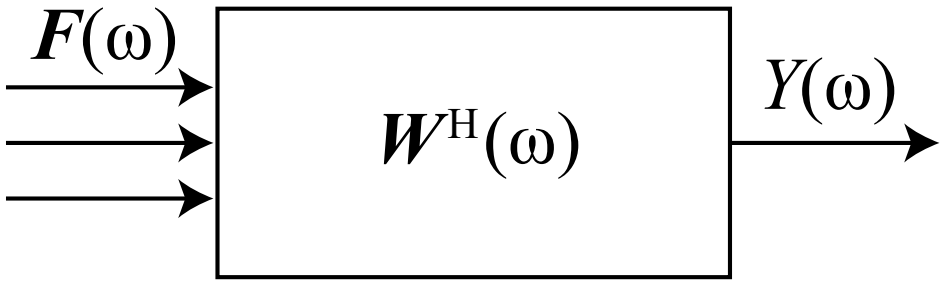

Figure 2.2: *Matrix operation in MPDR.*

2.2.2 Formulation of MPDR beamforming

In this subsection, we focus on the scenario involving a single plane-wave signal. The frequency-domain representation of the array signal, $\boldsymbol{F}(\omega)$, includes two components: the received source signal $\boldsymbol{F}_s(\omega) \in \mathbb{C}^{N\times 1}$ and an additive noise component $\boldsymbol{N}(\omega) \in \mathbb{C}^{N\times 1}$. Mathematically, the array signal can be expressed as

$$\boldsymbol{F}(\omega) = \boldsymbol{F}_s(\omega) + \boldsymbol{N}(\omega), \tag{2.9}$$

The source signal filtered by the array, $\boldsymbol{F}_s(\omega)$, can be written as

$$\boldsymbol{F}_s(\omega) = F_s(\omega)\boldsymbol{v}(\boldsymbol{a}), \tag{2.10}$$

which is identical to (2.8). Here, $F_s(\omega)$ represents the single source signal input to the array in the frequency domain. The parameter $\boldsymbol{a}$ specifies the direction from which the single plane wave arrives. The term $\boldsymbol{v}(\boldsymbol{a})$ is referred to as the steering vector associated with the direction $\boldsymbol{a}$.

In MPDR, the goal is to isolate the desired source signal $F_s(\omega)$ while minimizing the impact of the noise $\boldsymbol{N}(\omega)$. To achieve this, the observed array signal $\boldsymbol{F}(\omega)$ is processed using a spatial filter represented by the matrix operation $\boldsymbol{W}^{\mathsf{H}}(\omega) \in \mathbb{C}^{1\times N}$, as depicted

in Figure 2.2,

$$Y(\omega) = \boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{F}(\omega) = \boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{v}(\boldsymbol{a})F_s(\omega) + \boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{N}(\omega), \tag{2.11}$$

where H denotes the Hermitian transpose. The spatial filter $\boldsymbol{W}(\omega)$ combines the outputs of the individual sensors, weighted appropriately, to form a scalar output signal.

A key principle guiding MPDR beamforming is the distortionless criterion, which ensures that the desired signal remains unaffected in amplitude and phase after processing. It is required that, in the absence of noise,

$$Y(\omega) = F_s(\omega), \tag{2.12}$$

for any $F_s(\omega)$. This constraint of no distortion implies

$$\boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{v}(\boldsymbol{a}) = 1. \tag{2.13}$$

The distortionless criterion guarantees that the system maintains a gain of 1 for the desired direction while suppressing unwanted noise and interference from other directions. This balance between signal preservation and noise minimization is central to the MPDR approach, enabling effective beamforming in noisy environments.

The choice of $\boldsymbol{W}(\omega)$ is critical in MPDR. Under the distortionless criterion, $\boldsymbol{W}(\omega)$ is designed to achieve a distortionless response in the desired direction while minimizing the total power of the output signal $Y(\omega)$, which includes contributions from noise and interference. In other words, we wish to minimize the mean square of $Y(\omega)$, which is

$$\mathbb{E}\left[\left|Y(\omega)^2\right|\right] = \boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{S}(\omega)\boldsymbol{W}(\omega), \tag{2.14}$$

where $\boldsymbol{S}(\omega) = \mathbb{E}\left[\boldsymbol{F}(\omega)\boldsymbol{F}(\omega)^{\mathsf{H}}\right]$ is the spatial covariance matrix of $\boldsymbol{F}(\omega)$. We aim to

minimize $\mathbb{E}\left[|Y(\omega)^2|\right]$ while ensuring that the distortionless criterion (2.13) is satisfied [55, 56].

The solution to this constrained optimization problem is derived using the method of Lagrange multipliers. The objective function that we minimize is

$$\mathcal{F} \stackrel{\text{def}}{=} \boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{S}(\omega)\boldsymbol{W}(\omega)+\lambda(\omega)\left[\boldsymbol{W}^{\mathsf{H}}(\omega)\boldsymbol{v}(\boldsymbol{a})-1\right]+\lambda^{*}(\omega)\left[\boldsymbol{v}^{\mathsf{H}}(\boldsymbol{a})\boldsymbol{W}(\omega)-1\right]. \quad (2.15)$$

We take the complex gradient with respect to $\boldsymbol{W}^{\mathsf{H}}(\omega)$ and solve for the optimal weights

$$\boldsymbol{W}^{\mathsf{H}}(\omega) = -\lambda(\omega)\boldsymbol{v}^{\mathsf{H}}(\boldsymbol{a})\boldsymbol{S}^{-1}(\omega). \quad (2.16)$$

To determine the value of $\lambda(\omega)$, we apply the constraint in (2.13), which gives the following form

$$\lambda(\omega) = -\left[\boldsymbol{v}^{\mathsf{H}}(\boldsymbol{a})\boldsymbol{S}^{-1}(\omega)\boldsymbol{v}(\boldsymbol{a})\right]^{-1}. \quad (2.17)$$

As a result, $\boldsymbol{W}(\omega)$ is given by

$$\boldsymbol{W}^{\mathsf{H}}(\omega) = \frac{\boldsymbol{v}^{\mathsf{H}}(\boldsymbol{a})\boldsymbol{S}^{-1}(\omega)}{\boldsymbol{v}^{\mathsf{H}}(\boldsymbol{a})\boldsymbol{S}^{-1}(\omega)\boldsymbol{v}(\boldsymbol{a})}. \quad (2.18)$$

The matrix processor in (2.18) is referred to as the MPDR beamforming spatial filter. For notational simplicity, it is convenient to suppress the frequency ω and the direction $\boldsymbol{a}$, resulting in the following expression for the MPDR beamforming spatial filter

$$\boldsymbol{W}^{\mathsf{H}} = \frac{\boldsymbol{v}^{\mathsf{H}}\boldsymbol{S}^{-1}}{\boldsymbol{v}^{\mathsf{H}}\boldsymbol{S}^{-1}\boldsymbol{v}}. \quad (2.19)$$

As seen in (2.19), the spatial covariance matrix $\boldsymbol{S}$ and the steering vector $\boldsymbol{v}$ are critical for the calculation of the beamforming filter $\boldsymbol{W}$.

When the microphone array undergoes rotation, the spatial information changes, necessitating the recalculation of $\boldsymbol{S}$ and $\boldsymbol{v}$ to derive a new beamforming filter. However, the computation of $\boldsymbol{S}$ and $\boldsymbol{v}$ involves significant computational complexity, making it unsuitable for real-time processing applications. Therefore, it is essential to develop methods that avoid recalculating $\boldsymbol{W}$, which would improve the processing speed of the beamforming processing.

2.3 Sound field interpolation (SFI)

In this section, we review previous research on SFI [42,43] for addressing time-variant ATS utilizing an es-CMA. The discussion begins with the derivation process of the SFI technique, providing the theoretical foundation for its operation. Following this, we present a concise analysis of the singularity inherent in the rotation transformation matrix, a key component in modeling the rotation of the es-CMA. This matrix is critical for accurately generating the sound signals corresponding to their before-rotation positions. Finally, we illustrate the application of SFI to beamforming as an example, illustrating its utility in multichannel signal processing through both batch and online methodologies.

2.3.1 Formulation

It is important to emphasize that this earlier research has primarily focused on eliminating the need to update the spatial filter when the CMA undergoes rotation. Consequently, the spatial filter, represented by $\boldsymbol{W}$ in (2.11) and (2.19), is predetermined and remains constant throughout the process. By estimating the signal before rotation from the after-rotation observations, this method allows the direct application of the

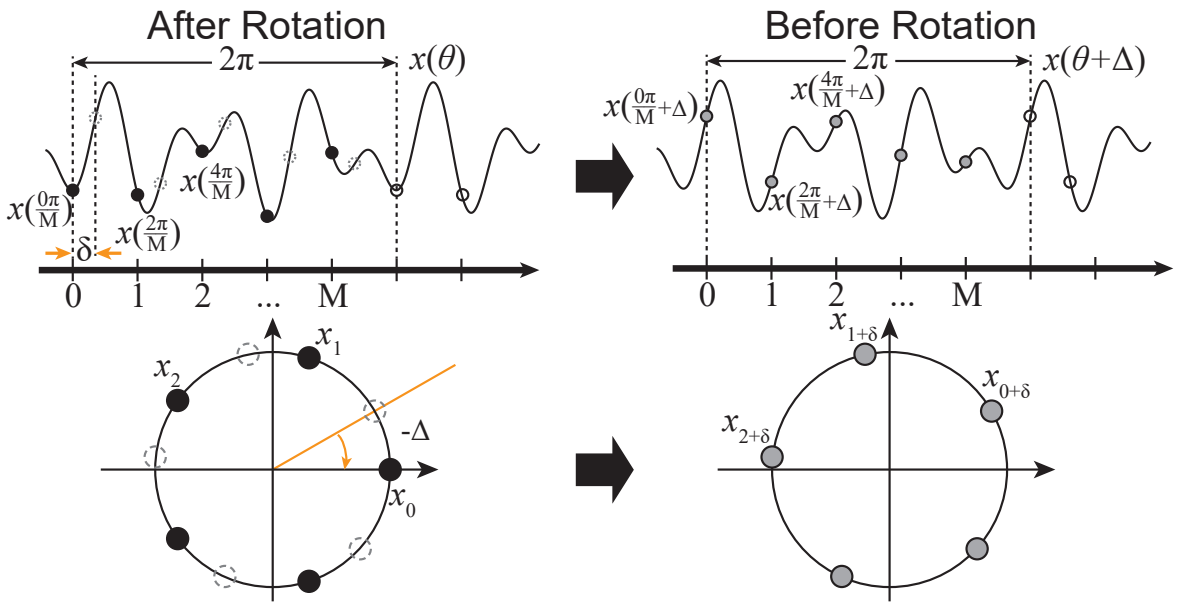

Figure 2.3: *Continuous sound field on a circle's circumference and the discretized sound field function with a δ sample shift.*

pre-existing spatial filter $\boldsymbol{W}$, thereby avoiding the computationally intensive process of re-estimation.

We consider a continuous sound field function $x(\theta)$ in the time–frequency domain, observed after the CMA undergoes a rotation of $-\Delta$ radians, as illustrated in Figure 2.3. The function $x(\theta)$ can be generally expressed as a sum of sinusoidal and cosinusoidal components using a Fourier series representation. Importantly, this function exhibits periodicity with a period of 2π, wherein $\theta \in [0, 2\pi)$ represents the angular position on the circle.

When an es-CMA is employed to capture the sound field, the continuous sound field function $x(\theta)$ is discretized by sampling it at discrete angular positions corresponding to the locations of the microphones. To preserve the periodicity of the discretized sound field, the microphones must be positioned at equidistant intervals along the circumference of the array. For an M-channel es-CMA, the interval between adjacent

microphones is $2\pi/M$, and the observed signal at the mth microphone is expressed as:

$$x_m = x\left(2\pi\frac{m}{M}\right),\ m = 0, ..., M-1, \tag{2.20}$$

where x_m represents the signal observed at the mth microphone after the rotation of the CMA.

It is important to note that a key advantage of this approach is its independence from any specific assumptions about the signal model. The observed signals x_m can represent any type of sound field in any acoustic environment, making the framework broadly applicable across diverse scenarios.

Assuming that the sampling theorem [57] holds, the continuous sound field function $x(\theta)$ can be reliably reconstructed from the discretized sound signal x_m. This capability underpins the feasibility of SFI, which leverages the non-integer sample shift theorem in the Fourier domain to achieve accurate reconstruction and manipulation of sound field. Specifically, the sound field before rotation, denoted by a discretized Δ-rad-rotated sound field function $x(2\pi m/M+\Delta)$, aligns consistently with the δ-sample-shifted sound signal $x_{m+\delta}$ observed using an es-CMA. Mathematically, this relationship is expressed as:

$$x_{m+\delta} = x\left(\frac{2\pi m}{M} + \Delta\right), \tag{2.21}$$

where the shift δ is given by

$$\delta = \frac{M\Delta}{2\pi}. \tag{2.22}$$

Using the non-integer sample shift theorem in the discrete Fourier transform (DFT) [58, 59], the shifted signal $x_{m+\delta}$ can be reconstructed in terms of the original discrete

sound signal $x_0, x_1, \ldots, x_{M-1}$. This reconstruction is expressed as:

$$x_{m+\delta} = \frac{1}{M} \sum_{k=-M/2+1}^{M/2} \left(\mathscr{F}_{\mathrm{D}} \left[x_m \right] e^{j\Delta k} \right) e^{j\frac{2\pi mk}{M}}, \tag{2.23}$$

or equivalently:

$$x_{m+\delta} = \sum_{n=0}^{M-1} x_n U_{m,n,\delta}, \tag{2.24}$$

where $\mathscr{F}_{\mathrm{D}} \left[x_m \right]$ represents the DFT of the discrete signal x_m. $U_{m,n,\delta}$ is the interpolation coefficient of SFI, which is computed by applying the *sinc* function, as demonstrated below:

$$U_{m,n,\delta} = \begin{cases} \frac{1-e^{jL\pi}}{M} + \frac{\mathrm{sinc}\left(\frac{L}{2}\right)\cos\left(\frac{M+2}{2M}L\pi\right)}{\mathrm{sinc}\left(\frac{L}{M}\right)}, & M \text{ is even}^1 \\ \\ \frac{1}{M} + \frac{M-1}{M}\frac{\mathrm{sinc}\left(\frac{L(M-1)}{2M}\right)\cos\left(\frac{M+1}{2M}L\pi\right)}{\mathrm{sinc}\left(\frac{L}{M}\right)}, & M \text{ is odd,} \end{cases} \tag{2.25}$$

where $L = n - m - \delta$ and $j = \sqrt{-1}$. According to Euler's formula, (2.25) in the case of an even M can also be written as

$$\mathrm{Re}\left(U_{m,n,\delta}\right) = \frac{1 - \cos L\pi}{M} + \frac{\mathrm{sinc}\left(\frac{L}{2}\right)\cos\left(\frac{M+2}{2M}L\pi\right)}{\mathrm{sinc}\left(\frac{L}{M}\right)}, \tag{2.26}$$

$$\mathrm{Im}\left(U_{m,n,\delta}\right) = -\frac{j\sin L\pi}{M}. \tag{2.27}$$

[1]Compared to the formulation of $U_{m,n,\delta}$ in [42], we have made slight modifications specifically for cases where M is even. In [43], the corrected formulation has been adopted.

In matrix representation, the SFI formulation (2.24) can also be defined as

$$\begin{aligned}\boldsymbol{x}(\Delta) &= \begin{bmatrix} x_{0+\delta} & \cdots & x_{M-1+\delta} \end{bmatrix}^{\mathsf{T}} \\ &= \begin{bmatrix} U_{0,0,\delta} & \cdots & U_{0,M-1,\delta} \\ \vdots & \ddots & \vdots \\ U_{M-1,0,\delta} & \cdots & U_{M-1,M-1,\delta} \end{bmatrix} \begin{bmatrix} x_0 \\ \vdots \\ x_{M-1} \end{bmatrix} \\ &= \boldsymbol{U}_M(\Delta)\boldsymbol{x}, \end{aligned} \tag{2.28}$$

where $\boldsymbol{U}_M(\Delta)$ is the rotation transform matrix, and $\boldsymbol{x} = \begin{bmatrix} x_0 & \cdots & x_{M-1} \end{bmatrix}^{\mathsf{T}}$ is the multichannel signal of the CMA after rotating $-\Delta$ rads in the time–frequency domain and is equal to $\boldsymbol{x}(0)$. Significantly, while all formulations are delineated within the confines of the time–frequency domain, it is noteworthy that this interpolation method in [42, 43] is not domain-restricted. In other words, even if the signal pertains to the time-domain, this interpolation technique remains applicable. It should also be emphasized that $\boldsymbol{U}_M(\Delta)$ takes the form of a cyclic matrix and remains independent of the frequency of the observed signal.

2.3.2 Analysis of rotation transform matrix

Given that our proposed methods in the following chapters rely on the application of the inverse matrix of the rotation transform matrix $\boldsymbol{U}_M(\Delta)$, it is essential to examine the properties of its inverse matrix. For clarity in the subsequent discussion, we define the es-CMA's position before rotation as the "reference position".

If an es-CMA undergoes an initial rotation by Δ rads, followed by a subsequent rotation by $-\Delta$ rads, it will return to its reference position. This relationship implies that the inverse of the rotation transform matrix, $\boldsymbol{U}_M(\Delta)^{-1}$, is equivalent to $\boldsymbol{U}_M(-\Delta)$.

Moreover, as established in [43], $\boldsymbol{U}_M(\Delta)$ is a unitary matrix.

In (2.25), when the number of microphones, M, is even, specific challenges arise due to the Nyqf component. In such cases, the numerator of the first term in (2.25) in the case of an even M, $e^{jL\pi} = (-1)^{n-m} e^{-j\delta\pi}$, represents the Nyqf component. This component demands careful handling to ensure accurate signal reconstruction and interpolation. In a previous paper [43], three distinct approaches were proposed to address this issue:

- Complex Nyqf (CoN): The Nyqf component is used in its complete complex form, retaining all phase and magnitude information.
- Real Nyqf (ReN): Only the real part of the Nyqf component is considered, simplifying the representation.
- Zero Phase Nyqf (ZPN): A zero value is substituted for δ exclusively in the Nyqf component, effectively neutralizing its phase shift.

In the subsequent part of this subsection, we focus on the scenario where M is even. This restriction allows us to explore the impact of the Nyqf component on the rotation matrix. By analyzing (2.26), (2.27), and leveraging the commutative property of matrix multiplication, we observe that in the absence of the imaginary part (2.27), the product $\boldsymbol{U}_M(\Delta) \cdot \boldsymbol{U}_M(-\Delta)$ does not equal the identity matrix under the ReN and ZPN. As a result, the es-CMA fails to return to its original position after an initial rotation of Δ rads followed by a subsequent rotation of $-\Delta$ rads. Additionally, the unitary property of the matrix is no longer preserved in these approaches, suggesting that both ReN and ZPN encounter limitations or inconsistencies when handling the Nyqf component. These findings indicate that while ReN and ZPN may provide practical simplifications when handling the Nyqf component, they introduce certain inaccuracies or deviations

that need to be carefully considered.

Essentially, the inverse matrix of $\boldsymbol{U}_M(\Delta)$ always exists, except in the exceptional case encountered with ReN. Returning to the initial stages of the derivation for SFI, according to [43], (2.23) can be simplified as

$$x_{m+\delta} = \mathscr{F}_{\mathrm{D}}^{-1}\left(\mathscr{F}_{\mathrm{D}}\left[x_m\right] e^{j\Delta k}\right). \tag{2.29}$$

By substituting the Fourier transform $\mathscr{F}_{\mathrm{D}}$ with the DFT matrix $\boldsymbol{F}$, we can translate (2.29) into the matrix representation

$$\boldsymbol{x}(\Delta) = \boldsymbol{F}^{-1}\boldsymbol{E}\left(\Delta\right)\boldsymbol{F}\boldsymbol{x}(0), \tag{2.30}$$

where

$$\boldsymbol{F} = \frac{1}{\sqrt{M}}\begin{bmatrix} e^{-j\frac{2\pi}{M}\cdot 0\cdot 0} & \dots & e^{-j\frac{2\pi}{M}\cdot 0\cdot(M-1)} \\ \vdots & \ddots & \vdots \\ e^{-j\frac{2\pi}{M}\cdot(M-1)\cdot 0} & \dots & e^{-j\frac{2\pi}{M}\cdot(M-1)\cdot(M-1)} \end{bmatrix}, \tag{2.31}$$

and

$$\boldsymbol{E}\left(\Delta\right) = \mathrm{diag}\left(e^{j\Delta\lceil 1-M/2\rceil}, ..., e^{j\Delta\lceil M/2\rceil}\right) \tag{2.32}$$

is a diagonal matrix of phase rotation for the δ-sample shift, where $\lceil\bullet\rceil$ indicates the ceiling function.

The specific case where Δ is equal to π/M results in the last element in $\boldsymbol{E}\left(\Delta\right)$ being $e^{j\pi/2}$. In ReN, where the imaginary part is neglected, the phase rotation matrix

$\boldsymbol{E}_{\mathrm{ReN}}(\pi/M)$ can be calculated as

$$\boldsymbol{E}_{\mathrm{ReN}}\left(\frac{\pi}{M}\right) = \mathrm{diag}\left(\cos\left(\frac{(1-M)\pi}{2M}\right), ..., 0\right), \tag{2.33}$$

which is a singular matrix. Therefore, in the ReN case, the rotation transform matrix $\boldsymbol{U}_M(\pi/M) = \boldsymbol{F}^{-1}\boldsymbol{E}_{\mathrm{ReN}}(\pi/M)\boldsymbol{F}$ does not have a corresponding inverse matrix. Specifically, when Δ is equal to π/M, the inverse matrix of $\boldsymbol{U}_M(\Delta)$ does not exist. This unusual situation requires further examination, and will be discussed in the following chapter.

2.3.3 MPDR beamforming with sound field interpolation

To apply SFI to the MPDR beamforming, we first assume the following conditions:

- The steering vector of the target sound source observed by an M-channel es-CMA at the reference position, denoted as $\boldsymbol{v}_f$, is given, where f indicates the frequency bin index. It is commonly assumed that the steering vector is either obtained or pre-estimated in advance for beamforming applications [3, 60, 61].
- The rotation angle for each time frame, θ_t, is provided, where t indicates the time frame index. The rotation angle can be readily obtained through various methods, such as using an acceleration sensor or through other estimation methods [45, 62, 63].

Under these conditions, we will introduce two approaches for applying SFI to beamforming. One approach involves the use of a pre-estimated spatial filter, while the other focuses on online spatial filtering. In the following discussions, we refer to the observation made by the es-CMA at the reference position, without any rotation, as

the "reference observation". We assume that at the start time, the es-CMA is located at the reference position.

Batch processing with a pre-estimated spatial filter

In this process, we use a fixed MPDR beamforming spatial filter $\boldsymbol{W}_f$, which is pre-estimated using (2.19) in advance. As mentioned earlier, the steering vector $\boldsymbol{v}_f$ is also obtained, so the only remaining requirement for calculating the spatial filter is the estimation of the covariance matrix. To achieve this, we need the reference observation over a sufficiently long time period to estimate the covariance matrix $\boldsymbol{S}_f$, assuming that neither the sound sources nor the CMA move during these time frames.

After the es-CMA rotates by θ_t rads, the reference observation is reconstructed before beamforming using (2.28), as follows

$$\hat{\boldsymbol{x}}_{\mathrm{ref},tf} = \boldsymbol{U}_M(-\theta_t)\boldsymbol{x}_{tf}, \tag{2.34}$$

which implies that by performing interpolation along the inverse rotation, we effectively restore the rotated es-CMA to its reference position. With the es-CMA virtually returned to the reference position, the spatial information can be considered unchanged, eliminating the need to re-estimate the covariance matrix $\boldsymbol{S}_f$ and the steering vector $\boldsymbol{v}_f$. Consequently, the target source can be directly enhanced using the pre-estimated filter $\boldsymbol{W}_f$ and the estimated reference observation, just as in conventional beamforming

$$y_{tf} = \boldsymbol{W}_f^{\mathsf{H}}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}. \tag{2.35}$$

Online Processing of Spatial Filter

While the batch processing approach outlined in Section 2.3.3 is effective when the ATS remains stationary aside from es-CMA rotation, this condition is rarely met in real-world scenarios. To address situations where the ATS undergoes minor variations in addition to es-CMA rotation, this subsection introduces an updated online processing algorithm incorporating SFI for beamforming. In this approach, a well-known smoothing (forgetting) factor α [64, 65] is employed to update the covariance matrix in real time. In addition, a matrix inversion lemma, the Sherman–Morrison formula [66–68], which can reduce the complexity of calculating the covariance matrix inversion that appears in the MPDR beamforming formulation, is utilized, making it more efficient for online processing applications.

Firstly, we estimate the reference observation by (2.34), as in the batch processing. By using the interpolated observation, we can estimate the covariance matrix at the t-th frame, $\hat{\boldsymbol{S}}_{tf}$, based on the covariance matrix from the previous frame, $\hat{\boldsymbol{S}}_{(t-1)f}$, and the smoothing factor α, as follows:

$$\hat{\boldsymbol{S}}_{tf} = \alpha \hat{\boldsymbol{S}}_{(t-1)f} + (1-\alpha)\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}. \tag{2.36}$$

This formulation, incorporating the smoothing factor α, is commonly used in online signal processing studies to adapt the covariance matrix over time [65,69]. Additionally, the Sherman–Morrison formula is utilized to efficiently compute the inverse of $\hat{\boldsymbol{S}}_{tf}$ with reduced complexity, facilitating its application in MPDR beamforming (2.19) for each

time frame. The inverse of $\hat{\boldsymbol{S}}_{tf}$ is calculated as follows

$$\begin{aligned}\hat{\boldsymbol{S}}_{tf}^{-1} &= \frac{1}{\alpha}\hat{\boldsymbol{S}}_{(t-1)f}^{-1} - \frac{\left(\alpha\hat{\boldsymbol{S}}_{(t-1)f}\right)^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\left(\alpha\hat{\boldsymbol{S}}_{(t-1)f}\right)^{-1}}{\frac{1}{1-\alpha} + \hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\left(\alpha\hat{\boldsymbol{S}}_{(t-1)f}\right)^{-1}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}} \\ &= \frac{1}{\alpha}\hat{\boldsymbol{S}}_{(t-1)f}^{-1} - \frac{\hat{\boldsymbol{S}}_{(t-1)f}^{-1}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\,\hat{\boldsymbol{S}}_{(t-1)f}^{-1}}{\frac{\alpha^2}{1-\alpha} + \alpha\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\,\hat{\boldsymbol{S}}_{(t-1)f}^{-1}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}},\end{aligned} \tag{2.37}$$

which is then employed to update the spatial filter, enabling online enhancement of the target source.

Algorithm 1: Online beamforming update algorithm with SFI

input : $\boldsymbol{x}_{tf} \in \mathbb{C}^{M\times 1}$, $\boldsymbol{v}_f \in \mathbb{C}^{M\times 1}$, $\theta_t \in \mathbb{R}$

output: y_{tf}

for $f \leftarrow 0$ **to** $F-1$ **do**
  initialize $\hat{\boldsymbol{S}}_f^{-1}$
for $f \leftarrow 0$ **to** $F-1$ **do**
  $\hat{\boldsymbol{x}}_{\mathrm{ref},tf} \leftarrow \boldsymbol{U}_M(-\theta_t)\boldsymbol{x}_{tf}$
  $\hat{\boldsymbol{S}}_{tf}^{-1} \leftarrow \frac{1}{\alpha}\hat{\boldsymbol{S}}_{(t-1)f}^{-1} - \frac{\hat{\boldsymbol{S}}_{(t-1)f}^{-1}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\,\hat{\boldsymbol{S}}_{(t-1)f}^{-1}}{\frac{\alpha^2}{1-\alpha} + \alpha\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}^{\mathsf{H}}\,\hat{\boldsymbol{S}}_{(t-1)f}^{-1}\,\hat{\boldsymbol{x}}_{\mathrm{ref},tf}}$
  $\boldsymbol{W}_{tf}^{\mathsf{H}} \leftarrow \frac{\boldsymbol{v}_f^{\mathsf{H}}\boldsymbol{S}_{tf}^{-1}}{\boldsymbol{v}_f^{\mathsf{H}}\boldsymbol{S}_{tf}^{-1}\boldsymbol{v}_f}$
  $y_{tf} \leftarrow \boldsymbol{W}_{tf}^{\mathsf{H}}\hat{\boldsymbol{x}}_{\mathrm{ref},tf}$

Algorithm 1 provides a detailed pseudo-code representation of the frame-wise online processing methodology, incorporating all essential formulations. This algorithm leverages the interpolated reference observation, smoothing factor, and efficient covariance matrix inversion using the Sherman–Morrison formula to adaptively enhance the target source.

It is important to emphasize the initialization of $\hat{\boldsymbol{S}}_{tf}^{-1}$, which plays a crucial role in ensuring the stability and convergence of the online algorithm. Various initialization strategies are available, including:

- **Random Matrix**: Assigning random values to $\hat{\boldsymbol{S}}_{tf}^{-1}$, which introduces variability but may lead to unpredictable behavior in early frames.
- **Identity Matrix**: Setting $\hat{\boldsymbol{S}}_{tf}^{-1}$ to the identity matrix, offering a neutral starting point and simplifying initial calculations.
- **Averaged Inversion**: Computing $\hat{\boldsymbol{S}}_{tf}^{-1}$ as the inverse of $\boldsymbol{x}_{tf}\boldsymbol{x}_{tf}^{\mathsf{H}}$ averaged over a predefined number of initial time frames. This approach incorporates prior observations, potentially improving initialization accuracy.

The choice of initialization method should be informed by the specific requirements of the application and the characteristics of the observed data. For instance, in environments with significant variability or noise, averaging over several frames may provide a more robust starting point. Conversely, for applications prioritizing computational simplicity, using the identity matrix might be preferable. Careful consideration and testing of the initialization strategy are essential to ensure optimal performance of the online processing framework.

2.4 Summary

This chapter has provided the foundational knowledge essential for understanding the concepts and methodologies presented in this thesis. Section 2.2 introduced the principles and implementation of MPDR beamforming, laying the groundwork for its application in sound field processing. Section 2.3 reviewed the SFI method, which is

integral to achieving rotation-robust beamforming with an es-CMA. Furthermore, we conducted an in-depth analysis of key SFI properties, including its singularity behavior, and demonstrated its practical application in both batch and online beamforming scenarios. The discussions and analyses in this chapter are intrinsically linked to the methods proposed later in this thesis. These foundational insights will be frequently referenced in subsequent chapters, serving as a basis for the proposed contributions. It is our intent that this chapter equips readers with a preliminary understanding of the research framework, thereby enhancing their comprehension of the advanced concepts and techniques presented in the following chapters.

3 Unequally Spaced Sound Field Interpolation

This chapter introduces an advanced method, named unequally spaced sound field interpolation (unes-SFI), specifically designed to enable rotation-robust beamforming with unequally spaced circular microphone arrays (unes-CMAs). The unes-SFI technique builds upon a modified sound field interpolation (SFI) framework to address the challenges posed by positional deviations of microphones in an unes-CMA. It estimates virtual after-rotation signals at equally spaced positions, compensating for microphone placement errors. The method leverages the previous SFI approach as an intermediate step to derive before-rotation equally spaced signals at the reference position. Subsequently, unes-SFI reconstructs the target before-rotation signal of the unes-CMA at its reference position, effectively enabling robust beamforming despite rotational transformations.

This chapter also provides a detailed analysis of the properties of unes-SFI. Simulated experiments, including online beamforming scenarios, demonstrate that unes-SFI significantly mitigates the adverse effects caused by unequal microphone spacing. It achieves substantial improvements in reconstructing the signal at the reference position under various conditions. Furthermore, unes-SFI consistently outperforms the previous SFI approach, delivering marked enhancements in beamforming accuracy and performance.

3.1 Overview

The proposed unes-SFI method is designed to tackle two critical challenges in rotation-robust beamforming: the time-variant acoustic transfer system (ATS) and the use of unes-CMAs. This method enables the estimation of a virtual signal of an unes-CMA at its reference position before rotation, using the observed signal obtained after rotation. In contrast to the previous SFI approach, which required an equally spaced circular microphone array es-CMA, unes-SFI accommodates non-uniformly spaced microphones. This capability makes unes-SFI particularly suited for practical scenarios, where unes-CMAs are more prevalent due to manufacturing tolerances or environmental constraints. It is worth emphasizing that unes-SFI shares a fundamental objective with previous SFI method [42, 43]: avoiding the need for recalculating or updating spatial filters after the array undergoes rotation. By addressing the limitations of traditional SFI, unes-SFI extends the scope of rotation-robust beamforming to include more diverse and realistic array configurations.

In unes-SFI, the error angle vector $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_M \end{bmatrix}^{\mathsf{T}}$ is assumed to be known beforehand. It is important to clarify that, in the before-rotation state, for a fixed-distribution M-channel unes-CMA, the microphone closest to the 0° position is designated as the 1st microphone. The subsequent microphones are then numbered sequentially in a counterclockwise direction as the 2nd microphone, ..., the mth microphone, ..., the $(M-1)$th microphone, and finally the Mth microphone. Here, $\epsilon_m \in (-2\pi/M, 2\pi/M)$ represents the angular deviation between the actual position of the mth microphone on the unes-CMA and its corresponding ideal position in a uniformly spaced distribution. As shown in Figure 3.1, the proposed method encompasses three distinct steps:

- **Compensate for the error angles**: Using the known error angle vector $\boldsymbol{\epsilon}$ and the observed signal from the unes-CMA after rotation, we estimate a pseudo-

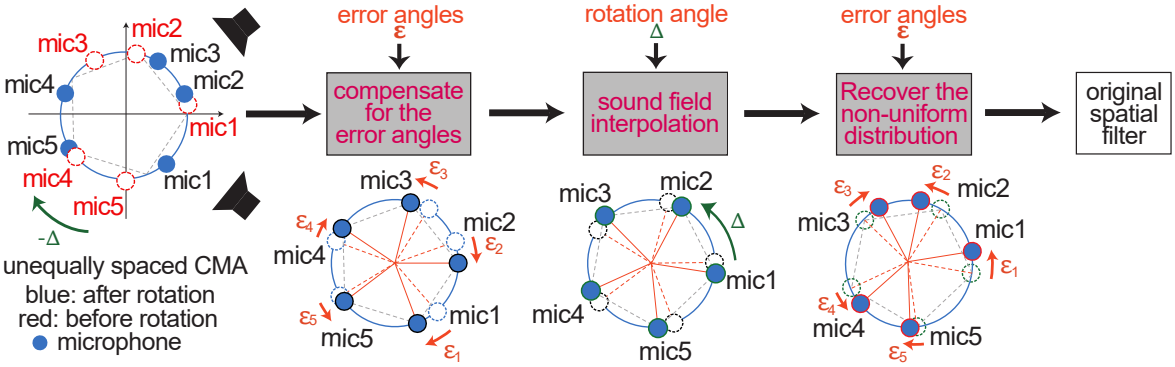

Figure 3.1: *Conceptual diagram of unes-SFI.*

signal that would have been captured by a virtual es-CMA. This step simplifies the inherently complex problem of handling an unes-CMA by reducing it to a more manageable equivalent on an es-CMA.

- **Sound field interpolation**: With the rotation angle Δ (obtained through various means such as accelerometers or estimation techniques [45, 62, 63]), we reconstruct the before-rotation signal observed by the virtual es-CMA. This step reverses the rotational effect on the sound field to provide a consistent reference observation.

- **Recover the non-uniform distribution**: Finally, we revert to the original non-uniform microphone arrangement, recovering the signal that the unes-CMA would have captured before rotation. This step ensures compatibility with the actual configuration of the unes-CMA, enabling direct application of the reconstructed signals to other array signal processing methods.

3.2 Formulation

In this section, we present the mathematical formulation of the proposed method.

Incorporating the error vector $\boldsymbol{\epsilon}$, the sound field function observed by the unes-CMA after a rotation of $-\Delta$ rads can be expressed as follows

$$\boldsymbol{x}(\boldsymbol{\epsilon}) = \left[x\left(\tfrac{2\pi\cdot 0}{M} + \epsilon_1\right) \quad \cdots \quad x\left(\tfrac{2\pi(M-1)}{M} + \epsilon_M\right) \right]^{\mathsf{T}}. \tag{3.1}$$

In (3.1), ϵ_m can be interpreted as the rotation angle from the ideal position of the mth microphone in the es-CMA to its actual position in the unes-CMA. This induces a rotated sound signal, denoted as $\boldsymbol{x}(\epsilon_m)$, which can be computed using the rotation transform matrix $\boldsymbol{U}_M(\epsilon_m)$.

From (2.28), it is evident that when extracting the mth channel signal from the rotated sound field $\boldsymbol{x}(\Delta)$, only the mth row of the rotation transform matrix $\boldsymbol{U}_M(\Delta)$ is necessary. Similarly, for the mth channel signal of the unes-CMA, $x\left(2\pi(m-1)/M + \epsilon_m\right)$, which aligns with the mth channel signal of $\boldsymbol{x}(\epsilon_m)$, we can establish its relationship with the pseudo observation $\hat{\boldsymbol{x}}(0)$ from a virtual es-CMA, which is

$$\hat{\boldsymbol{x}}(0) = \left[\hat{x}(0) \quad \cdots \quad \hat{x}\left(\tfrac{2\pi(M-1)}{M}\right) \right]^{\mathsf{T}}, \tag{3.2}$$

where $\hat{\bullet}$ represents a pseudo observation from a virtual CMA. The relationship is then given by:

$$x\left(\frac{2\pi(m-1)}{M} + \epsilon_m\right) = \boldsymbol{u}_m(\epsilon_m)\hat{\boldsymbol{x}}(0), \tag{3.3}$$

where $\boldsymbol{u}_m(\epsilon_l) \in \mathbb{C}^{1\times M}$ denotes the mth row of the rotation transform matrix $\boldsymbol{U}_M(\epsilon_l)$.

Referring to (3.3), the complete observation recorded using an unes-CMA after rotation can be expressed as

$$\boldsymbol{x}(\boldsymbol{\epsilon}) = \boldsymbol{U}_M(\boldsymbol{\epsilon})\,\hat{\boldsymbol{x}}(0), \tag{3.4}$$

where $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ is the compensation matrix defined as

$$\boldsymbol{U}_M(\boldsymbol{\epsilon}) \stackrel{\text{def}}{=} \left[\boldsymbol{u}_1^{\mathsf{T}}(\epsilon_1) \quad \cdots \quad \boldsymbol{u}_M^{\mathsf{T}}(\epsilon_M)\right]^{\mathsf{T}}, \tag{3.5}$$

and contains rows from different rotation transform matrices, each corresponding to the angular deviation of a specific microphone position. From (3.4), $\hat{\boldsymbol{x}}(0)$ can be calculated as

$$\hat{\boldsymbol{x}}(0) = \boldsymbol{U}_M(\boldsymbol{\epsilon})^{-1}\boldsymbol{x}(\boldsymbol{\epsilon}). \tag{3.6}$$

Consequently, in the first step, the microphone positional errors ($\boldsymbol{\epsilon}$) are compensated, yielding a virtual sound signal with a uniform distribution derived from the observation of an unes-CMA. In essence, the unes-CMA is virtually transformed into the es-CMA using the inverse matrix of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$.

In the second step, SFI is applied to $\hat{\boldsymbol{x}}(0)$ to calculate the Δ-rad-rotated result of the virtual es-CMA. This corresponds to the sound signal captured by the virtual es-CMA before rotation. The SFI step utilizes the rotation transform matrix for the rotation angle Δ to recover the before-rotation signal:

$$\hat{\boldsymbol{x}}(\Delta) = \left[\hat{x}(\Delta) \cdots \hat{x}\left(\tfrac{2\pi(M-1)}{M} + \Delta\right)\right]^{\mathsf{T}} = \boldsymbol{U}_M(\Delta)\hat{\boldsymbol{x}}(0). \tag{3.7}$$

The final step involves converting the virtual equally spaced signal before rotation, $\hat{\boldsymbol{x}}(\Delta)$, back to the real unequally spaced signal observed by the unes-CMA, represented as

$$\boldsymbol{x}(\Delta + \boldsymbol{\epsilon}) = \left[x(\Delta + \epsilon_1) \cdots x\left(\tfrac{2\pi(M-1)}{M} + \Delta + \epsilon_M\right)\right]^{\mathsf{T}}. \tag{3.8}$$

Using a similar approach to the first step (3.4), the before-rotation signal in the

unes-CMA can be calculated as:

$$\boldsymbol{x}(\Delta + \boldsymbol{\epsilon}) = \boldsymbol{U}_M(\boldsymbol{\epsilon})\hat{\boldsymbol{x}}(\Delta). \tag{3.9}$$

By combining these steps (3.6), (3.7), and (3.9), the relationship between the before-rotation signal and the after-rotation signal on the unes-CMA can be expressed as:

$$\boldsymbol{x}(\Delta + \boldsymbol{\epsilon}) = \boldsymbol{U}_M(\boldsymbol{\epsilon})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\boldsymbol{\epsilon})^{-1}\boldsymbol{x}(\boldsymbol{\epsilon}). \tag{3.10}$$

Upon completing these three sequential steps, the unes-CMA is effectively aligned with the original reference position before rotation. The pre-existing spatial filter designed for the reference position can then be directly applied to the reconstructed signal, $\boldsymbol{x}(\Delta + \boldsymbol{\epsilon})$, without requiring re-estimation of the filter. This capability significantly enhances the computational efficiency of online processing, making it practical for real-world scenarios.

3.3 Analysis of the compensation matrix

The compensation matrix $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ is fundamental to the unes-SFI method, as it addresses the angular deviations in microphone positions. By mapping the observed signal from an unes-CMA to a virtual es-CMA, $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ enables the subsequent operations, such as rotation and signal recovery, to be performed as if the observations were made with an equally spaced array. Therefore, it is necessary to analyze the properties of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, as these directly impact the effectiveness of the unes-SFI method.

When M, the number of microphones, is even, the Nyquist frequency (Nyqf) component plays a significant role in the behavior of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$. Handling this component appropriately is critical for maintaining the desired properties of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, such as in-

vertibility and unitarity, which directly affect the accuracy of the compensation. Three approaches for dealing with the Nyqf component, as previously introduced in Section 2.3.2, are: CoN, ReN, and ZPN. The subsequent analysis of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ will provide deeper insights into its behavior under different Nyqf approaches, guiding optimal design and application of the proposed method when M is even.

3.3.1 Periodicity

Evidently, in the previous SFI for an es-CMA [42, 43], the interpolation accuracy exhibits periodicity with respect to the rotation angle. This periodicity arises because a rotation of the es-CMA by an angle equivalent to the angular spacing between adjacent microphones ($2\pi/M$) results in a cyclic permutation of the microphone indices. Consequently, the rotation transform matrix $\boldsymbol{U}_M(\Delta)$ becomes an M-cyclic permutation matrix in such cases.

In our proposed unes-SFI, this periodicity is expected to persist, even when the angles between adjacent microphones differ from each other. For instance, consider the scenario where $\boldsymbol{x}(\boldsymbol{\epsilon})$ is observed after the unes-CMA undergoes a rotation of $-\Delta$ rads, which corresponds to the angular deviation between two specific channels, the ath and bth channels ($a \neq b$),

$$-\Delta = \left(\frac{2\pi(b-1)}{M} + \epsilon_b\right) - \left(\frac{2\pi(a-1)}{M} + \epsilon_a\right), \tag{3.11}$$

then the following equation holds:

$$x(\Delta + \frac{2\pi(b-1)}{M} + \epsilon_b) = x(\frac{2\pi(a-1)}{M} + \epsilon_a). \tag{3.12}$$

The left-hand side of (3.12) represents the signal of the bth channel before rotating by

$-\Delta$ rads, whereas the right-hand side corresponds to the signal of the ath channel after rotation. This indicates that the bth channel's position before rotation aligns with the ath channel's position after rotation when the rotation angle $-\Delta$ equals the angular deviation between these two channels. From this result, we can deduce that, even in an unes-CMA, the interpolation accuracy of a specific channel exhibits periodic behavior with respect to the rotation angle. The accuracy peaks whenever the rotation angle matches the angular deviation between two channels.

(3.12) can be analytically proven. According to the unes-SFI formulation (3.10), we can calculate the estimated signal of the bth channel before rotation as follows:

$$x(\Delta + \frac{2\pi(b-1)}{M} + \epsilon_b) = \boldsymbol{u}_b(\epsilon_b)\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\boldsymbol{\epsilon})^{-1}\boldsymbol{x}(\boldsymbol{\epsilon}), \tag{3.13}$$

where $\boldsymbol{u}_b(\epsilon_b)\boldsymbol{U}_M(\Delta)$ can be expressed as

$$\boldsymbol{u}_b(\epsilon_b)\boldsymbol{U}_M(\Delta) = \boldsymbol{u}_b(\epsilon_b + \Delta) = \boldsymbol{u}_b(\epsilon_a + \frac{2\pi(a-b)}{M}). \tag{3.14}$$

By substituting $m = b$ and $\delta = M(\epsilon_a + 2\pi(a-b)/M)/2\pi$ into L in (2.25), we obtain

$$L = n - b - \frac{M\left(\epsilon_a + \frac{2\pi(a-b)}{M}\right)}{2\pi} = n - a - \frac{M\epsilon_a}{2\pi}. \tag{3.15}$$

According to (2.25) and (2.28), it is easy to know that

$$\boldsymbol{u}_b(\epsilon_b)\boldsymbol{U}_M(\Delta) = \boldsymbol{u}_a(\epsilon_a). \tag{3.16}$$

Using (3.16), we can obtain (3.12) by simplifying the right-hand side of (3.13):

$$\begin{aligned} x(\Delta + \frac{2\pi(b-1)}{M} + \epsilon_b) &= \boldsymbol{u}_a(\epsilon_a)\boldsymbol{U}_M(\boldsymbol{\epsilon})^{-1}\boldsymbol{x}(\boldsymbol{\epsilon}) \\ &= x(\frac{2\pi(a-1)}{M} + \epsilon_a). \end{aligned} \tag{3.17}$$

However, (3.12) holds only in CoN and ReN but does not apply in ZPN. In ZPN, the treatment of the Nyqf component introduces an additional constraint that affects the equivalence relation described in (3.12). Specifically, because the parameter δ of L is ignored in the Nyqf component, the relationship in (3.15) no longer holds, but instead becomes

$$L = n - b \neq n - a = n - b + (b - a). \tag{3.18}$$

Hence, the Nyqf components for the bth and ath channels are given by

$$\begin{aligned} e_b^{jL\pi} &= (-1)^{n-b}, \\ e_a^{jL\pi} &= (-1)^{n-b+(b-a)} = (-1)^{n-b} \cdot (-1)^{b-a}. \end{aligned} \tag{3.19}$$

From the expressions, we observe that for $e_b^{jL\pi} = e_a^{jL\pi}$ to hold, the additional factor $(-1)^{b-a}$ must equal 1, which happens only if $b - a$ is an even number.

For L in the remaining terms of (2.25), where δ is not replaced by zero, the equivalence relation in (3.15) continues to hold. Consequently, in the ZPN approach, the equivalence relation described in (3.12) holds only when $b - a$ is even. For cases where $b - a$ is odd, the mismatch in the Nyqf component means that the equivalence breaks down. Thus, in ZPN, although the interpolation accuracy of a specific microphone exhibits periodicity with respect to the rotation angle, similar to other approaches,

the maximum interpolation accuracy can only be achieved when the rotated microphone aligns with the position of another microphone and simultaneously skips an odd number of intermediate microphones.

3.3.2 Singularity

As seen in (3.10), the inverse matrix of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ holds a crucial significance for the performance of unes-SFI. Therefore, the effectiveness of unes-SFI is heavily influenced by the singularity of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$. Under typical conditions, the inverse of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ exists, except in cases where two microphones are positioned at the same location on the unes-CMA, which is physically impossible in practical applications. However, as discussed in the preceding section, an exceptional situation arises in ReN, where $\boldsymbol{U}_M(\Delta)$ becomes singular. To address this, we will further examine whether a similar phenomenon occurs for unes-SFI in ReN and aim to derive a more universally applicable conclusion regarding $\boldsymbol{U}_M(\boldsymbol{\epsilon})$.

Firstly, when M is even, according to (2.26), the cos function in the latter term of $\mathrm{Re}\left(U_{m,n,\delta}\right)$ can be rewritten as

$$\begin{aligned}&\cos\left(\frac{M+2}{2M}L\pi\right)=\cos\left(\frac{1}{2}L\pi+\frac{1}{M}L\pi\right)\\&=\cos\left(\frac{1}{2}L\pi\right)\cos\left(\frac{1}{M}L\pi\right)-\sin\left(\frac{1}{2}L\pi\right)\sin\left(\frac{1}{M}L\pi\right),\end{aligned}\tag{3.20}$$

whereas the other part in the latter term is reformulated as

$$\frac{\mathrm{sinc}\left(\frac{L}{2}\right)}{\mathrm{sinc}\left(\frac{L}{M}\right)}=\frac{2}{M}\cdot\frac{\sin\left(\frac{L\pi}{2}\right)}{\sin\left(\frac{L\pi}{M}\right)}.\tag{3.21}$$

To ensure the validity of (3.21) for $\forall n,m\in[1,M]$, it is imperative to assume that

$\delta \neq 0$, thereby preventing a zero denominator when $n = m$.

Thus, (2.26) is simplified to

$$\begin{aligned}
&\mathrm{Re}\left(U_{m,n,\delta}\right) = \frac{(1-\cos L\pi)}{M} + \frac{2}{M}\cdot\frac{\sin\left(\frac{L\pi}{2}\right)}{\sin\left(\frac{L\pi}{M}\right)}\cdot \\
&\left[\cos\left(\frac{L\pi}{2}\right)\cos\left(\frac{L\pi}{M}\right) - \sin\left(\frac{L\pi}{2}\right)\sin\left(\frac{L\pi}{M}\right)\right] \\
&= \frac{1}{M}\left[1-\cos\left(L\pi\right) + \frac{\sin\left(L\pi\right)\cos\left(\frac{L\pi}{M}\right)}{\sin\left(\frac{L\pi}{M}\right)} - 2\sin^2\left(\frac{L\pi}{2}\right)\right] \\
&= \frac{1}{M}\cdot\frac{\sin\left(L\pi\right)\cos\left(\frac{L\pi}{M}\right)}{\sin\left(\frac{L\pi}{M}\right)} = -\frac{\sin\left(\delta\pi\right)}{M}\cdot V_{m,n,\delta},
\end{aligned} \tag{3.22}$$

where $V_{m,n,\delta}$ is defined as

$$V_{m,n,\delta} = \frac{(-1)^{n-m}\cos\left(\frac{L\pi}{M}\right)}{\sin\left(\frac{L\pi}{M}\right)} = (-1)^{n-m}\cot\left(\frac{L\pi}{M}\right). \tag{3.23}$$

In ReN, as the imaginary part (2.27) is neglected, $U_{m,n,\delta}$ is simplified to $\mathrm{Re}\left(U_{m,n,\delta}\right)$. Consequently, $\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right)$ in (3.5) can be redefined as

$$\begin{aligned}
\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right) = &\begin{bmatrix} -\frac{\sin(\delta_1\pi)}{M} & & \\ & \ddots & \\ & & -\frac{\sin(\delta_M\pi)}{M} \end{bmatrix} \\
&\cdot\begin{bmatrix} V_{0,0,\delta_1} & \cdots & V_{0,M-1,\delta_1} \\ \vdots & \ddots & \vdots \\ V_{M-1,0,\delta_M} & \cdots & V_{M-1,M-1,\delta_M} \end{bmatrix} \\
=&\boldsymbol{D}_M\left(\boldsymbol{\epsilon}\right)\cdot\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right),
\end{aligned} \tag{3.24}$$

where $\delta_i = M\epsilon_i/2\pi \in (-1,1)$, $i \in [1,M]$. Then, the determinant of $\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right)$ can be

calculated as

$$\det\left(\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right)\right) = \det\left(\boldsymbol{D}_M\left(\boldsymbol{\epsilon}\right)\right) \cdot \det\left(\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)\right). \tag{3.25}$$

As previously indicated, it is assumed that δ_i is not equal to zero for $\forall i \in [1, M]$, consequently rendering $\det\left(\boldsymbol{D}_M\left(\boldsymbol{\epsilon}\right)\right)$ as a non-zero constant. As a result, the matrix $\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right)$ is singular only when $\det\left(\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)\right) = 0$.

Here, for simplicity, we give a concrete example with $M = 2$ and $\boldsymbol{\epsilon} = \begin{bmatrix}\epsilon_1 & \epsilon_2\end{bmatrix}^{\mathsf{T}}$. Then, $\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)$ is calculated as

$$\begin{aligned}\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right) &= \begin{bmatrix} V_{0,0,\delta_1} & V_{0,1,\delta_1} \\ V_{1,0,\delta_2} & V_{1,1,\delta_2} \end{bmatrix} \\ &= \begin{bmatrix} -\cot\left(\frac{\delta_1}{2}\pi\right) & \cot\left(\frac{(\delta_1-1)}{2}\pi\right) \\ \cot\left(\frac{(\delta_2+1)}{2}\pi\right) & -\cot\left(\frac{\delta_2}{2}\pi\right) \end{bmatrix}.\end{aligned} \tag{3.26}$$

The determinant of $\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)$ can be obtained as

$$\det\left(\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)\right) = -\frac{4\cos\left(\frac{(\delta_1+\delta_2)}{2}\pi\right) \cdot \cos\left(\frac{(\delta_1-\delta_2)}{2}\pi\right)}{\cos\left(\frac{(\delta_1+\delta_2)}{2}\pi\right)^2 - \cos\left(\frac{(\delta_1-\delta_2)}{2}\pi\right)^2}. \tag{3.27}$$

Obviously, when $\delta_1 + \delta_2 = 1$ or $\delta_1 - \delta_2 = 1$, the determinant of $\boldsymbol{V}_M\left(\boldsymbol{\epsilon}\right)$ will be zero. However, $\delta_1 - \delta_2 = 1$, which corresponds to $\epsilon_1 - \epsilon_2 = 2\pi/M$, indicates that two microphones are overlapped and placed in the same position. These conditions were previously noted as physically implausible. Thus, $\boldsymbol{U}_M\left(\boldsymbol{\epsilon}\right)$ becomes singular when $\delta_1 + \delta_2 = 1$, which can also be expressed as $\epsilon_1 + \epsilon_2 = \pi$.

When there is a zero-value δ, *e.g.*, $\delta_i = 0$ for the ith channel, the ith rows of $\boldsymbol{D}_M\left(\boldsymbol{\epsilon}\right)$

and $\boldsymbol{V}_M(\boldsymbol{\epsilon})$, denoted by $\boldsymbol{d}_i(\epsilon_i)$ and $\boldsymbol{v}_i(\epsilon_i)$, respectively, will be adjusted as

$$\boldsymbol{d}_i(\epsilon_i) = \boldsymbol{v}_i(\epsilon_i) = \boldsymbol{I}_i, \tag{3.28}$$

where $\boldsymbol{I}_i$ is the ith row of an $M \times M$ identity matrix.

Taking $M = 4$ and $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_4 \end{bmatrix}^\mathsf{T}$ with $\epsilon_4 = 0°$ as an example, the determinant of $\boldsymbol{V}_M(\boldsymbol{\epsilon})$ can be calculated as

$$\begin{aligned} \det\left(\boldsymbol{V}_M(\boldsymbol{\epsilon})\right) = & \frac{4\cos\left(\frac{\delta_2-\delta_3+1}{4}\pi\right)}{\cos\left(\frac{\delta_1+1}{4}\pi\right)\sin\left(\frac{\delta_1}{2}\pi\right)} \cdot \frac{\sin\left(\frac{\delta_1-\delta_2-1}{4}\pi\right)}{\sin\left(\frac{\delta_2}{4}\pi\right)\cos\left(\frac{\delta_2}{2}\pi\right)} \\ & \cdot \frac{\cos\left(\frac{\delta_1-\delta_3}{4}\pi\right)}{\sin\left(\frac{\delta_3+1}{4}\pi\right)\sin\left(\frac{\delta_3}{2}\pi\right)} \cdot \cos\left(\frac{\delta_1+\delta_2+\delta_3}{4}\pi\right). \end{aligned} \tag{3.29}$$

Similarly, $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ is singular only when $\delta_1 + \delta_2 + \delta_3 = 2$, which can be reformulated as $\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 = \pi$.

Revisiting the earlier conclusion in Section 2.3.2, we can evidently see that in ReN, $\boldsymbol{U}_M(\Delta)$ becomes singular when Δ is equal to π/M, which in turn implies that the sum of Δ values from all M channels is also equal to π. As a result, a more general conclusion can be drawn, stating that $\boldsymbol{U}_M(\boldsymbol{\epsilon})^{-1}$ will not exist when

$$\sum_{i=1}^{M} \epsilon_i = \pi. \tag{3.30}$$

We will further experimentally investigate whether (3.30) holds when M is set to larger than 4 in Section 3.4.2.

It is important to note that, even in the absence of the previously discussed abnormal situation, $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ may still occasionally be a nearly singular matrix. In such instances, we apply singular value decomposition (SVD) [70] to $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, disregarding the eigenvalues and eigenvectors associated with exceptionally small condition num-

bers. Subsequently, the factorized coefficients of this truncated SVD [71] are employed to compute the inverse matrix of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$.

3.4 Simulated experimental evaluation

3.4.1 Setup

Dataset and preprocessing

To evaluate the performance and robustness of the proposed method to the rotation of an unes-CMA, simulation experiments were conducted using the SiSEC database [72]. Each utterance in the database was sampled at 16 kHz. Eight speech signals were selected, consisting of four female and four male voices, with sound sources positioned at various angles, as illustrated in Figure 3.2. To simulate a reverberant environment, the sound signals were convolved with room impulse responses (RIRs) simulated by an RIR generator [73] on the basis of the image method [74]. This process produced microphone signals with an approximate reverberation time of 100 ms. For analysis in the time–frequency domain, the STFT was applied using a 1/8-shifted Blackman window with a length of 64 ms.

Simulated experimental setup for unes-CMAs

The sound signals were recorded using an M-channel CMA with a radius of 0.05 m in a noise-free room. To create an unes-CMA, an error angle, denoted by $\epsilon_i(^\circ)$, $i \in \{1, ..., M\}$, was introduced to the position of each microphone. During the construction of the unes-CMA, it is possible that, after introducing error angles to the positions of two microphones, their positions may become swapped. For example, when creating

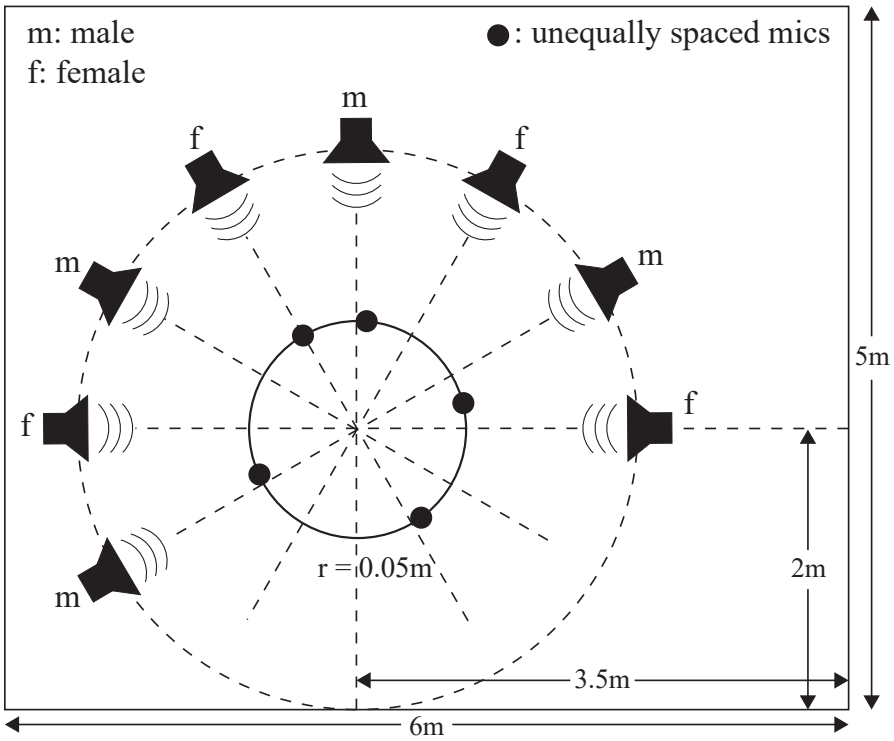

Figure 3.2: *Simulated environment in the experiments.*

a 6-channel unes-CMA, suppose a positive error angle of $\epsilon_1 = 40^\circ$ is added to the 1st microphone initially located at 0°, and a negative error angle of $\epsilon_2 = -40^\circ$ is added to the 2nd microphone initially located at 60°. As a result, the 1st microphone is now positioned at 40°, while the 2nd microphone is positioned at 20°. According to the convention described in Section 3.1, we always designate the microphone closest to 0° as the 1st microphone. Therefore, we exchange the indexing of the two microphones. In the final configuration, the microphone located at 20° is considered the 1st microphone with an error angle of $\epsilon_1 = 20^\circ$, and the microphone located at 40° is considered the 2nd microphone with an error angle of $\epsilon_2 = -20^\circ$.

The angle error for each microphone followed a Gaussian distribution with zero mean and variances ranging from $(0^\circ)^2$ to $(\sqrt{500}^\circ)^2$ in increments of $(\sqrt{10}^\circ)^2$. All the errors

were independently and identically distributed. For each Gaussian distribution with a specific variance, 100 samples were generated. The simulation process proceeded as follows: initially, the sound field was simulated after the unes-CMA rotated Δ rads. Subsequently, this sound signal was employed to estimate the observation signals before rotation at the reference position, with the rotational angle $\phi = \Delta\pi/180^\circ$ being a known value.

Evaluation criteria

In the initial experiment, we assessed the performance in a scenario involving a single source, where sound sources were not mixed. The evaluation was based on the signal-to-error ratio (SER) [42, 43, 75–77] defined as

$$\mathrm{SER}_{m,k} = 10\log_{10}\left(\frac{\sum_t |x_{m,t,k}|^2}{\sum_t |\hat{x}_{m,t,k} - x_{m,t,k}|^2}\right), \tag{3.31}$$

where $x_{m,t,k}$ is the time–frequency domain signal and $\hat{x}_{m,t,k}$ is its estimate. m, t, and k denote the channel, time frame, and frequency bin, respectively. We conducted this experiment by varying the number of microphones, M, within the range of 3–6, and manipulating the rotation angle ϕ.

In the second experiment, we employed the Minimum Power Distortionless Response (MPDR) beamformer, which has been introduced in Section 2.2.2, to compare the source enhancement performance characteristics of different methods. The evaluation was based on the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [78]. From [42, 43], we utilized the covariance matrix of the interference signal and the relative transfer function (RTF) [17, 79] to estimate the beamformer's filter. The RTF was calculated using the RIR from the target source to each microphone.

Then, two sources were randomly selected and mixed into the observation, with an angular separation between them set at $30°, 60°, \ldots, 180°$. This enabled us to simulate 12 environments, with two patterns at each angle.

3.4.2 Results of sound field interpolation

Interpolation accuracy

Initially, we focus solely on the sound source in the direction of 0°. Figure 3.3 presents several examples of SER results obtained using the previous SFI method [42, 43] and the proposed unes-SFI when the rotation angle ϕ ranges from 20° to 30° and M is varied from 4 to 6. The mean SER of all M channels is shown for a specific standard deviation (10°) of the error angle ϵ_i. The results in Figure 3.3 highlight that, in general, the proposed unes-SFI demonstrates superior capability to estimate the spectrum compared with the previous SFI method, achieving a performance increase ranging from a minimum of 5 dB to a maximum of 15 dB. However, it should be noted that higher-frequency components are relatively challenging for both methods because the high-frequency components of speech signals exhibit weaker energy and more rapid variations, making them difficult to estimate accurately. To simplify the analysis, we limited the frequency range to 0–3 kHz for SER evaluation and averaged the SERs in decibels in all subsequent experiments.

Effect of the Nyqf component

To present the results clearly and concisely, we consider two specific distributions of the unes-CMA with $M = 5$ and 6, with a standard deviation of 10°: $\left[-8°, 85°, 150°, 221°, 295°\right]$ and $\left[2°, 46°, 117°, 171°, 253°, 295°\right]$. Here, we focus solely on the sound source in the di-

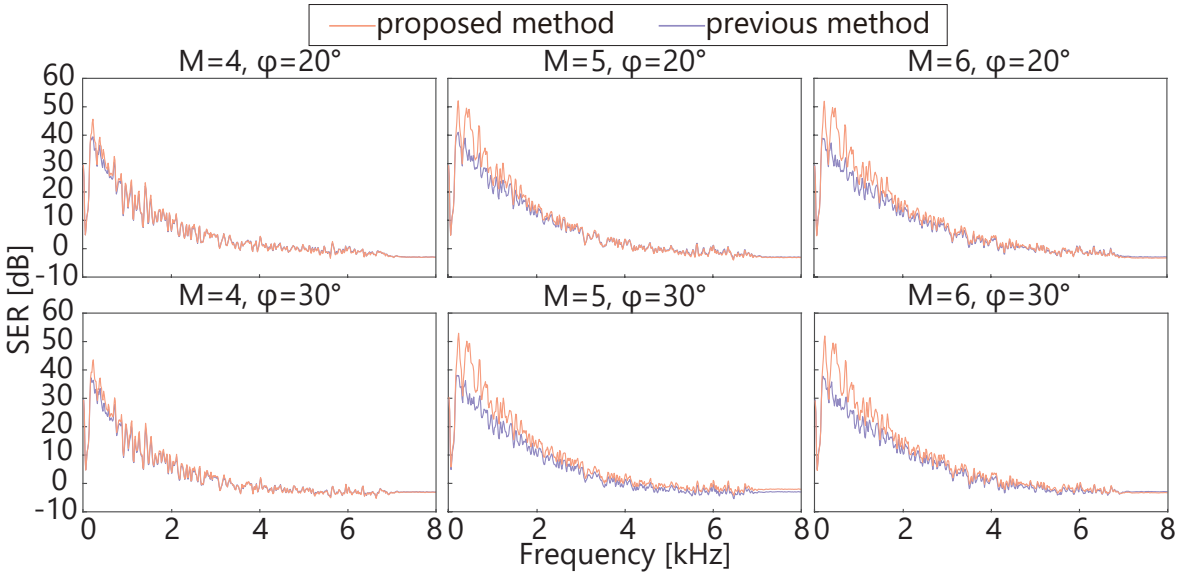

Figure 3.3: *Examples of SERs as a function of frequency.*

rection of 0°. Figures 3.4 and 3.5 show the variation of the SER concerning the rotation angle, with the vertical axis representing the average SER over the frequency range of 0–3 kHz for the first channel, and the horizontal axis illustrating the rotation angle of the unes-CMA. The baseline curve represents the SER without any interpolation. When $M = 6$, the results of the previous SFI method [42, 43] with ReN are obtained, which has been proven to be the most effective method in previous research [43]. Additionally, the SERs of unes-SFI with CoN, ReN, and ZPN are also displayed.

As observed in the figures, owing to erroneously treating the unes-CMA as an es-CMA, the previous SFI method consistently performed worse than unes-SFI when $M = 5$ and unes-SFI with CoN and ReN when $M = 6$. However, it occasionally outperforms unes-SFI with ZPN only at a few rotation angles around 169° and 293°. We also find that the SER of unes-SFI exhibits periodicity, as previously analyzed. When the first channel is rotated to the same position as another channel before rotation, the SER becomes maximum. However, in ZPN with $M = 6$, the SERs significantly

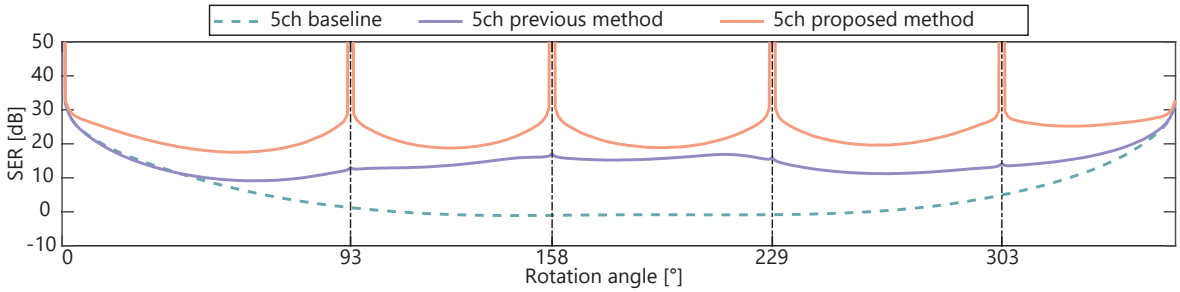

Figure 3.4: *Dependence of SER on the rotation angle with* $M = 5$*, where the baseline indicates the cases without any interpolation.*

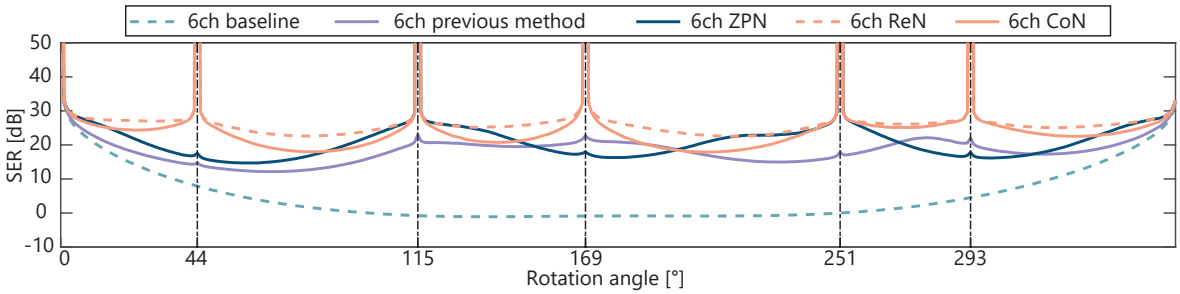

Figure 3.5: *Dependence of SER on the rotation angle with* $M = 6$*, where the baseline indicates the cases without any interpolation, and* ZPN*,* ReN*, and* CoN *indicate neglecting the Nyqf component, considering only the real part of the Nyqf component and employing the Nyqf component's complex value form.*

differ from those in CoN and ReN when the first channel rotates to the second, fourth, and sixth channels, whereas they remain the same for the first, third, and fifth channels. This reflects the effect of ignoring the Nyqf component and supports the earlier conclusion that in ZPN, the interpolation accuracy of a specific microphone cyclically is maximum when the microphone rotates to the position of other microphones before rotation and skips an odd number of microphones simultaneously. Furthermore, when $M = 6$, ReN exhibits slightly higher performance at any rotation angle than CoN. As explained in [42, 43], the reason is that the presence of the complex-valued Nyqf component in CoN may adversely impact the performance of SFI. And it should be noted that the time consumption of ZPN, ReN, and CoN is generally the same be-

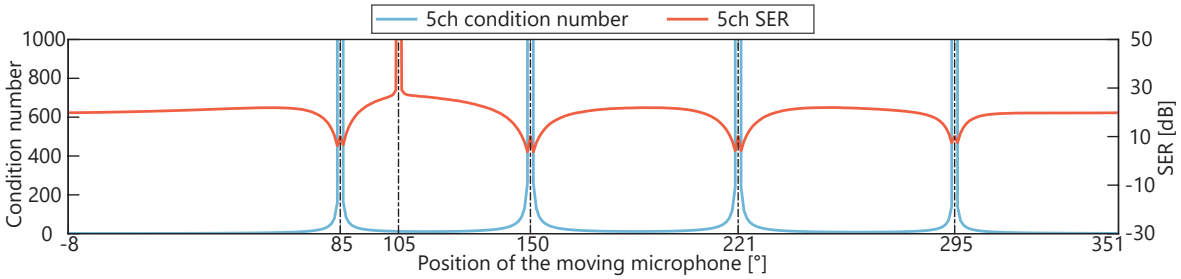

Figure 3.6: *Dependences of condition number and SER on the position of the moving microphone with* $M = 5$.

cause these three methods share the same formulation for calculation, as depicted in (2.25), and the difference among these three methods lies solely in how the Nyqf component is handled. Different approaches to handling this Nyqf component are unlikely to significantly impact the time consumption.

To assess the impact of the Nyqf component on the singularity of the compensation matrix $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, we maintained the same microphone distributions as in the previous experiment, but allowed the position of the first channel's microphone to vary along the unes-CMA. Specifically, we moved it around a circle along the unes-CMA, ranging from -8° to 351° when $M = 5$ and from 2° to 361° when $M = 6$. The condition number of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ was employed to quantify the singularity of $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, where a larger condition number indicates a more singular matrix [80]. At each new position of the moving microphone, we rotated the unes-CMA by 20° and calculated both the condition number and SER.

Figures 3.6 and 3.7 show the dependences of the condition number of the compensation matrix and the SER on the moving microphone position. The vertical axes on the left and right sides, respectively correspond to the condition number and the average SER of the microphone in the second channel of the initial distribution before relocating the microphone in the first channel. Irrespective of the moving microphone position,

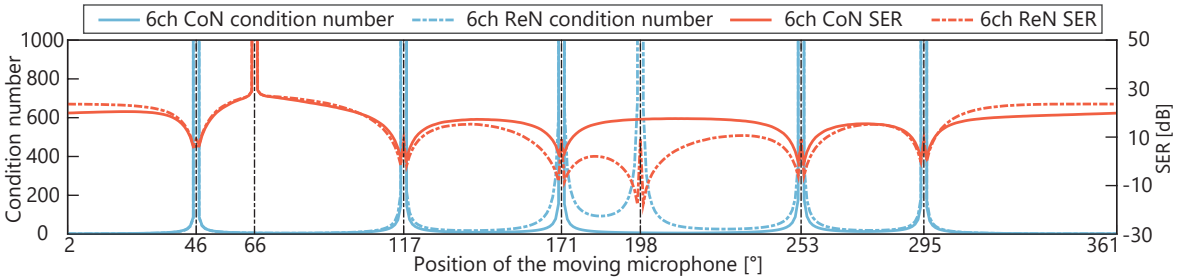

Figure 3.7: *Dependences of condition number and SER on the position of the moving microphone with* $M = 6$*, where* ReN *and* CoN *indicate considering only the real part of the Nyqf component and employing the Nyqf component's complex value form, respectively.*

we consistently utilize the microphone located at 85° ($M = 5$) and 46° ($M = 6$) before rotation to compute the SER result. Note that this choice is maintained even if the moving microphone alters the position order of channels, given the possibility of the channel position being not second after the microphone moved. The horizontal axis represents the moving microphone's position before rotation. For $M = 6$, only results for CoN and ReN are shown, as ZPN was previously found to be ineffective and unsuitable.

As observed in the figures, when the moving microphone is close to another microphone, the corresponding two rows in $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ become similar to each other. This leads to a larger condition number, indicating a more singular $\boldsymbol{U}_M(\boldsymbol{\epsilon})$, owing to which the SER decreases. At certain positions, such as 105° in Figure 3.6 and 66° in Figure 3.7, the SER results are extraordinarily high, because after rotating by 20°, the microphone used to calculate the SER aligns with the position of the moving microphone before rotation.

For $M = 5$ and $M = 6$ in CoN, there are four and five positions of the moving microphone causing the abnormally large condition number, respectively, where these positions coincide with the other fixed microphones. However, when $M = 6$

in ReN, there are six such positions, although we expected only five. The additional position angle is 198°. At this position, the 6-channel distribution in unes-CMA is $\left[46°, 117°, 171°, 198°, 253°, 295°\right]$. Thus, the error vector $\boldsymbol{\epsilon}$ is $\left[46°, 57°, 51°, 18°, 13°, -5°\right]$. This arrangement leads to the sum of all error angles being equal to 180°, which validates our previous conclusion in (3.30).

Note that the SERs at these singular positions are slightly higher than those nearby, owing to the application of truncated SVD at these positions, which reduces $\boldsymbol{U}_M(\boldsymbol{\epsilon})$'s singularity.

From Figure 3.5, when $M = 6$, the interpolation accuracy of CoN is slightly lower than that of ReN, but the degree of degeneracy is minimal and acceptable. From Figure 3.7, it is evident that ReN can lead to an unexpected abnormal situation where the proposed unes-SFI fails owing to a singular compensation matrix. Consequently, CoN is deemed the most reasonable approach to handling the Nyqf component compared with the other two methods. Therefore, in subsequent evaluations, CoN will be employed when M is an even number.

Robustness to the variance of angle error

Figure 3.8 shows the relationship between the variance of the error angles and the SER improvement for cases where Ms are 5 and 6, and ϕ is 20°, with the sound source located at 0°. The SER improvement quantifies the increase in SER achieved through signal processing. The baseline used in SER improvement is obtained without any processing, where the SER is computed by comparing the uninterpolated signal after rotation with the target signal before rotation. Each box in the graph represents the mean SER improvement over M channels for each sample, resulting in 100 data points in each box.

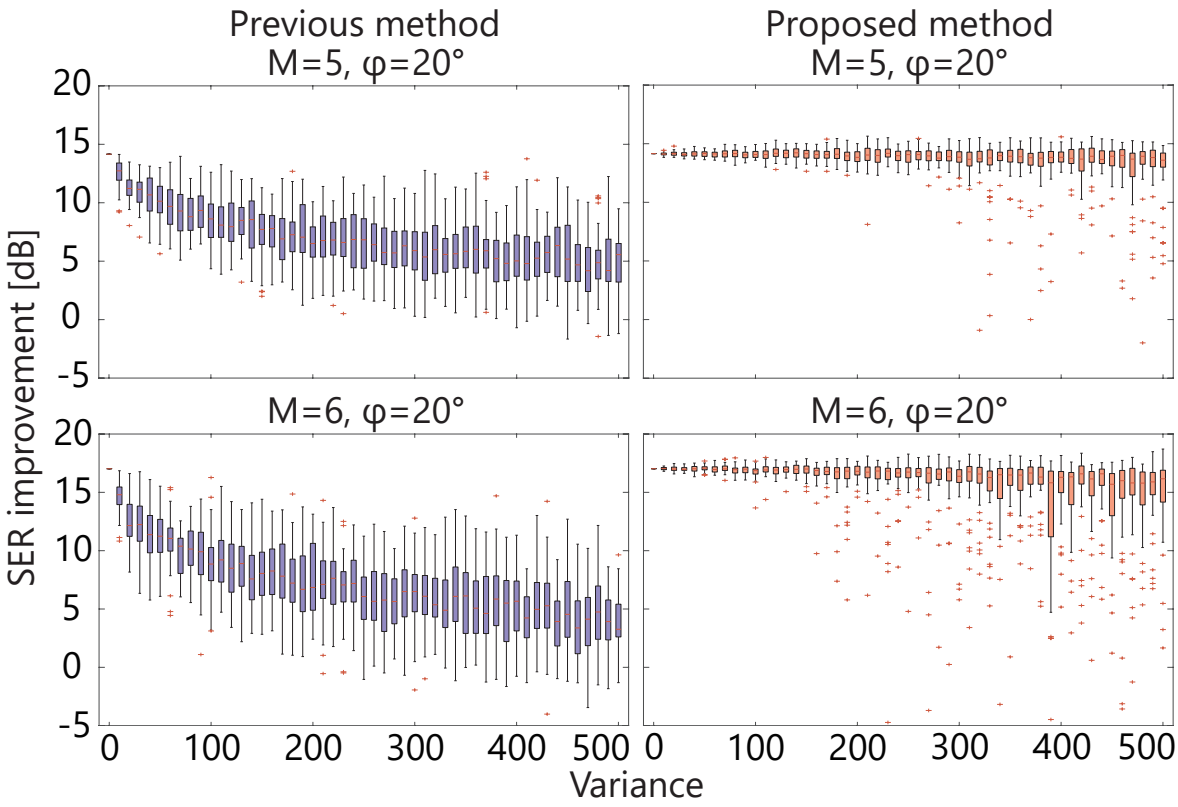

Figure 3.8: *Boxplots of the relationship between the variance of the error angle and the SER improvement at frequencies up to 3 kHz relative to the cases without interpolation.*

The results clearly demonstrate that as the variance of the errors of angles increases, the SER improvement of the previous SFI method [42,43] experiences significant degradation from approximately 15 dB to as low as 4 dB. In contrast, our proposed method maintains its excellent interpolation performance at 15 dB, even with substantial errors of angles. This underscores the impracticality of directly applying ordinary SFI to an unes-CMA and highlights the advantages of employing our novel technique.

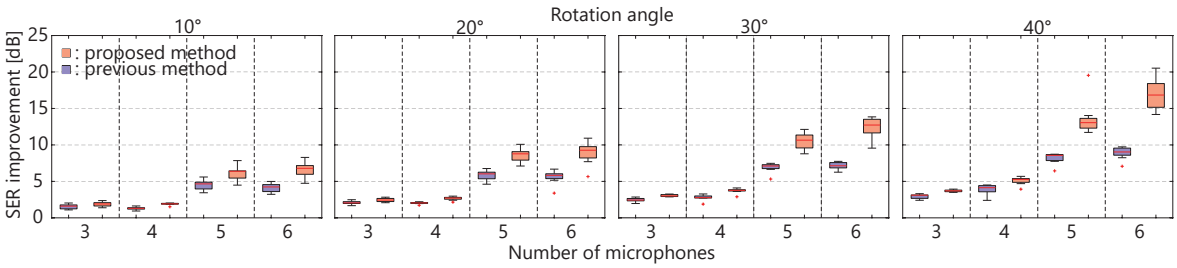

Figure 3.9: *Boxplots of mean SER improvement at frequencies up to 3 kHz for M channels in eight situations.*

Channelwise SER improvements

Figure 3.9 illustrates the channelwise SER improvements obtained for various numbers of microphones M and rotation angles ϕ with the standard deviation of error set to 10°. Here, we use eight sound sources situated in different directions. The mean SER improvement relative to cases without interpolation is calculated over M channels. Each box in Figure 3.9 contains eight samples, corresponding to the mean SER improvement of eight sound sources.

The results demonstrate that the proposed method consistently exhibits greater SER improvement than the previous technique [42,43] across all situations. Specifically, the minimum improvement occurs at $M = 3$ and $\phi = 10°$ with an increase of 1 dB, while the maximum improvement is observed at $M = 6$ and $\phi = 40°$, reaching approximately 8 dB. As anticipated, in the proposed method, an increase in the number of microphones leads to enhanced performance owing to the higher spatial sampling rate. Conversely, in the previous SFI method, using more microphones does not always lead to an improved SER and may even result in a poorer performance. This is observed when changing the number of microphones M from 3 to 4 and from 5 to 6. This phenomenon can be attributed to the introduction of more errors in the rotation transform matrix $\boldsymbol{U}_M(\Delta)$ with more microphones, where the benefits of a higher spatial sampling

rate do not outweigh the adverse effects of errors.

Furthermore, the proposed method achieves a greater SER improvement with an increase in the rotation angle. This can be attributed to the inferior performance of the case without any processing at higher rotation angles, where the proposed method's capability to compensate for microphone positions becomes more crucial.

Robustness to the error in rotation angle

The rotation angle is a critical prior knowledge that must be known in advance for our proposed method. In cases where the rotation angles are inaccurately measured or estimated, the performance of our method could be affected. In this subsection, we explore a scenario where the rotation angle is not accurately measured and investigate the robustness of our proposed method to errors in the rotation angle. In various methods for rotation angle estimation, the error in rotation angle estimation can be limited to within 5°. Consequently, we present the channelwise SER improvement under varying errors in rotation angle estimation, ranging from $-5°$ to $5°$. For simplicity, here we only focus on the situations where the number of microphones M is 5 and 6, rotation angle ϕ is 10° and 30°. We continue to employ eight sound sources positioned in various directions. The mean SER improvement, relative to cases without interpolation, is computed over M channels. Each box encompasses eight samples, corresponding to the mean SER improvement of eight sound sources. As depicted in Figure 3.10, a degradation in SER improvement is observed with an increase in the absolute value of the error in rotation angle estimation, aligning with our expectations. However, the extent of degradation is not considerable, with a reduction of less than 3 dB. Despite this, our method remains effective to a certain degree in reconstructing the sound signal before rotation.

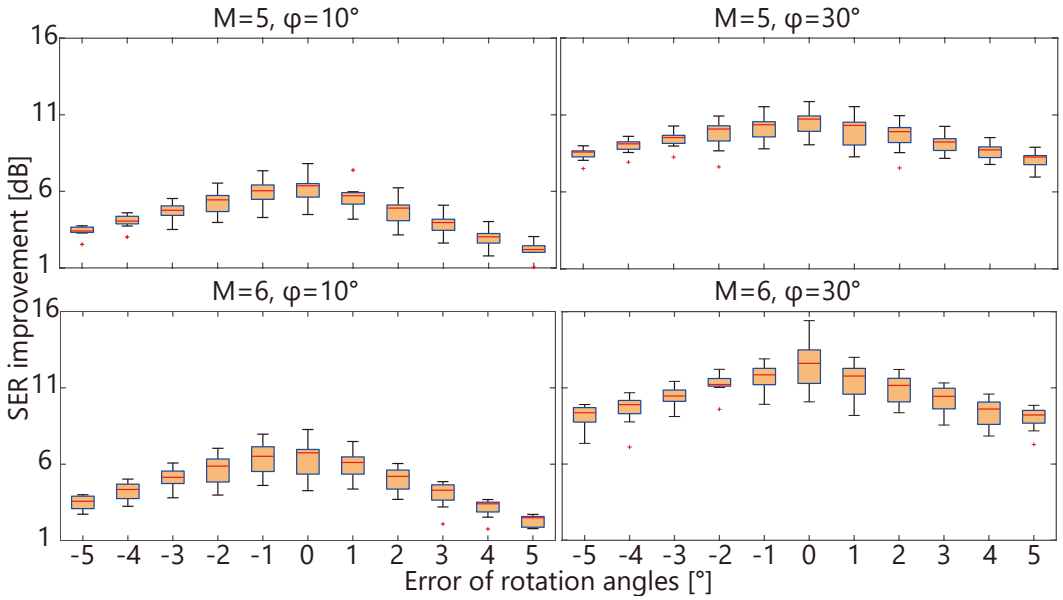

Figure 3.10: *Boxplots of mean SER improvement under different errors of rotation angle estimation.*

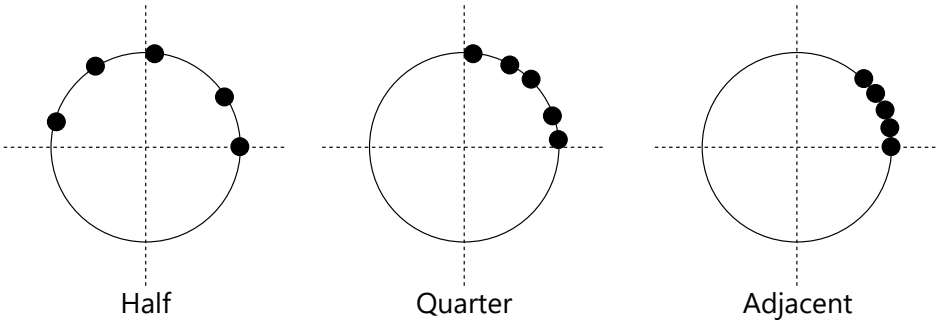

Figure 3.11: *Examples of extreme distributions.*

Influence of microphone distributions

The microphones in the unes-CMA are generally distributed unevenly throughout the circular array, as shown in Figure 3.2. However, in some atypical cases, the microphones may exhibit central clustering within certain regions of the unes-CMA. In this study, we considered three such extreme distributions when the number of microphones is 5: distributions spanning only half of the circle (**Half**), distributions spanning only

a quarter of the circle (**Quarter**), and distributions with microphones placed next to each other at an angular interval of 1° (**Adjacent**). Examples of these extreme distributions are shown in Figure 3.11. The baseline configuration corresponds to the typical scenario where microphones are unequally spaced throughout the entire circle (**Baseline**).

To evaluate the generalization capacity of the proposed method, we analyzed the SER improvement results at rotation angles of 20°, 100°, and 190°, as illustrated in Figure 3.12. The rotation angles of 100° and 190° were chosen to examine the performance of **Half** and **Quarter** when the microphones' positions after rotation fall outside the before-rotation distribution range. The results indicate that the proposed method can accurately estimate the target signal with **Half** and **Quarter** at a rotation angle of 20°. It is also noteworthy that the proposed method performs better under the **Quarter** than under the **Half**, and both outperform the **Baseline**. This is because a rotation angle of 20° does not cause all microphones to fall outside the original distribution range. Moreover, the more concentrated microphone distributions in the **Quarter** and **Half** provide a higher spatial sampling rate around the after-rotation positions, thereby leading to more accurate interpolation results. However, when rotating by 100° and 190°, the proposed method demonstrates some effectiveness only for **Half**, although the SER improvement is notably $5-10\,\mathrm{dB}$ smaller than that for the **Baseline**. **Adjacent** presents a significant challenge. In the case of **Adjacent**, it is observed that the proposed method cannot provide satisfactory results regardless of the rotation angle. These behaviors are expected as **Quarter** and **Adjacent** exhibit a small sampling range and limited spatial information available for interpolation beyond the original distribution.

Surprisingly, **Adjacent**'s performance is less inferior than **Quarter**'s when the ro-

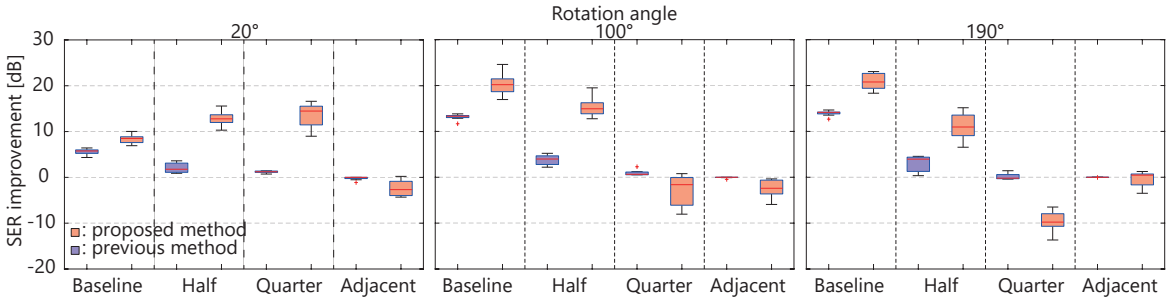

Figure 3.12: *Boxplots of mean SER improvement at frequencies up to 3 kHz relative to the cases without interpolation in extreme distributions, where* ***Baseline****,* ***Half****,* ***Quarter****, and* ***Adjacent*** *represent distributions throughout the entire circle, distributions spanning only half of the circle, distributions spanning only a quarter of the circle, and distributions with microphones placed next to each other, respectively.*

tation angles are 100° and 190°, contrary to our initial expectations. This unexpected outcome can be attributed to the application of truncated SVD in 3.3.2, which effectively mitigated the issues with the compensation matrix for **Adjacent**, preventing extremely erratic performance. If the truncated SVD were not applied, all rows in $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ of **Adjacent** would be similar to each other, leading to a nearly singular $\boldsymbol{U}_M(\boldsymbol{\epsilon})$ and a significantly deteriorating SER improvement. Thus, the observed instances where the performance of the previous SFI method [42, 43] is not as poor as that of the proposed method can be attributed to the previous SFI method's lack of necessity to address the sampling range. In addition, the previous SFI method does not handle the ill-conditioned compensation matrix.

3.4.3 Results of source enhancement with batch processing

In this experiment, we evaluate source enhancement performance using the MPDR beamformer. We fix the number of microphones at $M = 5$ and vary the rotation angle ϕ to 10°, 20°, 30°, and 40°. Firstly, we compute the filter weight $\boldsymbol{w}$ for the MPDR

Table 3.1: *Abbreviations for spectrograms from different evaluation settings*

Spectrograms from different evaluation settings	Status of evaluation conditions					
	B	R	I	S	L	U
B0: No beamforming	0	0	0	0	0	0
R0: No rotation	1	0	0	0	0	0
I0: No interpolation	1	1	0	0	0	0
S1: SFI	1	1	1	1	0	0
L1: CSM-LI	1	1	1	0	1	0
U1: proposed unes-SFI	1	1	1	0	0	1

beamformer using the RTF and the multichannel STFT spectrogram obtained from the unes-CMA at its original microphone position before any rotation. Then, the weight $\boldsymbol{w}$ is applied to this spectrogram without rotation; this reference performance is denoted by R0. This serves as the most favorable scenario, as it uses true signals instead of interpolated signals for the MPDR beamformer. In the subsequent content, for simplicity in naming spectrograms from different methods, we use a set of abbreviations to represent the evaluation conditions: B represents Beamforming, R for Rotation, I for Interpolation, S for SFI [42, 43], L for CSM-LI method [40] which employs the linear interpolation method only on the two neighboring microphones in the time domain, U for unes-SFI, Index 0 for off, and Index 1 for On.

Therefore, the various spectrograms from different evaluation settings, which we subsequently postprocessed using the MPDR beamformer's weight $\boldsymbol{w}$ to generate the estimated target signal, are summarized in the Table 3.1. We used the unprocessed case (B0), where the microphone signal was mistakenly treated as the target signal without applying beamforming, and R0 as baselines for comparison.

Additionally, we recalculated a new MPDR beamformer using the same RTF before

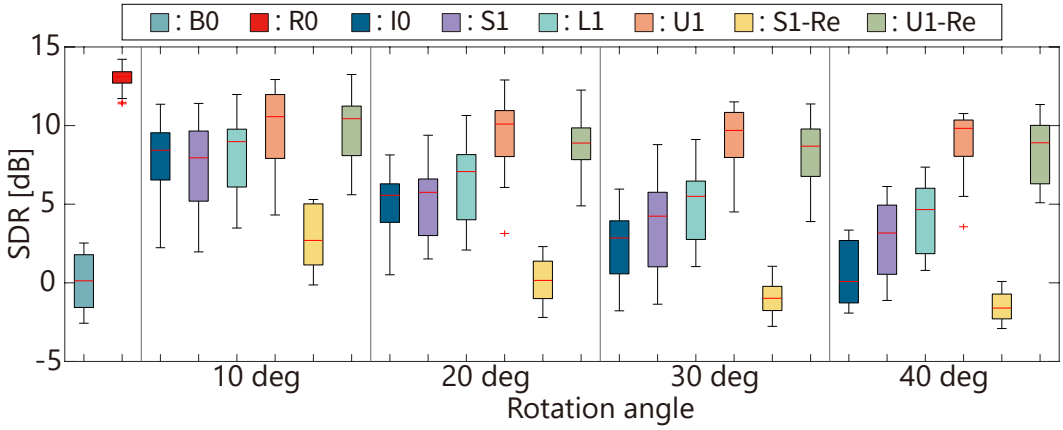

Figure 3.13: *Boxplots of SDR obtained by MPDR beamformer in different situations: unprocessed* (B0)*, no rotation of the CMA* (R0)*, without interpolation when the CMA rotates* (I0)*, with ordinary SFI when the CMA rotates* (S1)*, CSM-LI when the CMA rotates* (L1)*, unes-SFI when the CMA rotates* (U1)*, re-estimation of the filter after ordinary SFI when CMA rotates* (S1-Re)*, and re-estimation of the filter after unes-SFI when CMA rotates* (U1-Re)

rotation and the interpolated spectrograms from S1 and U1, and applied this newly developed MPDR beamformer to the interpolated spectrograms; the results were denoted by S1-Re and U1-Re, respectively. These results provide insights into the performance of online beamforming described in the next experiment.

The SDR and SIR results for different scenarios with a standard deviation of error set to 10° are presented in Figure 3.13 and Figure 3.14. As expected, B0 exhibits the lowest SDR and SIR (0 dB), whereas R0 achieves the most significant source enhancement performance since the ATS remains time-invariant. Interestingly, the S1 approach does not perform as well as anticipated. In most environments, S1's SDR and SIR show little difference from I0's SDR and SIR, and in some circumstances, S1's performance is even inferior to I0's performance. These findings indicate that the previous SFI method is ineffective when the CMA undergoes rotation owing to the non-uniformity of microphone spacing. In source enhancement using the MPDR beamformer, the

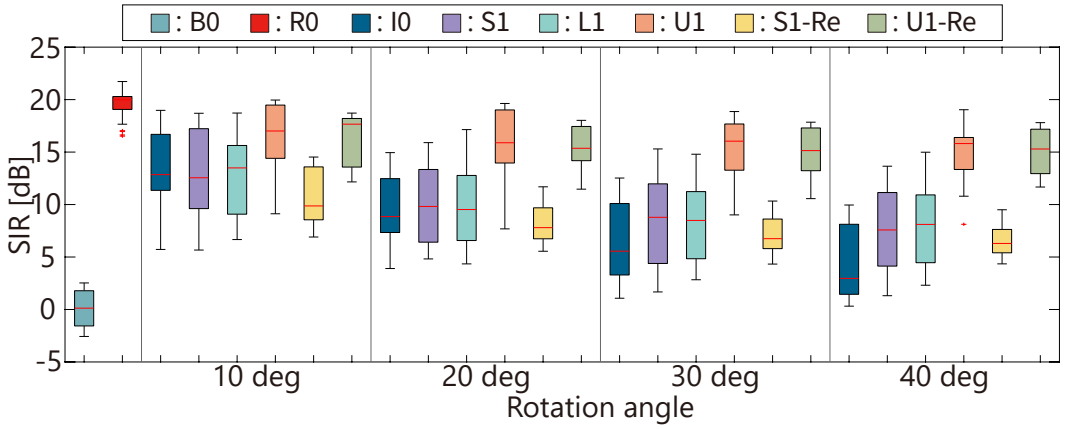

Figure 3.14: *Boxplots of SIR obtained by MPDR beamformer in different situations: unprocessed* (B0)*, no rotation of the CMA* (R0)*, without interpolation when the CMA rotates* (I0)*, with ordinary SFI when the CMA rotates* (S1)*, CSM-LI when the CMA rotates* (L1)*, unes-SFI when the CMA rotates* (U1)*, re-estimation of the filter after ordinary SFI when CMA rotates* (S1-Re)*, and re-estimation of the filter after unes-SFI when CMA rotates* (U1-Re)

previous interpolation provides only a slight improvement, and it is likely that a better source enhancement can be achieved without employing the previous SFI method. In contrast, the proposed method (U1) outperforms in both the case without interpolation (I0) and that with the previous interpolation technique (S1 and L1) by approximately $3-9\,\mathrm{dB}$ and $2-7\,\mathrm{dB}$, respectively. Moreover, it approaches the performance of the best-case scenario (R0), with a difference of less than $4\,\mathrm{dB}$, regardless of the type of simulated environment used. The unequally spaced interpolation method demonstrates robustness to the non-uniform distribution of microphones on the CMA, significantly enhancing the array signal processing performance.

Furthermore, U1-Re performs similarly to U1, suggesting that unes-SFI can likely improve online processing as well, with only a slight decline of less than $1\,\mathrm{dB}$ due to the small mismatch between the covariance matrix estimated from the interpolated spectrogram and the pre-estimated RTF. Conversely, S1-Re exhibits poor source en-

hancement results, and is almost as ineffective as B0. One of the main reasons for such degraded performance is that the previous SFI method cannot precisely interpolate the spectrogram before rotation, resulting in a covariance matrix that is entirely mismatched with the RTF.

3.4.4 Results of source enhancement with online processing

In this experiment, we propose the utilization of SFI for beamforming in an online processing scenario, taking into account continuous dynamic changes in the ATS. Online processing is designed to effectively handle minor variations in the ATS. As outlined in Section 2.3.3, to achieve this objective, we introduce a common smoothing factor denoted by α [64, 65], which enables the updating of spatial covariance during online processing. Additionally, we use the matrix inversion lemma, particularly the Sherman–Morrison formula [66–68], to alleviate the computational complexity associated with the covariance inversion in the MPDR formulation. By employing these techniques, we aim to increase the efficiency and effectiveness of the beamforming process in the presence of ATS variations.

It is noteworthy that the algorithm employed for online beamforming in this experiment bears a similarity to that utilized in the previous research [43]. The experimental conditions closely resemble those described in 3.4.1. However, there are some differences as follows. Two source signals were utilized, each with a duration of 40 s. Additionally, we simulated the impulse response with reverberation times of 100 ms and 500 ms. The positions of the two sources were located at angles of 60° and 150°, following the alignment shown in Figure 3.2. The frame length was set to 256 ms, and a segmental SDR with a length of 1 s was employed to evaluate source enhancement performance. For the smoothing factor α, we selected a value of 0.99, which has been

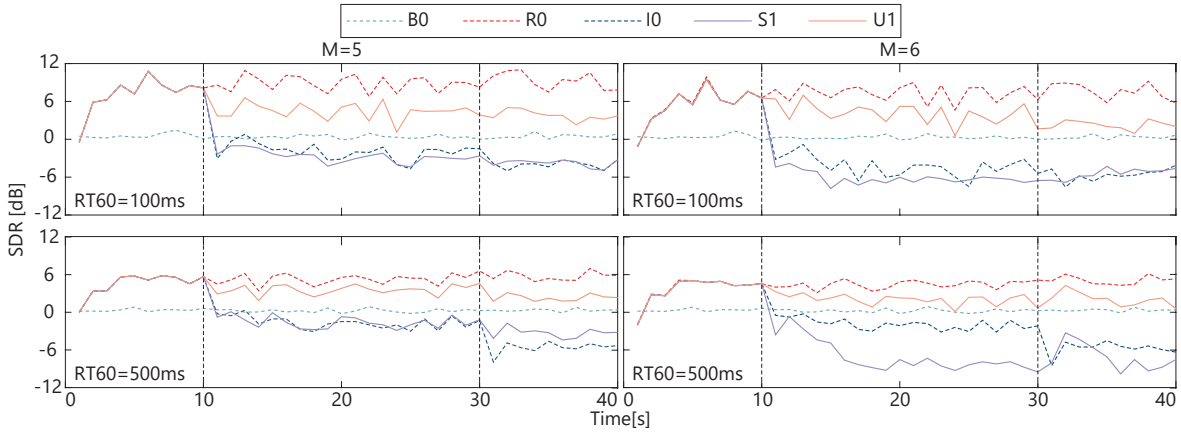

Figure 3.15: *Segmental SDR every* 1 s *with* $M = 5$ *and* 6 *and RT60* = 100 ms *and* 500 ms*, where the two vertical dashed lines indicate the time points when the rotation started:* $0^\circ \Rightarrow 20^\circ \Rightarrow 40^\circ$. B0 *shows the mixture itself,* R0 *shows the case where rotation does not occur,* I0 *and* S1 *respectively show online processing without and with ordinary interpolation, and* U1 *shows online processing with the unes-SFI.*

empirically validated to produce the highest segmental SDR. To initialize $\hat{\boldsymbol{V}}_f^{-1}$, we used the inversion of the covariance matrix over the first 10 frames. During the experiment, the unes-CMA underwent two rotations: the first rotation commenced at 10 s, progressing from 0° to 20°, and the second rotation began at 30 s, spanning from 20° to 40°. Notably, the unes-CMA did not instantaneously rotate at 10 s or 30 s but rather underwent a gradual rotation at a uniform speed of 0.01° per time sample (equivalent to 160° per second). This rotation speed aligns with the typical average rotation speed for humans or humanoid robots. The observations in this experiment were generated by concatenating the observations of the unes-CMA after rotating at different angles in the simulation.

Figure 3.15 presents the segmental SDR results obtained with $M = 5$ and 6, and reverberation times of 100 ms and 500 ms. As shown, the R0 scenario, where the unes-CMA does not rotate, consistently achieves the most effective source enhancement performance. Surprisingly, unlike batch processing, the B0 method does not

yield the lowest SDRs, contrary to initial expectations. The S1 method exhibits the poorest performance at about −6 dB, sometimes even worse than the I0 method in some scenarios. These observations clearly indicate that the previous SFI technique [42, 43] is entirely ineffective when applied to an unes-CMA in online beamforming processing.

In comparison, our proposed method (U1), which achieved an average improvement of 9 dB in SDR score, exhibits a significantly improved performance compared with S1, achieving results closest to the highest performance (R0) even in challenging environments with long reverberation time. This demonstrates the robustness and effectiveness of our approach in the context of online beamforming processing.

3.5 Conclusions

In this chapter, we proposed a novel framework for rotation-robust beamforming on an unes-CMA, building upon and extending prior research. By enhancing the simple SFI method, we developed the unes-SFI approach, incorporating a compensation matrix and adapting the SFI technique. This framework effectively enables the transformation of the time-variant ATS on an unes-CMA into a time-invariant ATS on an es-CMA, allowing for the estimation of the unes-CMA signal before rotation and achieving rotation-robust beamforming. We also conducted an in-depth analysis of the compensation matrix's properties and the influence of the Nyqf component.

Through a series of comprehensive simulated experiments, we have systematically evaluated the performance and robustness of the proposed interpolation framework across several critical dimensions. First, we demonstrated that the framework achieves high interpolation accuracy, successfully reconstructing before-rotation signals. Further analysis confirmed that the presence of Nyqf component has a significant impact on interpolation performance, thereby validating our theoretical analysis.

The proposed framework also exhibits strong robustness. It maintains stable interpolation performance even under increasing error angles and shows resilience to inaccuracies in the estimated rotation angle. Channelwise SER evaluations further confirmed that the method provides notable improvements in signal estimation compared to previous approaches across a range of scenarios. Moreover, we investigated the influence of microphone distribution on interpolation performance. The results indicate that microphone configuration plays a critical role in overall performance, highlighting the need for future research to mitigate this dependency.

Finally, the proposed method demonstrated consistent improvements in downstream array signal processing tasks, such as source enhancement using the MPDR beamformer, further underscoring its practical effectiveness and adaptability for rotation-robust processing.

However, the method has limitations that warrant further exploration. In this study, we assume prior knowledge of microphone position errors, which may not be available in practical scenarios. Investigating the application of SFI on an unes-CMA without access to such information is a compelling research direction, which we aim to pursue in the next chapter.

4 Generalized Sound Field Interpolation with Unsupervised Position Calibration

In this chapter, we introduce a novel method called Generalized Sound Field Interpolation (GSFI) designed to achieve rotation-robust beamforming using circular microphone arrays (CMAs) with unknown microphone distributions. While the unequally spaced SFI (unes-SFI) method presented in the previous chapter provides a robust solution for non-uniform arrays, it relies on prior knowledge of the microphone positions. This requirement limits its practical applicability in real-world scenarios where such information is typically unavailable. To overcome this limitation, we propose a method that integrates unsupervised calibration with the unes-SFI approach, enabling effective interpolation and beamforming for unequally spaced circular microphone arrays (unes-CMAs) with unknown microphone positions.

Unsupervised calibration employs a novel iterative optimization technique to estimate microphone positional errors without any pre-existing information about their locations. This process iteratively adjusts the estimated positions, leveraging observed data to converge toward an accurate representation of the microphone distribution on the unes-CMA. Once the positional errors are determined, the unes-SFI framework can reconstruct the target signal for the unes-CMA before rotation, effectively mitigating

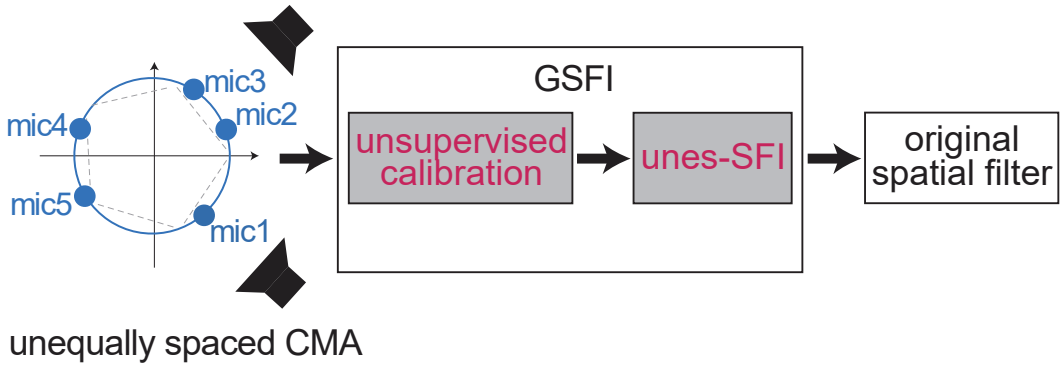

Figure 4.1: *Conceptual diagram of GSFI. GSFI on an unes-CMA encompasses two components: unsupervised calibration and unes-SFI.*

the effects of rotation and microphone positioning errors.

Additionally, we reduce the computational complexity of the unsupervised calibration process by refining the cost function used during optimization. Simulation experiments were conducted to assess the performance and robustness of the proposed approach under various conditions. The results demonstrated that the method effectively mitigates the adverse effects of unknown microphone placements. It achieved substantial improvements in signal estimation before rotation and beamforming, outperforming previous approaches and highlighting its potential for practical applications in real-world scenarios.

4.1 Overview

The GSFI method is specifically designed to address the critical challenge of time-variant Acoustic Transfer System (ATS) on a microphone array with an unknown microphone distribution. As emphasized earlier, GSFI shares the same overarching goal with the approaches introduced in Chapters 2 and 3: eliminating the need for spatial filter re-estimation following the rotation of a CMA.

As illustrated in Figure 4.1, the GSFI framework for an unes-CMA consists of two primary components: unsupervised calibration and unes-SFI. At the heart of this framework lies the concept of unsupervised calibration, illustrated in Figure 4.2, which involves iterative optimization of the error vector, $\boldsymbol{\epsilon}$. This approach facilitates calibration for each microphone using only the multichannel microphone signals, without requiring any prior information about the microphone distribution.

Unsupervised calibration unfolds in three sequential steps:

- **Compensate for the angular deviation**: An initial set of error values is assumed, and one channel signal is designated as the reference signal. The remaining $(M-1)$ channel signals are treated as pseudo-observations. Then we compensate for the angular deviations between each microphone, whose signal is chosen as a pseudo-observation, and its corresponding microphone on a virtual equally spaced $(M-1)$-channel CMA.

- **Reference signal estimation via SFI**: The signals from the virtual equally spaced $(M-1)$-channel CMA are interpolated using SFI to estimate the reference signal.

- **Error vector optimization**: The estimated reference signal is compared to the actual reference signal, and the error vector, $\boldsymbol{\epsilon}$, is optimized by minimizing a predefined cost function that quantifies the discrepancy between the two signals.

Once the error vector is obtained through this process, it can be seamlessly integrated into the unes-SFI framework to reconstruct the before-rotation signal of the unes-CMA

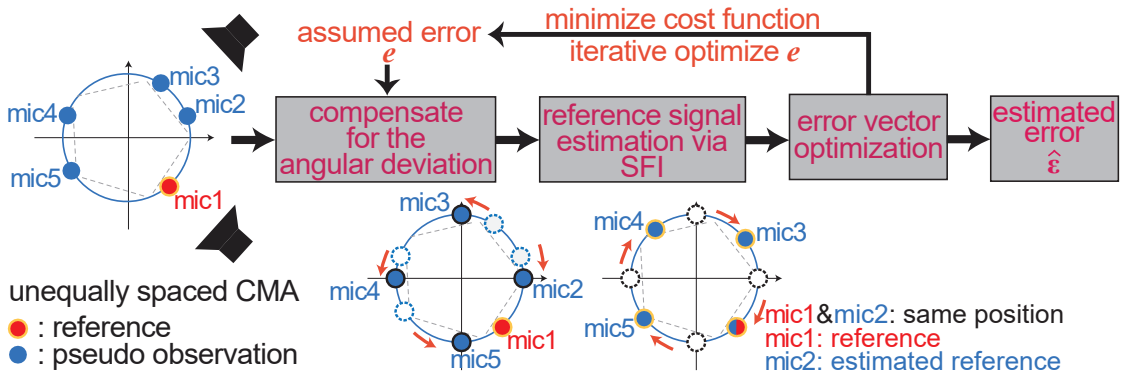

Figure 4.2: *Conceptual diagram of unsupervised calibration. Unsupervised calibration comprises three steps: compensate for the angular deviations, estimate the reference by SFI, minimize the cost function to optimize* $\boldsymbol{e}$.

4.2 Formulation

In this section, we address the challenge of interpolation for a freely spaced unes-CMA with an unknown distribution. Recognizing the physical constraints of practical applications, we assume that microphones cannot occupy identical positions simultaneously, which ensures a valid array configuration. Unlike the unes-SFI method, which relies on precise knowledge of each microphone's position, our approach eliminates this requirement. Instead, it operates with minimal prior information, necessitating only the number of microphones in the array and the rotation angle, which can be readily determined [45, 62, 63], making GSFI more adaptable to real-world scenarios where detailed positional information may not be available.

The observation obtained from the unes-CMA is represented by the same equation as (3.1):

$$\boldsymbol{x}(\boldsymbol{\epsilon}) = \left[x\left(\tfrac{2\pi\cdot 0}{M} + \epsilon_1\right) \quad \cdots \quad x\left(\tfrac{2\pi(M-1)}{M} + \epsilon_M\right) \right]^{\mathsf{T}}. \tag{4.1}$$

Given the unknown actual microphone distribution, we initially assume a distribution

for the unes-CMA. To facilitate this assumption, we introduce a known error vector, denoted as $\boldsymbol{e} = \begin{bmatrix} e_1 & \cdots & e_M \end{bmatrix}^{\mathsf{T}}$, which is initially set to a zero vector to substitute for the unknown actual error angle vector $\boldsymbol{\epsilon}$. Consequently, using $\boldsymbol{e}$, the initial position of each microphone is determined as

$$\boldsymbol{P}(\boldsymbol{e}) = \begin{bmatrix} 0 + e_1 & \frac{2\pi \cdot 1}{M} + e_2 & \cdots & \frac{2\pi(M-1)}{M} + e_M \end{bmatrix}^{\mathsf{T}}. \tag{4.2}$$

Subsequently, the process begins by partitioning the multichannel microphone signals into two distinct groups: the reference signal and the pseudo-observations. For example, the signal from the first microphone channel is designated as the reference signal, $x_{\text{ref}} = x(0 + \epsilon_1)$, while the signals from the remaining $M - 1$ channels are categorized as pseudo-observations, which are illustrated as

$$\boldsymbol{x}_{\text{psd}}(\boldsymbol{\epsilon}) = \begin{bmatrix} x\left(\frac{2\pi}{M} + \epsilon_2\right) & \cdots & x\left(\frac{2\pi(M-1)}{M} + \epsilon_M\right) \end{bmatrix}^{\mathsf{T}}. \tag{4.3}$$

These pseudo-observations are treated as if they were recorded by an $(M-1)$-channel unes-CMA. Within this context, the initial step of unes-SFI is applied to $\boldsymbol{x}_{\text{psd}}(\boldsymbol{\epsilon})$, using (3.6) and the positional information $\boldsymbol{P}(\boldsymbol{e})$ to compensate for the positional errors of this $(M-1)$-channel unes-CMA. As a result, the signal corresponding to a virtual $(M-1)$-channel es-CMA, denoted by

$$\boldsymbol{x}_{\text{psd}}(\boldsymbol{0}) = \begin{bmatrix} x(0) & \cdots & x\left(\frac{2\pi(M-2)}{M-1}\right) \end{bmatrix}^{\mathsf{T}}, \tag{4.4}$$

can be computed as

$$\boldsymbol{x}_{\text{psd}}(\boldsymbol{0}) = \boldsymbol{U}_{M-1}(\boldsymbol{e}_{\setminus e_1})^{-1}\boldsymbol{x}_{\text{psd}}(\boldsymbol{\epsilon}), \tag{4.5}$$

where $\boldsymbol{e}_{\setminus e_1}$ denotes a vector that excludes e_1 from $\boldsymbol{e}$, and the compensation matrix

$\boldsymbol{U}_{M-1}(\boldsymbol{e}_{\setminus e_1})$ is calculated as

$$\boldsymbol{U}_{M-1}(\boldsymbol{e}_{\setminus e_1}) = \begin{bmatrix} \boldsymbol{v}_1\left(\frac{2\pi\cdot 1}{M} + e_2 - 0\right) \\ \vdots \\ \boldsymbol{v}_{M-1}\left(\frac{2\pi(M-1)}{M} + e_M - \frac{2\pi(M-2)}{M-1}\right) \end{bmatrix}. \tag{4.6}$$

Here, $\boldsymbol{v}_m(\epsilon_n) \in \mathbb{C}^{1\times(M-1)}$ represents the mth row of the rotation matrix $\boldsymbol{U}_{M-1}(\epsilon_n) \in \mathbb{C}^{(M-1)\times(M-1)}$.

Upon acquiring the equally spaced signal $\boldsymbol{x}_{\mathrm{psd}}(\mathbf{0})$, conventional SFI can be employed to estimate the reference signal, expressed as

$$\hat{x}_{\mathrm{ref}} = \boldsymbol{v}_1(e_1)\boldsymbol{x}_{\mathrm{psd}}(\mathbf{0}). \tag{4.7}$$

It is crucial to underscore that in this method, the computation of $\boldsymbol{U}_{M-1}(\boldsymbol{e}_{\setminus e_1})$ and $\boldsymbol{v}_1(e_1)$ is based on the assumed error vector $\boldsymbol{e}$, rather than the actual error vector $\boldsymbol{\epsilon}$. This distinction arises because $\boldsymbol{e}$ is the only accessible variable, while $\boldsymbol{\epsilon}$ remains unknown.

The cost function can be defined as the difference between $\hat{x}_{\mathrm{ref}}$ and x_{ref}, expressed as follows:

$$\mathcal{L}_1(\boldsymbol{e}) = 10\log_{10}\Big(\sum_{\substack{t\in\{1,\dots,T\}\\ f\in\{1,\dots,F\}}} |\hat{x}_{\mathrm{ref},t,f} - x_{\mathrm{ref},t,f}|^2\Big), \tag{4.8}$$

where t and f denote the indices for the time frame and frequency bin, respectively, while T and F represent the total number of time frames and frequency bands, respectively.

Subsequently, we designate the signal from the second channel, $x(2\pi/M + \epsilon_2)$, as the reference and apply the same method described earlier to compute a new cost

function, $\mathcal{L}_2(\boldsymbol{e})$. This process is repeated for all M channels, resulting in M individual cost functions denoted as $\mathcal{L}_1(\boldsymbol{e}), \mathcal{L}_2(\boldsymbol{e}), \ldots, \mathcal{L}_M(\boldsymbol{e})$, each corresponding to a different reference signal. By aggregating these cost functions, we construct a composite cost function

$$\mathcal{L}(\boldsymbol{e}) = \sum_{i=1}^{M} \mathcal{L}_i(\boldsymbol{e}). \tag{4.9}$$

Although $\mathcal{L}(\boldsymbol{e})$ does not lend itself to a closed-form solution, it is differentiable with respect to $\boldsymbol{e}$, enabling the application of backpropagation to estimate the error vector $\boldsymbol{\epsilon}$. To minimize $\mathcal{L}(\boldsymbol{e})$, we utilize gradient descent with the Adam optimizer [81] for iterative optimization:

$$\hat{\boldsymbol{\epsilon}} = \arg\min_{\boldsymbol{e}} \mathcal{L}(\boldsymbol{e}). \tag{4.10}$$

Despite the nonconvex nature of (4.9), which allows for the existence of multiple local minima, the function possesses a global minimum. Achieving this global minimum through gradient descent can be challenging; the optimization process typically converges to a local minimum that, while not the absolute lowest point, still demonstrates effective performance.

Notably, this calibration approach avoids the use of deep neural networks. Instead, it relies solely on framewise observations to optimize $\boldsymbol{e}$. Once the estimated error vector $\hat{\boldsymbol{\epsilon}}$ is determined, the unsupervised calibration process is concluded, and the previously described unes-SFI can be employed to compute the Δ-rad-rotated signal:

$$\boldsymbol{x}(\Delta + \boldsymbol{\epsilon}) = \boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}})^{-1}\boldsymbol{x}(\boldsymbol{\epsilon}). \tag{4.11}$$

A key practical advantage of this method is its adaptability to real-world conditions.

Since the non-uniform microphone distribution of an unes-CMA is typically static, the error vector remains unique to the array. Consequently, unsupervised calibration is required only once. The resulting estimated error vector $\hat{\boldsymbol{\epsilon}}$ can be reused consistently, eliminating the need for further optimization, even under changing environmental conditions.

Finally, this process enables accurate estimation of the sound field as it would have been recorded in its original, before-rotation state, without requiring the exact value of $\boldsymbol{\epsilon}$.

4.3 Modification of cost functions

As previously discussed, calculating the final cost function $\mathcal{L}(\boldsymbol{e})$ requires the computation of M individual cost functions, denoted as $\mathcal{L}_i(\boldsymbol{e})$. For each $\mathcal{L}_i(\boldsymbol{e})$, the computational complexity is $O(M^2 \cdot T \cdot F)$. Thus, the overall time complexity for computing $\mathcal{L}(\boldsymbol{e})$ is $O(M^3 \cdot T \cdot F)$.

In practical scenarios, the unsupervised calibration process, which estimates the error vector $\hat{\boldsymbol{\epsilon}}$, is a one-time operation. Once $\hat{\boldsymbol{\epsilon}}$ is derived, the iterative optimization does not need to be repeated, minimizing the impact of the computational cost associated with calculating $\mathcal{L}(\boldsymbol{e})$. This ensures that the relatively high computational complexity does not adversely affect the real-time applicability of the GSFI method. However, there remains significant potential for further reducing the computational burden of $\mathcal{L}(\boldsymbol{e})$. Accelerating the estimation of the error vector would enhance the practicality and efficiency of the GSFI method in real-world applications.

In this subsection, we focus on the efforts and advancements made to reduce the computational complexity of the cost function. Since the complexity of $\mathcal{L}(\boldsymbol{e})$ is primarily influenced by two factors—M^3 (related to the number of microphones) and

$T \cdot F$ (related to the total number of time frames and frequency bins)—we address the reduction in computational complexity from both of these perspectives.

4.3.1 Simplification of the calculation of $\mathcal{L}(\boldsymbol{e})$

The computational complexity of $\mathcal{L}(\boldsymbol{e})$, $O(M^3 \cdot T \cdot F)$, clearly highlights a major drawback: as the number of microphones M increases, the computational cost grows cubically, which becomes impractical for arrays with a large number of microphones. However, by carefully analyzing the structure of the cost functions, we can identify opportunities to reduce this complexity while maintaining the accuracy of the optimization process.

The computation of each individual cost function $\mathcal{L}_i(\boldsymbol{e})$ involves a complexity of $O(M^2 \cdot T \cdot F)$, where the $O(M^2)$ component is dominated by matrix multiplication operations. For a given microphone array, where the number of microphones M is fixed, this component cannot be further reduced. Therefore, if the need to calculate all M cost functions could be bypassed, the overall computational complexity for $\mathcal{L}(\boldsymbol{e})$ can reach a theoretical minimum value of $O(M^2 \cdot T \cdot F)$.

During the iterative minimization of the final cost function $\mathcal{L}(\boldsymbol{e})$ using the gradient descent method, each individual cost function $\mathcal{L}_i(\boldsymbol{e})$ independently converges towards its own local minimum. To reduce the computational burden, we propose selecting only one of the M cost functions as the final cost function. This can be achieved by identifying the most representative cost function based on its initial value and using it for the entire optimization process, that is,

$$\mathcal{L}(\boldsymbol{e}) = \mathcal{L}_i(\boldsymbol{e}),\ i = \underset{i \in \{1,\ldots,M\}}{\arg\min}\, \mathcal{L}_i(\mathbf{0}), \tag{4.12}$$

where $\boldsymbol{e}$ is initialized to the zero vector. Once the representative cost function is selected, it is used exclusively throughout the iterative optimization process. This approach eliminates the need to compute all M cost functions at every iteration, achieved without compromising the accuracy or integrity of the unsupervised calibration process and significantly reducing the computational complexity of the overall optimization.

4.3.2 Simplification of the calculation of $\mathcal{L}_i(\boldsymbol{e})$

In the aforementioned analysis, we highlighted that the value of M in the computational complexity of $\mathcal{L}_i(\boldsymbol{e})$ cannot be altered, and due to the matrix multiplication operation, the computational complexity of $O(M^2)$ is also inevitable. Therefore, to optimize the computational complexity of $\mathcal{L}_i(\boldsymbol{e})$, we must focus on reducing the values of T (the total number of time frames) and F (the total number of frequency bins). As highlighted, many time–frequency (TF) components in a sound signal—such as silent segments in the time domain or low-power frequency components—do not contribute significantly to the cost function computation. Thus, leveraging a strategy that eliminates these irrelevant components can substantially reduce computational requirements without compromising performance.

To achieve this reduction, a binary mask $\mathbf{B}$ is applied to the TF representation of the reference signal. This binary mask compresses the TF matrix by selecting only the most relevant components based on a predefined threshold. The binary mask $\mathbf{B}$ can be defined as:

$$\mathbf{B}(t,f) = \begin{cases} 1, & \text{if } |x_{\text{ref},t,f}| \geq \text{threshold} \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

Therefore, applying $\mathbf{B}$, components with magnitudes below the threshold are set to zero, while those above the threshold are retained. After sparsification, the calculation of the individual cost function $\mathcal{L}_i(\boldsymbol{e})$ is confined to the reduced set of TF components. This significantly decreases the dimensions of the TF matrix used in the computation. The revised cost function is given by:

$$\mathcal{L}_i(\boldsymbol{e}) = 10\log_{10}\Big(\sum_{(t,f)\in\mathbb{TF}} |\hat{x}_{\mathrm{ref},t,f} - x_{\mathrm{ref},t,f}|^2\Big), \tag{4.14}$$
$$\text{where } \forall(t,f)\in\mathbb{TF}, \quad \mathbf{B}(t,f) = 1.$$

Here, $\mathbb{TF}$ is an ordered pair set composed of all TF index pairs (t, f) retained after sparsification ($\mathbf{B}(t, f) = 1$).

By applying the binary mask $\mathbf{B}$, the number of TF components involved in the computation is reduced from $T \cdot F$ to $|\mathbb{TF}|$, where $|\mathbb{TF}| \ll T \cdot F$. Consequently, the computational complexity of $\mathcal{L}_i(\boldsymbol{e})$ (or $\mathcal{L}(\boldsymbol{e})$) becomes $O(M^2 \cdot |\mathbb{TF}|)$. This is substantially smaller than the original complexity, $O(M^2 \cdot T \cdot F)$, especially for sparse signals, resulting in faster computations for the cost function.

4.4 Simulated experimental evaluation

4.4.1 Setup

Dataset and preprocessing

We use the same methodology as described in Chapter 3 to construct the experimental dataset. From the SiSEC database [72], which offers high-quality speech recordings sampled at a precision of 16 kHz, we curated a balanced dataset consisting of eight

distinct speech signals—equally divided between female and male speakers. To emulate the complexities of a reverberant environment, these speech signals were convolved with room impulse responses (RIRs), following the same approach as in Chapter 3 and enabling us to generate microphone signals with a reverberation time of approximately 100 ms. For a thorough analysis within the TF domain, we applied the STFT, using a Blackman window of 64 ms length with a 1/8 overlap. To rigorously evaluate the efficacy of the proposed GSFI technique on unes-CMAs, we utilized an M-channel unes-CMA with a radius of 0.05 m placed within a noise-free room to record sound signals.

Simulated experimental setup for unes-CMAs

To emulate the inherent non-uniform distribution in practical scenarios, a stochastic element denoted as $\omega_i(^\circ)$, $i \in \{0, ..., M-1\}$, was artificially imposed on each microphone location. This variable ω_i adhered to a Gaussian distribution characterized by a mean of zero and a standard deviation equivalent to $\sqrt{200}^\circ$. Moreover, an additional layer of complexity was added by including an unknown error, $\epsilon_i(^\circ)$, associated with each microphone placement, rendering the actual distribution unspecified. It is pertinent to note that during the previous unes-SFI without the newly proposed unsupervised calibration procedure, we could only assume ω_i to represent the angular error because only ω_i was available and ϵ_i was unknown to us. However, the genuine angular error encompassed $\omega_i + \epsilon_i$. We hypothesized that such mismatches might impair the performance of unes-SFI, yet the proposed GSFI method potentially alleviates the detrimental impact of these unknown mismatches. The unknown mismatch error ϵ_i also conformed to a Gaussian distribution, with a mean of zero and variances systematically spanning from $(0^\circ)^2$ to $(\sqrt{500}^\circ)^2$ in discrete steps of $(\sqrt{10}^\circ)^2$. It is crucial to

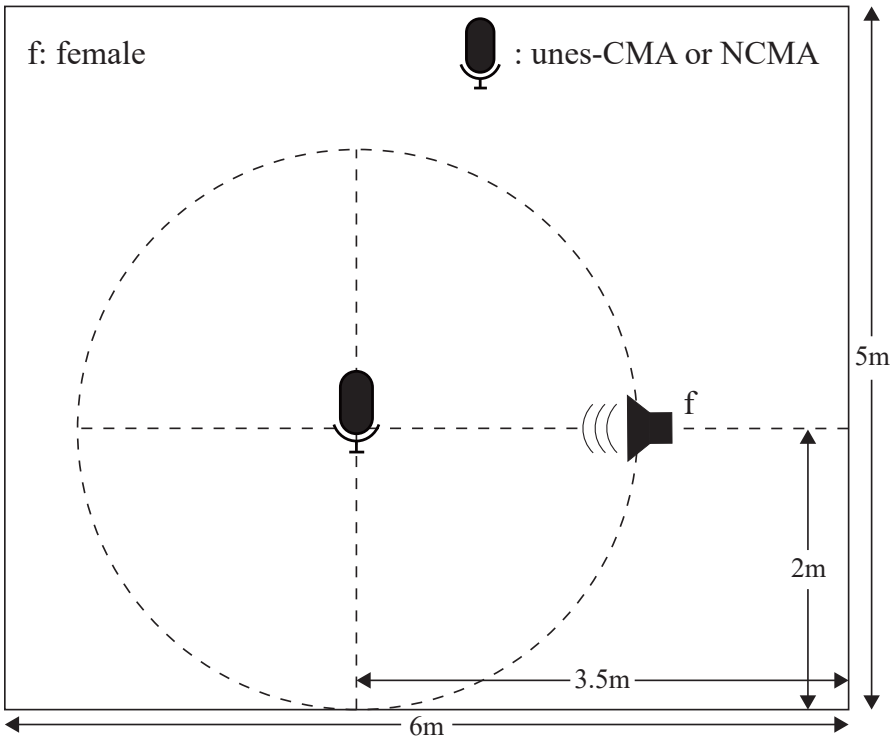

Figure 4.3: *Simulated environment during unsupervised calibration. One of the speech signals (female) is placed at the position of* 0°*, with the microphone array located at the center of the circle.*

emphasize that all introduced errors were independently and identically distributed, ensuring statistical integrity. To fortify the robustness of our analytical framework, for every specified variance within the Gaussian distribution, a comprehensive set of 100 random samples was generated, thereby guaranteeing a rigorous and reliable statistical evaluation.

As mentioned earlier, the advantage of GSFI lies in the fact that once the estimated error vector is obtained through unsupervised calibration, there is no need for re-calibration even if the acoustic environment changes, such as a change in sound source signal or the relocation of the microphone array and sound source. In our experiments,

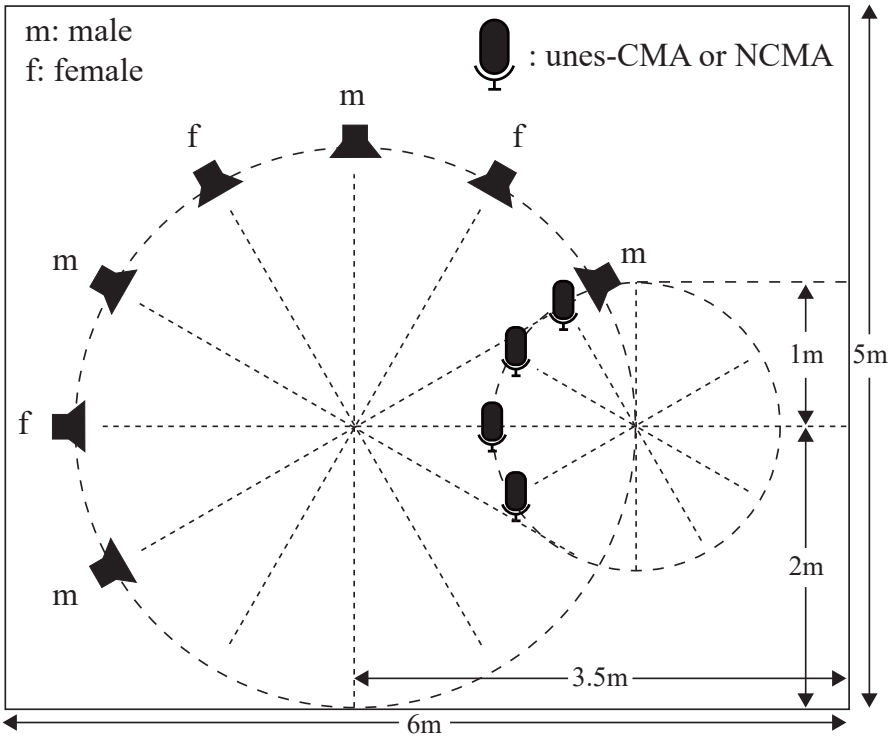

Figure 4.4: *Simulated environment for evaluation. The remaining seven speech signals are positioned at* $30°, \ldots, 210°$*, respectively. There are four optional positions of the microphone array, located at* $120°, \ldots, 210°$*. In the experiments to evaluate the performance of GSFI, one speech signal and one position of the microphone array are selected each time, resulting in 28 combinations. In the experiments to evaluate the performance of source enhancement, two speech signals are chosen and mixed as the observed signal, with the microphone array positioned at* $180°$*.*

we selected a single speech signal from the dataset for unsupervised calibration. Initially, the sound field was simulated with the microphone array positioned as shown in Figure 4.3. This setup was used to perform unsupervised calibration, thereby obtaining the estimated error vector. To evaluate the performance of the interpolation and the robustness of our method, we altered the positions of both the microphone array and the sound source after the initial calibration. Despite these changes, the interpo-

lation process utilized the error vector estimated from the initial placement depicted in Figure 4.3. The new positions of the microphone array and sound source during interpolation are illustrated in Figure 4.4. The simulation process was structured as follows:

- Initially, the sound field was simulated with the microphone array positioned, as shown in Figure 4.3, and a single speech signal was used to perform unsupervised calibration, yielding the estimated error vector.
- Subsequently, the microphone array and sound source were relocated to one of the new positions, as illustrated in Figure 4.4.
- Finally, at these new positions, a new sound field was simulated after the microphone array rotated by Δ rads. This newly simulated sound field was then used to estimate the observation signals as if they were captured at the original reference position before rotation, with the rotational angle $\phi = \Delta\pi/180^\circ$ being a known value.

Evaluation criteria

In our initial experiments to evaluate the performance of GSFI on unes-CMAs, the experiments were conducted in a controlled scenario involving a single sound source, ensuring that no sound sources were mixed. The performance assessment was based on the same metric as introduced in Chapter 3, signal-to-error ratio (SER). We varied the number of microphones, M, from 4 to 7 to examine the impact of array size on the performance of GSFI. Additionally, we manipulated the rotation angle ϕ to simulate different array orientations and assess the robustness of the method under varying conditions.

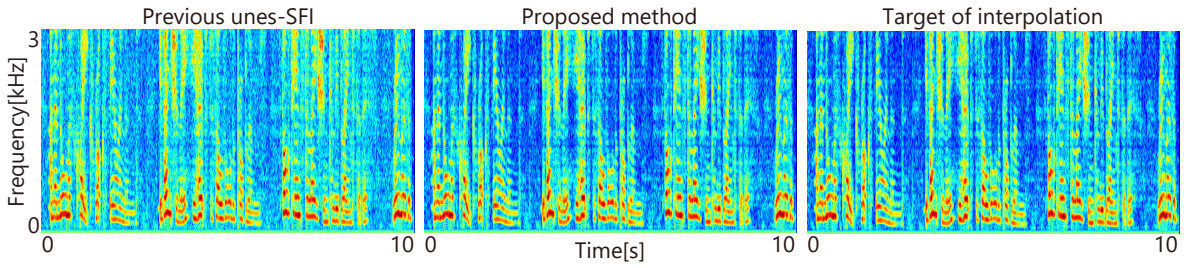

Figure 4.5: *Spectrograms of signals interpolated using the previous unes-SFI method and the newly proposed GSFI method, alongside the spectrogram of the target signal expected from interpolation.*

In the second set of experiments, we aimed to compare the source enhancement performance characteristics of various methods using the Minimum Power Distortionless Response (MPDR) beamformer. The evaluation metrics remained consistent with Chapter 3 and were based on the source-to-distortion ratio (SDR) and the source-to-interference ratio (SIR). To calculate the filter of the beamformer, we adopted the same approach as described in Chapter 3, relying on two key components: the covariance matrix of the interference signal and the relative transfer function (RTF), which was derived using the RIR from the target source to each microphone. For these experiments, we selected two sound sources randomly from Figure 4.4 and mixed them into the observation signal. The angular separation between the two sources was varied systematically at intervals of $30°, 60°, \ldots, 180°$.

4.4.2 Effectiveness of the proposed method

Initially, we focus on the scenario illustrated in Fig. 4.3. Fig. 4.5 presents an example of direct acoustic results obtained through interpolation, including spectrograms of signals interpolated using the previous unes-SFI method and the newly proposed GSFI method, alongside the spectrogram of the target signal expected from interpolation.

In this case, the number of microphones is $M = 7$, and the rotation angle is 30°.

To better highlight the differences among the spectrograms, we primarily examine interpolation results within the 0–3 kHz frequency range. As shown in Fig. 4.5, the interpolated signal obtained using our proposed GSFI method with unsupervised calibration closely matches the target signal, exhibiting minimal differences. In contrast, the interpolation results from the previous unes-SFI method without calibration exhibit noticeable discrepancies in the spectrogram compared to the target signal, particularly at high frequencies. Additionally, the previous interpolation method introduces error noise into the interpolated signal. These direct acoustic results clearly demonstrate the significant advantages of our GSFI method in accurately reconstructing the before-rotation signal, further validating its effectiveness over the previous method.

4.4.3 Comparison between the cost functions (4.9) and (4.12)

In this experiment, we evaluated the impact of the two cost functions introduced in Sections 4.2 and Section 4.3 on the performance of GSFI. In this comparison, we aimed to determine how these cost functions affect the efficiency and effectiveness of unsupervised calibration. For this experiment, we focused on a scenario where the unes-CMA and the sound source are positioned as depicted in Figure 4.3. Figure 4.6 presents some examples of the variations in the values of the two cost functions during the iterative optimization of unsupervised calibration with $M = 5$ and 6. Both cost functions exhibit similar rates of change, demonstrating that their convergence behavior is nearly identical. After an equivalent number of iterations, both cost functions reach a local minimum. However, a key distinction lies in the computational complexity associated with each cost function. Calculating the cost functions (4.12) and (4.14) is considerably less computationally complex than calculating the cost functions (4.8)

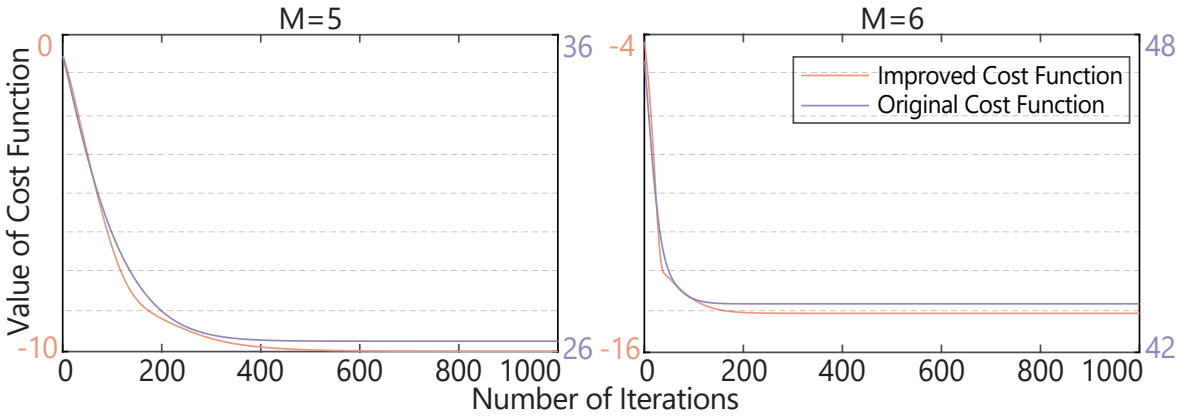

Figure 4.6: *Variation in the values of the two cost functions during the iterative optimization process of unsupervised calibration.*

and (4.9). This reduction in complexity translates to a significant improvement in the execution speed of unsupervised calibration when the improved cost function is employed. It is important to note, as illustrated in Figure 4.6, that there are differences in the cost function values and local minima between cost functions (4.12) and (4.9). These discrepancies arise because, in cost function (4.12), we simplified the formulation by avoiding the aggregation of multiple cost functions and reducing the dimensionality of the matrix involved in the computation. Consequently, cost function (4.12) yields a lower final value than cost function (4.9). However, a lower value of cost function (4.12) than of (4.9) does not necessarily indicate superior optimization performance. The reduction in cost function value is a result of simplifications made during calculation, rather than an improvement in the iterative optimization results.

Building upon our previous examination of cost functions, we now turn to their impact on interpolation performance. We conducted this experiment by varying the rotation angle ϕ from 10° to 40° and adjusting the number of microphones M from 5 to 6. For each configuration, we used a standard deviation of 10° for the unknown mismatch ϵ_i. Figure 4.7 displays the mean SER across all M channels, providing a

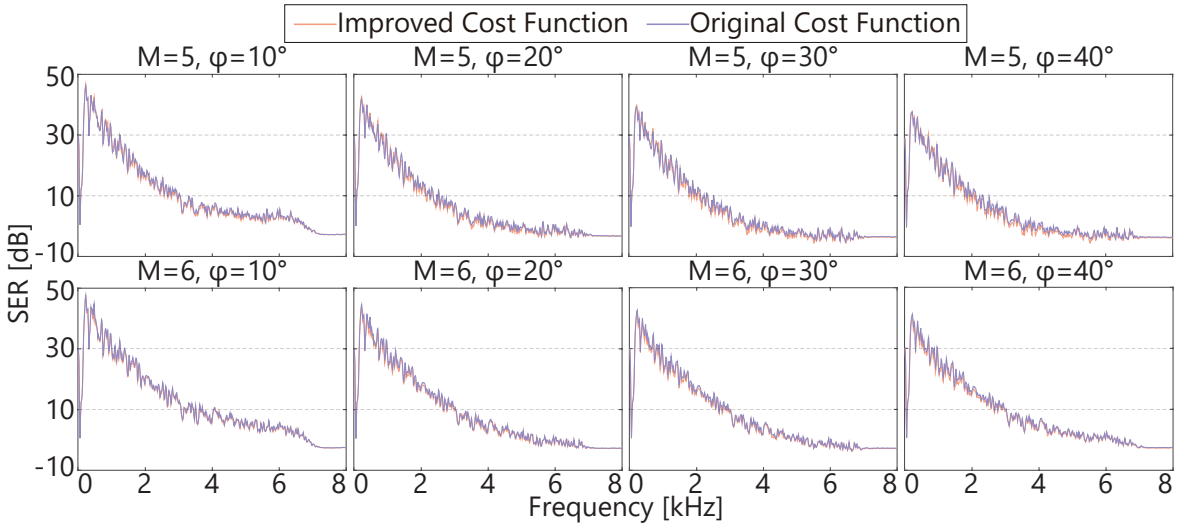

Figure 4.7: *SER results for two cost functions with different M and ϕ values.*

comprehensive assessment of interpolation accuracy. The results demonstrate that the interpolation performance is nearly identical for both cost functions across the specified ranges of rotation angles and microphone counts. Despite the difference in computational complexity, the SER outcomes remain consistent, indicating that the simplified cost functions ((4.12) and (4.14)) do not compromise accuracy. Given the marked improvement in computational speed and the negligible performance difference, it becomes clear that the improved cost function is preferable for practical applications. Therefore, for all subsequent experiments, we adopted the improved cost function. Notably, higher-frequency components pose a challenge. To streamline the analysis and focus on the most critical performance aspects, we limited the frequency range to 0–1 kHz for SER evaluation. The SERs were averaged in decibels across all subsequent experiments to provide a clear and concise comparison.

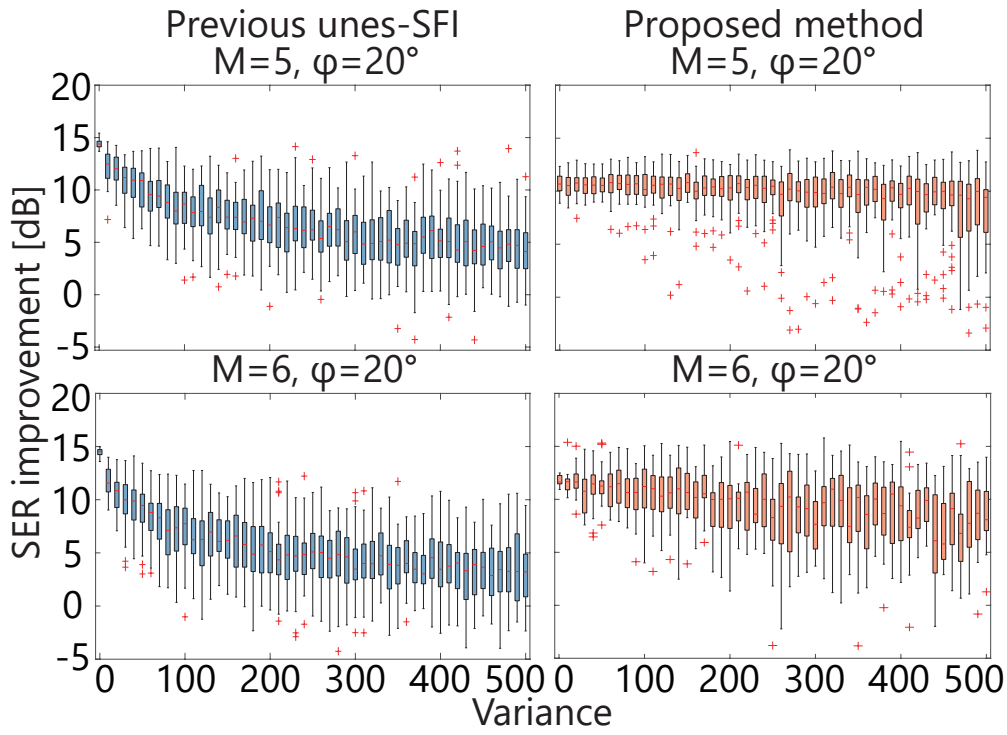

Figure 4.8: *Boxplots of the relationship between the variance of the unknown mismatch ϵ_i and the SER improvement at frequencies up to 1 kHz relative to the cases without interpolation.*

4.4.4 Robustness to the variance of angle error

To assess the robustness of our proposed method, we investigated the relationship between the variance of the unknown mismatch ϵ_i and the SER improvement. The SER improvement was used to measure the enhancement achieved through signal processing, with the baseline SER calculated by comparing the uninterpolated signal after rotation with the target signal before rotation. In this analysis, we focused on scenarios with five and six microphones ($M = 5$ and $M = 6$) and a rotation angle (ϕ) of 20°, with the unes-CMA and sound source positioned as depicted in Figure 4.3.

Figure 4.8 illustrates the relationship between the variance of the unknown mismatch ϵ_i and the SER improvement. Each box plot in Figure 4.8 represents the mean

SER improvement over M channels for each sample, totaling 100 data points per box. The results indicate that as the variance of the unknown mismatches increases, the SER improvement provided by the previous unes-SFI method significantly deteriorates from approximately 13 dB to as low as 5 dB. Conversely, our proposed GSFI method consistently demonstrates superior interpolation performance, achieving a stable SER improvement of 13 dB, even with large unknown mismatches. The findings underscore the limitations of applying the previous unes-SFI method directly to an unes-CMA with an unknown microphone distribution. In contrast, our proposed GSFI method exhibits robust performance under these challenging conditions, highlighting its practical advantages and effectiveness in handling substantial unknown mismatches.

4.4.5 Channelwise SER improvements

We analyzed the channelwise SER improvements across various scenarios involving different M and ϕ values, with the standard deviation of the mismatch set to 10°. The positions of the sound sources and unes-CMAs were varied, as depicted in Figure 4.4. Figure 4.9 presents the mean SER improvement relative to cases without interpolation calculated over M channels. Each box plot contains 28 samples, representing the mean SER improvement of seven sound sources at four distinct unes-CMA locations.

The results clearly indicate that our proposed method consistently delivers greater SER improvements than the previous unes-SFI method in all evaluated scenarios. Specifically, the minimum improvement occurs at $M = 4$ with an increase of 3 dB, while the maximum improvement is observed at $M = 7$, reaching approximately 15 dB. As expected, increasing M in the proposed method enhances performance owing to the higher spatial sampling rate. However, in the previous unes-SFI method, adding more microphones does not always lead to improved SER and can even degrade performance.

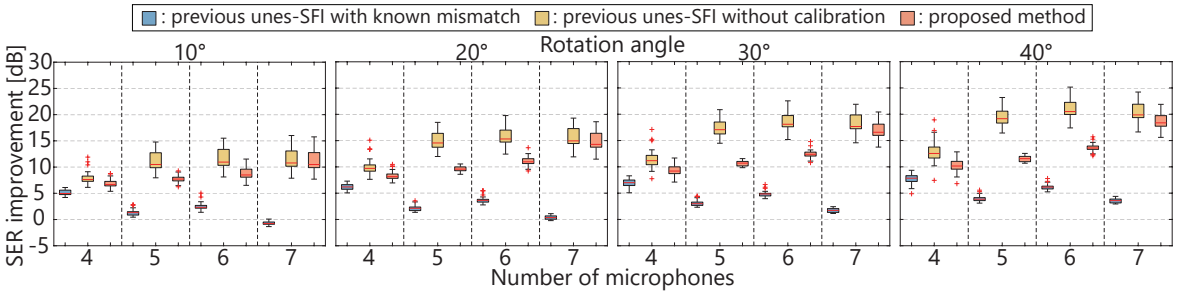

Figure 4.9: *Boxplots of mean SER improvement at frequencies up to 1 kHz across various scenarios with different M and ϕ values.*

This degradation is observed when M is increased from 4 to 5 and from 6 to 7, likely due to the introduction of additional errors in the compensation matrix $\boldsymbol{U}_M(\boldsymbol{\epsilon})$. In such cases, the advantages of a higher spatial sampling rate are negated by the detrimental effects of these errors. We also assessed the performance of the unes-SFI method under conditions where the mismatch ϵ_i is known. Although the proposed method exhibited a slightly smaller improvement in such scenarios, the performance reduction was minimal and within acceptable limits, with decreases as small as 1 dB in some cases, underscoring the robustness of our method even in the absence of prior information.

4.4.6 Effect of the signal duration on the unsupervised calibration

In practical applications, although unsupervised calibration is only performed once and does not significantly hinder real-time processing, there remains a need to expedite the estimation of the error vector. Additionally, online beamforming typically relies on very short signal segments to ensure timely responses. Consequently, it is desirable for unsupervised calibration to accurately estimate the error vector even when using these shorter segments. In our previous experiments, a 10 s sound signal was employed

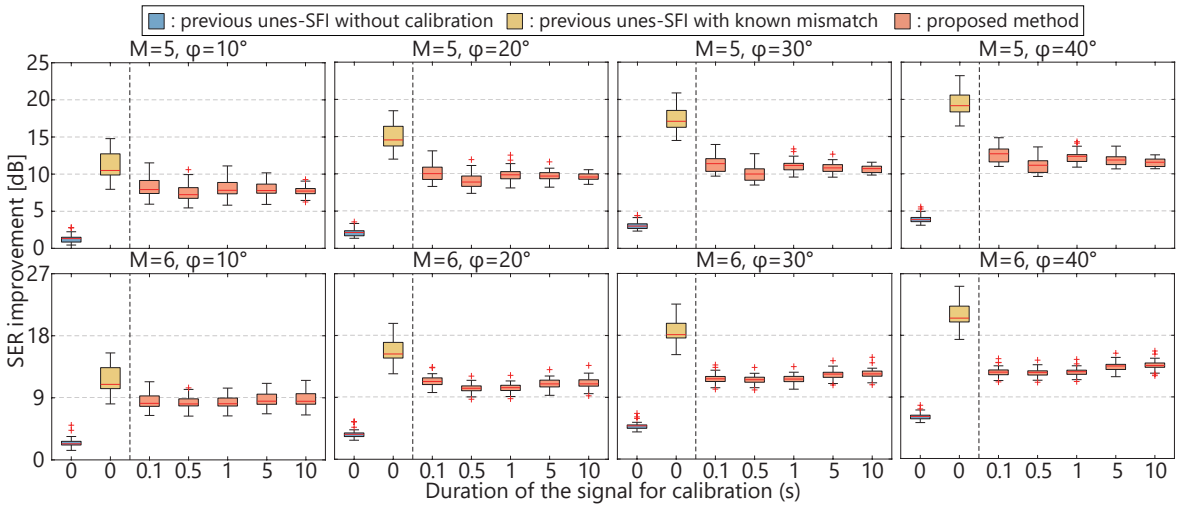

Figure 4.10: *Relationship between the duration of the sound signal used for calibration and the channelwise SER improvement across various scenarios with different M and ϕ values.*

for unsupervised calibration. Here, we explored the effect of signal segment length on the accuracy and efficiency of unsupervised calibration using sound signals of varying lengths: 10 s, 5 s, 1 s, 500 ms, and 100 ms. The estimated error vectors obtained from these calibrations were then used for the interpolation and the estimation of the signal before rotation. Figure 4.10 presents the relationship between the duration of the sound signal used for calibration and the channelwise SER improvement for scenarios with $M = 5$ and 6 and ϕ ranging from 10° to 40°, with the unes-CMA and sound source positioned as illustrated in Figure 4.4. In the previous unes-SFI, there is no calibration process; therefore, the duration of the sound signal used for calibration is 0 s. The results indicate that the duration of the signal has negligible impact on the performance of unsupervised calibration and subsequent interpolation. Signal segments of varying lengths yielded fluctuations in SER improvement of less than 1 dB. Using shorter signals does not lead to significantly worse results, demonstrating the feasibility of the proposed method for real-time processing.

Table 4.1: *Abbreviations for spectrograms from different evaluation settings*

Spectrograms from different evaluation settings	Status of evaluation conditions				
	B	R	I	K	C
B0: No beamforming	0	0	0	0	0
R0: No rotation	1	0	0	0	0
I0: No interpolation	1	1	0	0	0
C0: unes-SFI without calibration	1	1	1	0	0
K1: unes-SFI with known mismatches	1	1	1	1	0
C1: proposed method with calibration	1	1	1	0	1

4.4.7 Results of source enhancement with batch processing

In this experiment, we assessed the source enhancement capabilities using the MPDR beamformer under various conditions. The experiment was conducted with a fixed number of microphones, $M = 6$, whereas the rotation angle ϕ was varied at 10°, 20°, 30°, and 40°. Initially, the filter weight $\boldsymbol{w}$ for the MPDR beamformer was calculated using the RTF and the multichannel STFT spectrogram derived from the unes-CMA situated at its original, unrotated position. This scenario, denoted as R0, serves as the benchmark, as it employs true, uninterpolated signals for the beamformer. For clarity and consistency, we adopt a set of abbreviations to denote the evaluation conditions when naming spectrograms from different methods, similar to the naming convention used in Chapter 3: B represents Beamforming, R for Rotation, I for Interpolation, K for Known Mismatches, C for Calibration, Index 0 for off, and Index 1 for On.

Therefore, the various spectrograms from different evaluation settings, which were then processed using the precomputed MPDR beamformer weight $\boldsymbol{w}$ to generate the estimated target signal, are summarized in the Table 4.1. For a baseline comparison, we

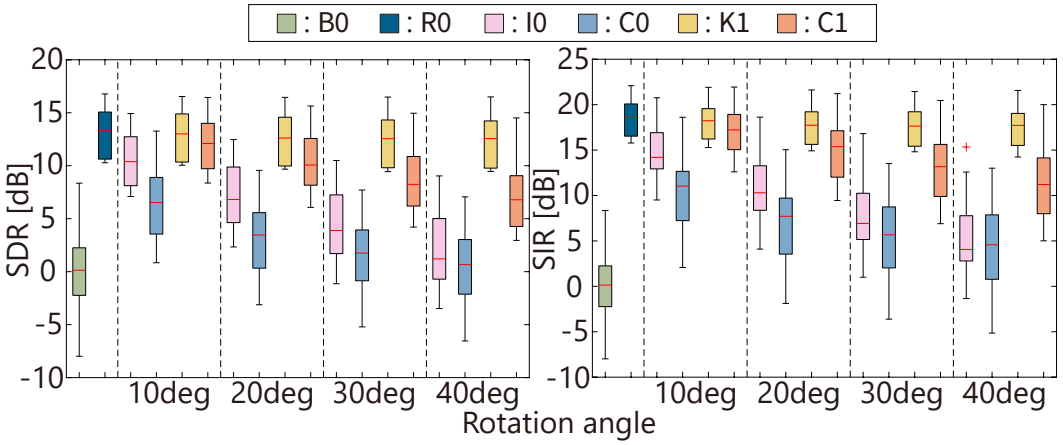

Figure 4.11: *Boxplots of SDR and SIR obtained by MPDR beamformer in six situations: unprocessed* (B0)*, no rotation of the CMA* (R0)*, without interpolation when the CMA rotates* (I0)*, with previous unes-SFI without calibration when the CMA rotates* (C0)*, previous unes-SFI with known mismatches when the CMA rotates* (K1)*, and our proposed method when CMA rotates* (C1)*.*

included an unprocessed scenario (B0), where the raw microphone signal was incorrectly treated as the target signal, alongside R0.

The SDR and SIR outcomes for various scenarios, with the mismatch standard deviation set to 10°, are depicted in Figure 4.11. Predictably, B0 demonstrates the lowest SDR and SIR (0 dB), whereas R0 achieves the highest source enhancement, attributed to the time-invariant nature of the ATS. The proposed method (C1) generally outperforms both the non-interpolated case (I0) and the previous interpolation technique (C0) by approximately 4 dB and 7 dB, respectively, suggesting its capability to mitigate the performance degradation associated with unknown microphone distributions. The performance difference between K1 and C1 remains marginal and acceptable, with the smallest difference being only 1 dB. These findings indicate that our proposed method with unsupervised calibration maintains robustness to unes-CMA rotations and enhances array signal processing performance, even when the precise microphone

distribution is unknown.

4.4.8 Results of source enhancement with online processing

In this experiment, we examined the application of the GSFI technique to online beamforming, focusing on scenarios where the ATS was subjected to continuous, dynamic changes. As described in Section 2.3.3, we also used a smoothing factor α to facilitate the online updating of the spatial covariance matrix (SCM) and the Sherman–Morrison formula to reduce the computational complexity involved in inverting the covariance matrix within the MPDR beamforming framework.

The algorithm and experimental conditions used in this study mirrored those from Section 3.4.4. We used two source signals, each with a duration of 40 s, positioned at angles of 30° and 210°, and one microphone array positioned at an angle of 180°, as depicted in Figure 4.4. The frame length was set to 256 ms, and a segmental SDR at intervals above 1 s was employed to evaluate source enhancement performance. Unsupervised calibration was performed using only the first frame (256 ms) of the sound signal. The smoothing factor α was empirically set to 0.99, which yielded the highest segmental SDR. The inversion of the covariance matrix $\hat{\boldsymbol{V}}_f^{-1}$ was initialized using the first 10 frames. During the experiment, the unes-CMA underwent two rotations: the first rotation from 0° to 20° began at 10 s, and the second rotation from 20° to 40° started at 30 s. These rotations were gradual, occurring at a constant speed of 0.01° per time sample (equivalent to 160° per second), simulating typical human or humanoid robot rotation speeds. Observations were generated by concatenating the observations of the unes-CMA at various angles.

Figure 4.12 presents the segmental SDRs for $M = 6$. As expected, the R0 scenario, where the unes-CMA did not rotate, consistently achieved the highest source enhance-

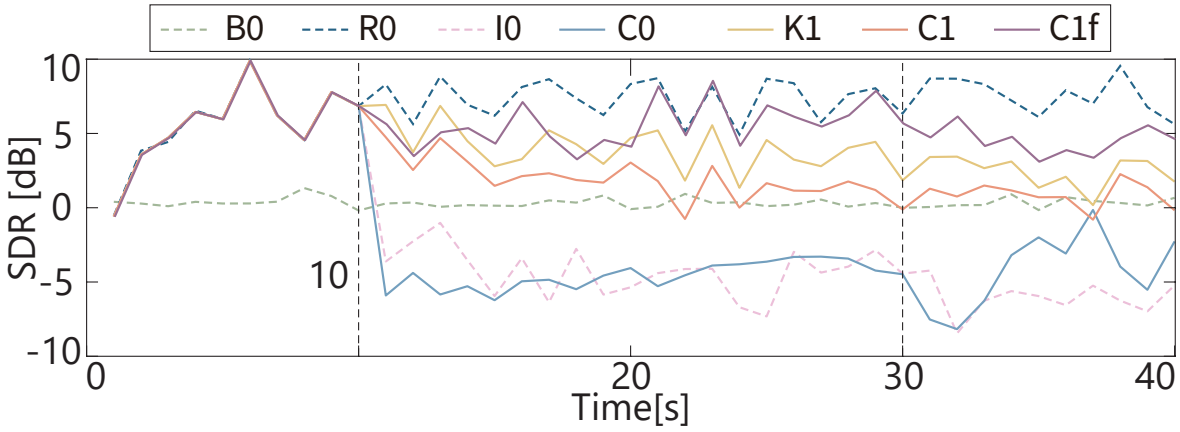

Figure 4.12: *Segmental SDR every* 1 s *with* $M = 6$ *on an unes-CMA, where the two vertical dashed lines indicate the time points when the rotation started:* $0° \Rightarrow 20° \Rightarrow 40°$. B0 *shows the mixture itself,* R0 *shows the case where no rotation occurs,* I0 *shows online processing without interpolation,* C0 *and* K1 *respectively show online processing using previous unes-SFI without calibration and with known mismatches,* C1 *shows online processing with the newly proposed method, and* C1f *shows online processing with the newly proposed method and the frozen SCM.*

ment performance. Contrary to initial expectations, the B0 method does not yield the lowest SDRs. The C0 method exhibited the poorest performance at about −5 dB, often worse than the I0 method. These findings indicate that the previous SFI technique is ineffective for an unes-CMA with an unknown microphone distribution in online beamforming scenarios. In contrast, our proposed method (C1), which achieved an average improvement of 8 dB in SDR score, demonstrated an improved performance compared with C0 and I0. Despite some instances where the SDRs of our proposed method (C1) occasionally approach those of the B0 scenario, the overall trend indicates that our method can still provide an improvement to some extent.

During online beamforming, we applied the signal processed by GSFI to update the SCM online. However, estimation errors in the GSFI-estimated signal led to inaccuracies in the SCM, which likely contributed to the inaccurate calculation of the

filter of the beamformer and the observed performance being similar to B0. We additionally introduced a method, where the SCM was iteratively updated only before the unes-CMA rotation. Once the unes-CMA rotated, further updates to the SCM were halted. Subsequently, the update process for the filter of the beamformer was also ceased owing to the frozen SCM (fSCM). The results obtained from this method, denoted as C1f, revealed a significant improvement of 4 dB in SDR compared to C1. By avoiding the introduction of inaccurate SCM calculations, our proposed method no longer shows performance similar to B0. This underscores the capability of the proposed GSFI method with unsupervised calibration for online beamforming.

4.5 Conclusion

In this chapter, we introduced the GSFI method to achieve rotation-robust beamforming for unes-CMAs with unknown microphone distributions. At the core of GSFI is an innovative unsupervised calibration framework, which estimates the positional errors of microphones without requiring any prior knowledge of their locations. This calibration process determines the microphone distribution in an unsupervised manner. Once the positional errors are estimated, GSFI incorporates the previously established unes-SFI method to reconstruct the target signal of the unes-CMA in its before-rotation state.

Simulation results validated the robustness of GSFI against unknown microphone distributions and demonstrated its strong performance in array signal processing tasks. First, we showed that the proposed method enables accurate signal estimation even in the absence of prior knowledge about microphone distributions. Furthermore, we demonstrated that the improved cost function achieves nearly undiminished interpolation performance while significantly reducing computational complexity.

In addition, GSFI maintains stable signal estimation performance compared to previous methods across various experimental conditions. The duration of the signal used for unsupervised calibration was also found to have minimal impact on both the calibration process and subsequent interpolation accuracy. This finding highlights the feasibility of applying the proposed method to real-time scenarios, as shorter signal durations do not substantially degrade performance.

Finally, the proposed method consistently improved performance in downstream source enhancement task, further underscoring its practical utility.

Nevertheless, the current method is limited to circular arrays. Extending GSFI to nearly circular microphone arrays (NCMAs), which approximate the shape of the robot head rather than a perfect circle, presents an intriguing and practical research direction. This extension will be explored further in the next chapter.

5 Two-stage Generalized Sound Field Interpolation on a Nearly Circular Microphone Array

In this chapter, we address a more complex scenario involving head-mounted microphone arrays. Given the need to conform to the shape of the robot head, and the fact that the head is not a perfect sphere, maintaining a standard circular shape for head-mounted circular microphone arrays (CMAs) becomes challenging. As a result, nearly-circular microphone arrays (NCMAs) are more commonly used in the application scenarios discussed in this thesis.

Building on the Generalized Sound Field Interpolation (GSFI) method introduced in Chapter 4, we extend the approach to accommodate NCMAs. In this chapter, we explore how GSFI can be applied to NCMAs and present simulation experiments that validate the feasibility of interpolation in these non-ideal array configurations.

5.1 Overview

Up to this point, we have focused exclusively on CMAs. In techniques such as sound field interpolation (SFI), unequally spaced sound field interpolation (unes-SFI), and the GSFI method for an unes-CMA mentioned in the previous chapters, the microphone

array is assumed to be circular. However, in real-world applications, since the robot head is not a perfect sphere, maintaining a strictly circular array when worn on the head is challenging. For an NCMA, previous SFI and unes-SFI techniques cannot be directly applied as their effectiveness will be severely impacted, and it is also theoretically inappropriate to directly use unsupervised calibration because the NCMA does not exhibit the periodicity and other properties of a CMA.

Given the prevalence of NCMAs over strictly circular ones, we propose a new framework to address the challenges associated with their non-ideal geometry. This framework is built upon two main ideas:

- **Simplify the problem on the NCMA**: To simplify the signal processing challenges associated with an NCMA and leverage the methods developed for CMAs in previous chapters, we first construct a virtual pseudo-CMA (pCMA). This pCMA is designed to approximate the NCMA by ensuring that the signals it receives closely resemble those captured by the NCMA. By doing so, the pCMA serves as a functional substitute for the NCMA, allowing us to transform the problem of interpolation on an NCMA into the more familiar CMA framework.

- **Two-stage approach between pseudo-CMAs**: When the microphone array rotates, the correspondence between the pCMA and the NCMA changes. Specifically, the pCMA constructed to replace the before-rotation NCMA differs from the one required to represent the after-rotation NCMA. Consequently, constructing a single pCMA is insufficient to accurately estimate the signals of the NCMA in its before-rotation state. To address this issue, we adopt a two-stage approach: the states of the NCMA before and after rotation are separately mapped to two distinct pCMAs. These pCMAs are then used collaboratively to estimate the signals of the NCMA in its original, before-rotation position.

These two ideas ensure that the challenges posed by the NCMA's non-circular geometry and rotational variations are effectively managed, thereby extending the applicability of the GSFI method to nearly-circular arrays.

5.2 Simplification of the NCMA problem to a CMA problem

In this section, we describe how the NCMA problem can be simplified into a CMA problem. This is achieved by constructing a pCMA based on the NCMA and determining the optimal microphone distribution for the pCMA. The goal is to ensure that the signals received by the pCMA closely approximate those received by the NCMA, thereby enabling the pCMA to serve as an effective substitute for the NCMA. This transformation allows us to apply the existing CMA-based methods to address challenges associated with NCMAs.

5.2.1 How to construct a pseudo-CMA

As previously mentioned, for a unes-CMA, the microphone positions are initially unknown. However, since all microphones are constrained to lie on the same circle, unsupervised calibration can be employed to estimate the angular error of each microphone, thereby determining their precise positions on the circle.

In contrast, an NCMA introduces a more complex challenge. The microphones in an NCMA can be considered as being distributed across multiple concentric circles, each with the same center but different radii. The rotational movement in this context pertains solely to rotation around the common center of these concentric circles, which

we designate as the center of the NCMA. This nearly circular distribution disrupts the periodicity inherent in circular arrays, making it difficult to directly estimate accurate angular errors through unsupervised calibration.

To address this issue, we construct a pCMA for substituting the NCMA according to the following ideas:

- **Transforming the NCMA into a pCMA**: The microphones on the NCMA are hypothetically moved radially such that they are positioned on a common circle, forming a pCMA. This transformation standardizes the radial distances of the microphones to enable the calibration process.

- **Signal consistency in the transformation**: Due to this radial transformation, the signals received by the microphones on the pCMA evidently differ from those received on the original NCMA. Thus, the challenge lies in determining the microphone distribution on the pCMA such that the signals they receive are approximately equivalent to those received by the NCMA.

- **Determination of the distribution of microphones on the pCMA**: This step involves optimizing the angular positions on the pCMA to minimize the discrepancy between the signals observed on the NCMA and those hypothesized for the pCMA. A calibration process can be adapted to estimate the hypothetical microphone positions on the pCMA that would yield signals closely resembling those captured by the NCMA.

5.2.2 How to determine the distribution of microphones on the pCMA

To determine a distribution of microphones on the pCMA such that the signal it receives matches that of the original NCMA, we adopt the unsupervised calibration for the pCMA.

The goal is to simplify the complex challenge of processing an NCMA into the more tractable task of processing a CMA. Since the signals received on the pCMA are unknown, we cannot directly minimize the discrepancy between the signals of the NCMA and the pCMA in a supervised manner. Instead, we hypothesize that the pCMA already receives the same signal as the NCMA, denoted by $\boldsymbol{x}_{\text{NCMA}}$. This assumption allows us to apply unsupervised calibration to estimate the distribution of microphones on the pCMA, similar to the approach described in Chapter 4.

Although applying unsupervised calibration directly to the NCMA is unreasonable, we rationalize the use of unsupervised calibration on $\boldsymbol{x}_{\text{NCMA}}$ by indirectly applying it through a pCMA, which virtually aligns all microphones of the NCMA onto a common circle.

5.2.3 Details of pCMA construction

During unsupervised calibration, a single microphone on the pCMA is chosen as the reference microphone. The signal for this reference microphone is estimated using signals from the other microphones. Since we assume the signal received by the pCMA matches $\boldsymbol{x}_{\text{NCMA}}$, this process is equivalent to estimating the reference signal on the NCMA using signals from microphones positioned on different concentric circles. Therefore, the placement of microphones on the pCMA follows a logical and consistent

approach:

- The circle on which the pCMA resides coincides with that of the reference microphone, implying that the reference microphone is retained in its original position.
- The remaining microphones are radially moved to align with the reference microphone's circle, effectively collapsing the NCMA into a single circular array.

This transformation ensures that the pCMA captures the original spatial characteristics of the NCMA while simplifying the structure for subsequent processing.

5.2.4 Details of unsupervised calibration on the pCMA

During optimization, the reference channel is consistently chosen as the microphone channel corresponding to the smallest initial cost function value, as determined by (4.12). This ensures that:

- The pCMA's defining circle remains fixed throughout the iterative optimization process.
- The iterative updates during unsupervised calibration are applied relative to a stable spatial framework.

5.2.5 Summary

Figure 5.1 illustrates the detailed process for constructing the pCMA and determining the microphone distribution:

1. Assume the pCMA receives the same signal as the NCMA ($\boldsymbol{x}_{\mathrm{NCMA}}$).

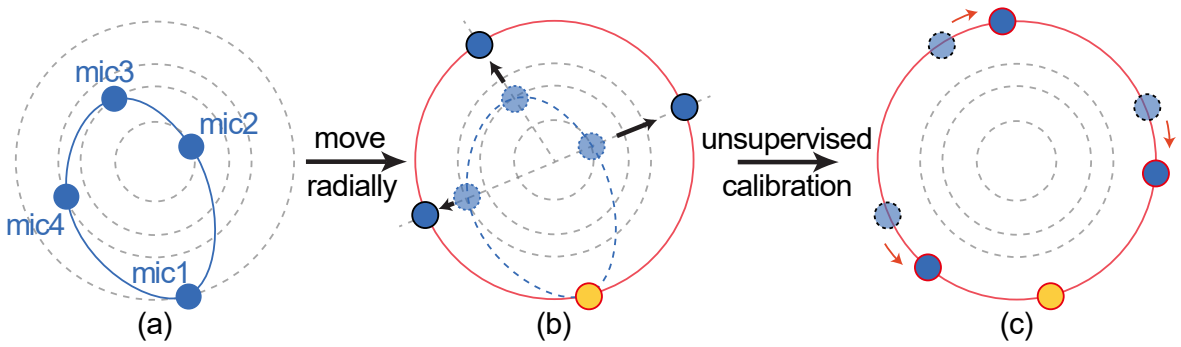

Figure 5.1: *Conceptual diagram of how to simplify an NCMA to a CMA. (a) NCMA on which each microphone can be considered positioned on different concentric circles. (b) If taking the first channel as the reference channel, the pCMA can be conceptually obtained by radially moving the microphones from other concentric circles to the circle of the reference channel. (c) After unsupervised calibration, the distribution of microphones on the pCMA that would result if the signals received by the NCMA were captured by the pCMA is determined.*

2. Taking one channel as the reference channel, consider that the pCMA is conceptually obtained by radially moving the microphones from other concentric circles to the circle of the reference channel.

3. Apply unsupervised calibration using $\boldsymbol{x}_{\mathrm{NCMA}}$ on the pCMA to estimate the microphone positions, iteratively optimizing the cost function while maintaining consistency in the reference channel.

This approach effectively reduces the complexity of processing an NCMA by leveraging the spatial simplicity of a pCMA while preserving the integrity of the original signals. Through this approach, we effectively address the challenges posed by the nearly circular distribution of microphones in an NCMA. By transforming the NCMA into a pCMA and leveraging unsupervised calibration, we can determine a suitable configuration that preserves signal consistency, enabling further processing such as SFI and beamforming.

5.3 Generalized sound field interpolation on an NCMA

While the previous section demonstrated how to simplify the NCMA problem into a CMA problem using a pCMA, relying solely on a single pCMA introduces several limitations that hinder accurate signal reconstruction on the NCMA. In this section, we will first explain the limitations of using only one pCMA, thereby highlighting the necessity of going beyond the single pCMA approach and employing a two-stage method to perform interpolation on the NCMA. Subsequently, we will provide a detailed description of the proposed two-stage method. This method ensures both reliability and computational efficiency by leveraging precomputed results and maintaining geometric consistency across stages.

5.3.1 Limitations of using only one pCMA

To achieve our goal of avoiding re-estimating the spatial filter after the microphone array has rotated, we need to develop a method that estimates the NCMA signal before rotation using the NCMA signal after rotation. The received signal of the M-channel NCMA after rotating by $-\Delta$ radians is denoted as

$$\boldsymbol{x}_{\mathrm{NCMA,a}} = \begin{bmatrix} x_{1,a} & \cdots & x_{M,a} \end{bmatrix}^{\mathsf{T}}. \tag{5.1}$$

Our task is to estimate what the sound signal, currently received by the NCMA ($\boldsymbol{x}_{\mathrm{NCMA,a}}$) would have been if it were observed at the original position before rotating by $-\Delta$ radians.

After the NCMA undergoes a rotational transformation, the corresponding pCMA can be derived using the previously outlined method, resulting in what we refer to as pCMA-a, where “a” signifies “after rotation”. To represent the non-uniform dis-

tribution of microphones on pCMA-a, we introduce the known error vector $\boldsymbol{e}_{\mathrm{a}}$, which is initially set to zero. Consequently, the angular positions of the microphones on pCMA-a can be expressed as

$$\boldsymbol{P}(\boldsymbol{e}_{\mathrm{a}}) = \left[0 - \Delta + e_{\mathrm{a},1} \quad \cdots \quad \tfrac{2\pi(M-1)}{M} - \Delta + e_{\mathrm{a},M}\right]^{\mathsf{T}}. \tag{5.2}$$

Using the observed signals $\boldsymbol{x}_{\mathrm{NCMA,a}}$ and the microphone distribution $\boldsymbol{P}(\boldsymbol{e}_{\mathrm{a}})$, unsupervised calibration is applied to iteratively optimize the error vector. Upon convergence, the estimated value of the true error vector of pCMA-a, denoted as $\hat{\boldsymbol{\epsilon}}_{\mathrm{a}}$, is obtained. This implies that when the microphones are positioned on pCMA-a according to the estimated error vector $\hat{\boldsymbol{\epsilon}}_{\mathrm{a}}$, the signals received by pCMA-a will closely approximate those received by the NCMA after rotation.

To reconstruct the NCMA signal before rotation, represented by the Δ-rad-rotated result of $\boldsymbol{x}_{\mathrm{NCMA,a}}$, a seemingly straightforward approach might involve rotating the corresponding pCMA-a by Δ rads and assuming that this rotated pCMA-a aligns with the Δ-rad-rotated NCMA. Using this assumption, the unes-SFI method would be applied to pCMA-a to estimate the Δ-rad-rotated signal, denoted as $\boldsymbol{x}(\Delta)_{\mathrm{pCMA-a}}$, according to the following equation:

$$\boldsymbol{x}(\Delta)_{\mathrm{pCMA-a}} = \boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}}_{\mathrm{a}})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}}_{\mathrm{a}})^{-1}\boldsymbol{x}_{\mathrm{NCMA,a}}. \tag{5.3}$$

This estimated signal would then be considered the before-rotation NCMA signal. However, this approach is fundamentally flawed due to a critical distinction: the microphones of pCMA-a and NCMA are positioned along different circular paths. Even for corresponding microphones—those whose received signals are approximately equivalent—the signal similarity does not persist after both pCMA-a and NCMA are rotated by Δ rads.

To clarify this issue, consider an extreme yet intuitive example involving two M-channel unes-CMAs with the same center but different radii. Although the microphones on these unes-CMAs are distributed differently, we assume they initially receive the same signal, denoted by $\boldsymbol{z}$. Given their respective known error vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ ($\boldsymbol{\alpha} \neq \boldsymbol{\beta}$), unes-SFI can be applied to each unes-CMA to compute the Δ-rad-rotated signal as follows:

$$\begin{aligned} \boldsymbol{z}(\Delta)_{\boldsymbol{\alpha}} &= \boldsymbol{U}_M(\boldsymbol{\alpha})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\boldsymbol{\alpha})^{-1}\boldsymbol{z} \\ \boldsymbol{z}(\Delta)_{\boldsymbol{\beta}} &= \boldsymbol{U}_M(\boldsymbol{\beta})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\boldsymbol{\beta})^{-1}\boldsymbol{z}. \end{aligned} \tag{5.4}$$

From these equations, it is evident that $\boldsymbol{z}(\Delta)_{\boldsymbol{\alpha}}$ and $\boldsymbol{z}(\Delta)_{\boldsymbol{\beta}}$ are not equal. This discrepancy highlights a key issue: microphones positioned on circles with different radii, even if they initially receive identical signals, will not maintain signal equivalence after rotation.

Consequently, the assumption that the signal obtained from the Δ-rad-rotated pCMA-a, $\boldsymbol{x}(\Delta)_{\mathrm{pCMA-a}}$, represents the signal before rotation for an NCMA is inherently flawed, as the rotated pCMA-a does not accurately correspond to the Δ-rad-rotated NCMA.

5.3.2 Two-stage method for estimating signal before rotation on an NCMA

Each rotation modifies the correspondence between the NCMA and its associated pCMA, necessitating a more sophisticated method. To address this challenge, we propose a two-stage method for estimating the signal on the NCMA before rotation. This approach accommodates the dynamic relationship between the NCMA and its corresponding pCMA after each rotation, providing a more robust and accurate solution

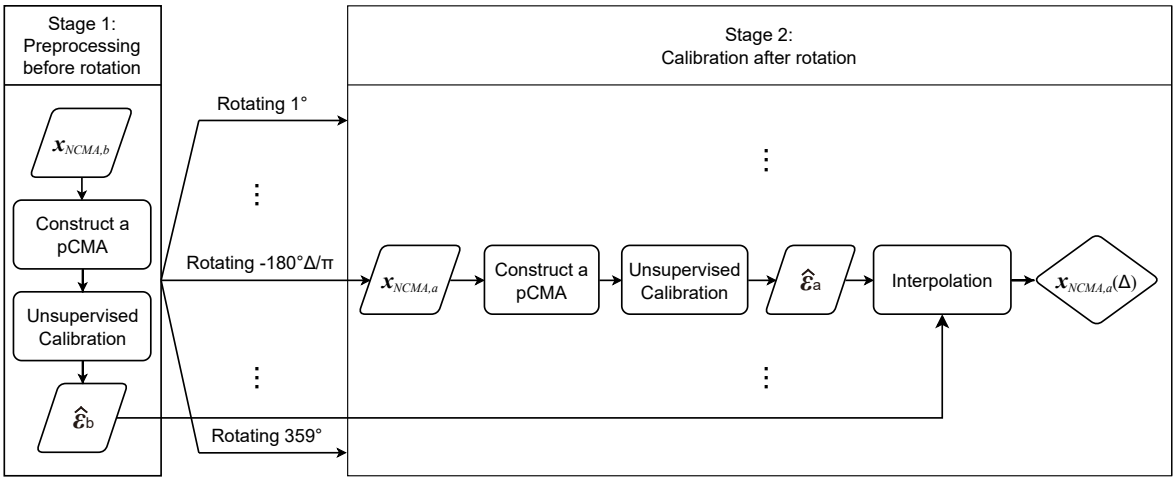

Figure 5.2: *Overall processing flow of the two-stage method.*

for signal estimation in these scenarios. Below, we outline the methodology and its advantages in handling this complex problem. The overall processing flow of the two-stage method is illustrated in Figure 5.2.

Stage 1: Preprocessing before rotation

In the first stage, preprocessing is conducted while the NCMA remains in its initial, unrotated position. The signals captured by the microphones at this position are denoted as

$$\boldsymbol{x}_{\mathrm{NCMA,b}} = \begin{bmatrix} x_{1,b} & \cdots & x_{M,b} \end{bmatrix}^{\mathsf{T}}. \tag{5.5}$$

It is important to note that $\boldsymbol{x}_{\mathrm{NCMA,b}}$ is not the target signal we aim to estimate through interpolation. Instead, it is used to compute the spatial filter required for beamforming. When the NCMA rotates by $-\Delta$ radians, the resulting microphone signals are represented as $\boldsymbol{x}_{\mathrm{NCMA,a}}$. Our objective is to estimate the signal $\boldsymbol{x}_{\mathrm{NCMA,a}}(\Delta)$, which corresponds to the state of $\boldsymbol{x}_{\mathrm{NCMA,a}}$ as if it were recorded at the original, before-rotation

position of the NCMA. This eliminates the need to recompute the spatial filter after each rotation.

In this before-rotation state, the NCMA is associated with a corresponding pCMA, referred to as pCMA-b, where "b" stands for "before rotation". For pCMA-b, we initially assume an error vector $\boldsymbol{e}_{\mathrm{b}}$ with a known value of zero, thereby defining the microphone distribution for this pCMA-b as

$$\boldsymbol{P}(\boldsymbol{e}_{\mathrm{b}}) = \left[0 + e_{\mathrm{b},1} \quad \cdots \quad \tfrac{2\pi(M-1)}{M} + e_{\mathrm{b},M}\right]^{\mathsf{T}}. \tag{5.6}$$

Using the signal $\boldsymbol{x}_{\mathrm{NCMA,b}}$ and the assumed error vector $\boldsymbol{e}_{\mathrm{b}}$, unsupervised calibration is then applied to obtain an estimate of the true error vector for pCMA-b denoted by $\hat{\boldsymbol{\epsilon}}_{\mathrm{b}}$. This calibration step forms the foundation for the subsequent signal estimation process.

Stage 2: Calibration after rotation

The second stage takes place after the NCMA has rotated by $-\Delta$ radians. In this stage, the microphone signals $\boldsymbol{x}_{\mathrm{NCMA,a}}$ and the initial error vector $\boldsymbol{e}_{\mathrm{a}}$ are used to estimate the error vector for pCMA-a, following the method outlined earlier.

A key aspect of this process is the consistent use of the same reference channel during the unsupervised calibration in both the first and second stages. By maintaining a fixed reference channel for computing the cost functions (4.12) and (4.14), we ensure that both pCMA-b and pCMA-a are aligned on the same circle. This geometric consistency is essential for the effectiveness of the interpolation method, as it allows for accurate signal reconstruction across the two stages and ensures the coherence of the overall approach.

It is important to note that, aside from using the same channel as the reference, the

unsupervised calibrations in the two stages are entirely independent and uncorrelated. Once $\hat{\boldsymbol{\epsilon}}_{\mathrm{b}}$ and $\hat{\boldsymbol{\epsilon}}_{\mathrm{a}}$ have been obtained, we can determine what $\boldsymbol{x}_{\mathrm{NCMA,a}}$ would be like if it were observed at the original position before rotation. This reconstruction is achieved using the following equation:

$$\boldsymbol{x}_{\mathrm{NCMA,a}}(\Delta) = \boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}}_{\mathrm{b}})\boldsymbol{U}_M(\Delta)\boldsymbol{U}_M(\hat{\boldsymbol{\epsilon}}_{\mathrm{a}})^{-1}\boldsymbol{x}_{\mathrm{NCMA,a}}. \tag{5.7}$$

Through this approach, the GSFI technique is successfully implemented on the NCMA, enabling accurate estimation of the before-rotation signal and avoiding the need for re-estimating spatial filters after rotations.

Optimization for practical applications

In practical applications, since the NCMA's position before rotation is fixed, the first stage of unsupervised calibration only needs to be executed once. The estimated result, $\hat{\boldsymbol{\epsilon}}_{\mathrm{b}}$, can then be saved for use in subsequent interpolation tasks. However, because the microphone distribution on pCMA-a changes significantly after each rotation, the second stage of unsupervised calibration must be performed for every new rotation to obtain the corresponding estimated error vector $\hat{\boldsymbol{\epsilon}}_{\mathrm{a}}$, just as Figure 5.2 shows.

To optimize efficiency, we can take advantage of the fact that, with an angular resolution of 1°, there are only 359 possible rotational positions for the NCMA. Instead of performing unsupervised calibration repeatedly, the error vectors for all 359 positions of pCMA-a can be pre-estimated and stored in advance. Each error vector corresponds to one rotational position of the NCMA.

By estimating all 359 error vectors of pCMA-a in advance, the second-stage unsupervised calibration is replaced by a lookup operation. During the second stage, rather than recalculating $\hat{\boldsymbol{\epsilon}}_{\mathrm{a}}$ through iterative optimization, we simply select the precomputed

error vector corresponding to the current rotational position. This selection is based on the rotation angle and the result of the first-stage calibration, $\hat{\boldsymbol{\epsilon}}_{\mathrm{b}}$. By avoiding iterative optimization after each rotation, this approach significantly improves the overall efficiency of the interpolation process while maintaining accuracy.

5.4 Simulated experimental evaluation

To validate the effectiveness of the proposed GSFI technique on NCMAs, we conducted simulation experiments that closely followed the methodology outlined in Chapter 4. These experiments aimed to demonstrate that the two-stage GSFI method can successfully interpolate sound fields on NCMAs, thereby enabling the accurate estimation of before-rotation signals.

In contrast to the experiments in Chapter 4, where we investigated various properties of unsupervised calibration (*e.g.*, Section 4.4.6 explored the relationship between the signal length used for unsupervised calibration and the interpolation accuracy), the focus in this chapter shifted entirely to evaluating the interpolation performance of the proposed two-stage GSFI method, specifically the accuracy of estimating the before-rotation signals. As such, experiments related to validating the properties of unsupervised calibration were omitted.

In addition to verifying GSFI's ability to estimate before-rotation signals on NCMAs, we also evaluated its effectiveness in enhancing source signals when integrated with beamforming techniques. The experimental setup was consistent with the setup used in Chapter 4, except for the substitution of unes-CMAs with NCMAs. Below, we detail the experimental setup and procedure.

5.4.1 Setup

Dataset and preprocessing

We utilized the SiSEC database [72], which provides high-quality speech recordings sampled at 16 kHz. From this database, we curated a dataset comprising eight distinct speech signals, balanced equally between male and female speakers. These speech signals were convolved with room impulse responses (RIRs) to simulate a reverberant environment with a reverberation time of approximately 100 ms, as described in Chapter 3. For time-frequency (TF) domain analysis, we applied the short-time Fourier transform (STFT) with the following parameters:

- **Window type**: Blackman window.
- **Window length**: 64 ms length.
- **Overlap**: 1/8 overlap.

Simulated experimental setup for NCMAs

We constructed 10 different M-channel NCMAs by randomly selecting microphones from various CMAs sharing the same center but with radii varying from 0.03 m to 0.1 m. These arrays were placed in a noise-free room to record the sound signals.

Like in Chapter 4, we still introduced a known angular error $\omega_i(^\circ)$ and an additional unknown error $\epsilon_i(^\circ)$, $i \in \{0, ..., M-1\}$, to simulate the non-uniform unknown microphone distribution. ω_i adhered to a Gaussian distribution characterized by a mean of zero and a standard deviation equivalent to $\sqrt{200}^\circ$, and ϵ_i also conformed to a Gaussian distribution, with a mean of zero and variances systematically spanning from $(0^\circ)^2$ to $(\sqrt{500}^\circ)^2$ in discrete steps of $(\sqrt{10}^\circ)^2$. All the errors were independently and

identically distributed. For every specified variance of ϵ_i, a comprehensive set of 100 random samples was generated, thereby guaranteeing a rigorous and reliable statistical evaluation.

It is crucial to emphasize that while ω_i and ϵ_i are used to construct the unknown microphone distribution, the purpose of unsupervised calibration in the GSFI method for the NCMAs is **not** to precisely determine the true error angles of each microphone, i.e., $\omega_i+\epsilon_i$. Instead, the aim is to identify a microphone distribution on the hypothetical pCMA such that the signals received by the pCMA closely approximate those captured by the NCMA.

The incorporation of the known error angle ω_i and the unknown error angle ϵ_i serves two key purposes. First, the inclusion of both ω_i and ϵ_i ensures the arbitrariness and uncertainty of the microphone distribution, aligning the experimental setup more closely with real-world applications, where the microphone arrangement is often irregular and partially unknown. Second, through comparative experiments, we can evaluate the performance of the previous unes-SFI method, which lacks an unsupervised calibration process, on the NCMA under distinct conditions: when the actual error angle $(\omega_i + \epsilon_i)$ is unknown, and ω_i is mistakenly treated as the actual error angle; and when the actual error angle $(\omega_i + \epsilon_i)$ is known.

Once again, the simulation setup leveraged two distinct simulated environments for performing unsupervised calibration and evaluation, respectively, to emphasize a key advantage of the proposed method: its ability to perform robustly even when the acoustic environment changes. As noted earlier, even dealing with NCMAs, once the error vector for the pCMA is estimated, no additional calibration is required—–even when applied to new environments. Figure 5.3 and Figure 5.4 illustrate the two simulated environments.

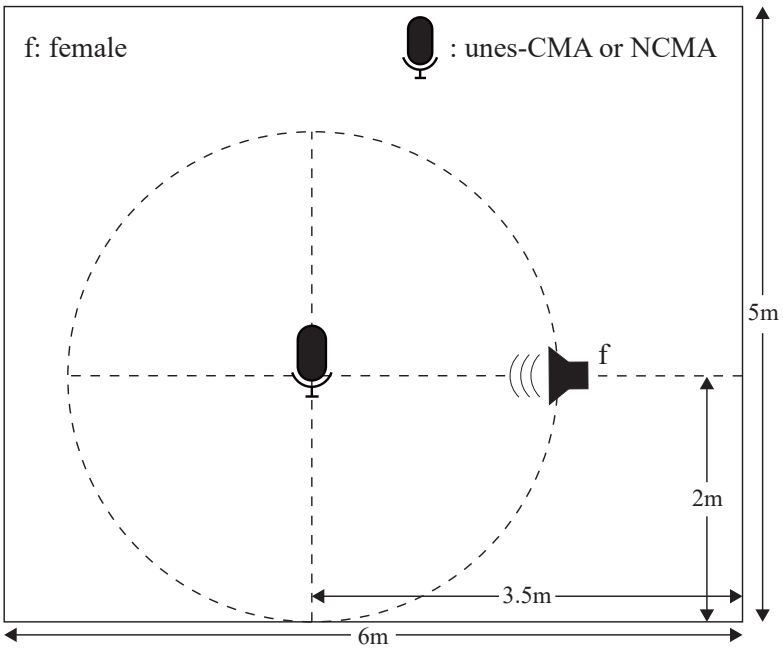

Figure 5.3: *Simulated environment during unsupervised calibration. One of the speech signals (female) is placed at the position of* 0°*, with the microphone array located at the center of the circle.*

The simulation process was structured as follows: initially, the sound field was simulated using the microphone array configuration depicted in Figure 5.3. A single speech signal was utilized to perform unsupervised calibration, yielding all of the estimated error vectors for the pCMA-b and pCMA-a in the two-stage method. Subsequently, after calibration, the microphone array and sound source were relocated to one of the new positions shown in Figure 5.4. At these new positions, a new sound field was simulated after the microphone array rotated by Δ radians. Using the newly simulated sound field, the observation signals were estimated as if they were captured at the original reference position before rotation, with the rotational angle $\phi = \Delta\pi/180^\circ$ being a known value.

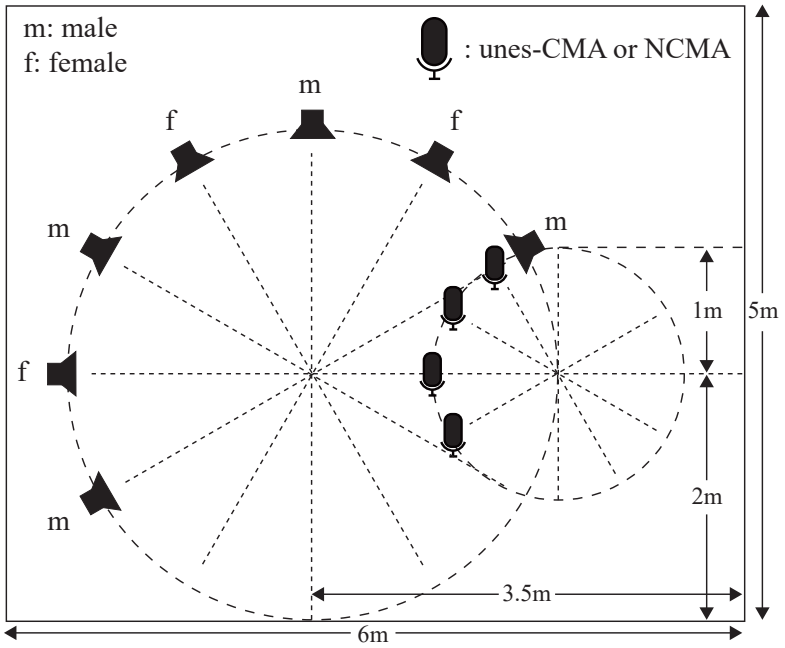

Figure 5.4: *Simulated environment for evaluation. The remaining seven speech signals are positioned at* $30°, \ldots, 210°$*, respectively. There are four optional positions of the microphone array, located at* $120°, \ldots, 210°$*. In the experiments to evaluate the performance of GSFI, one speech signal and one position of the microphone array are selected each time, resulting in 28 combinations. In the experiments to evaluate the performance of source enhancement, two speech signals are chosen and mixed as the observed signal, with the microphone array positioned at* $180°$*.*

Evaluation criteria

Similarly, the performance evaluation of GSFI on NCMAs was carried out in two distinct sets of experiments to assess its effectiveness in controlled and mixed-source scenarios.

The initial experiments were designed to evaluate the performance of GSFI in a controlled setting, involving a single, unmixed sound source. The performance was

measured using the Signal-to-Error Ratio (SER) [42, 43, 75–77], which is defined as

$$\mathrm{SER}_{m,f} = 10\log_{10}\left(\frac{\sum_t |x_{m,t,f}|^2}{\sum_t |\hat{x}_{m,t,f} - x_{m,t,f}|^2}\right), \tag{5.8}$$

where $x_{m,t,f}$ represents the TF domain signal for the mth channel at time frame t and frequency bin f, and $\hat{x}_{m,t,f}$ denotes $x_{m,t,f}$'s estimate. The number of microphones, M, was varied between 4 to 7 to investigate the impact of array size on performance. Additionally, the array was rotated to simulate different orientations, enabling an assessment of GSFI's robustness to changes in rotation angles.

The second set of experiments aimed to evaluate the source enhancement performance of GSFI and compare it with other methods in scenarios involving multiple mixed sound sources. The Minimum Power Distortionless Response (MPDR) beamformer [46, 47, 82] was used to enhance the target source. The performance metrics are based on the source-to-distortion ratio (SDR), which quantifies the overall quality of the enhanced signal, and the source-to-interference ratio (SIR), a metric for measuring the suppression of interfering sources in the enhanced signal [78]. The MPDR beamforming filter was computed using the covariance matrix of the interference signal and the relative transfer function (RTF) [17, 79] as described in [42, 43]. The RTF was determined from the RIR of the target source to each microphone. For these experiments, two sound sources were randomly selected from Figure 5.4 and mixed into the observation signal. The angular separation between the two sources was systematically varied at intervals of $30^\circ, 60^\circ, \ldots, 180^\circ$.

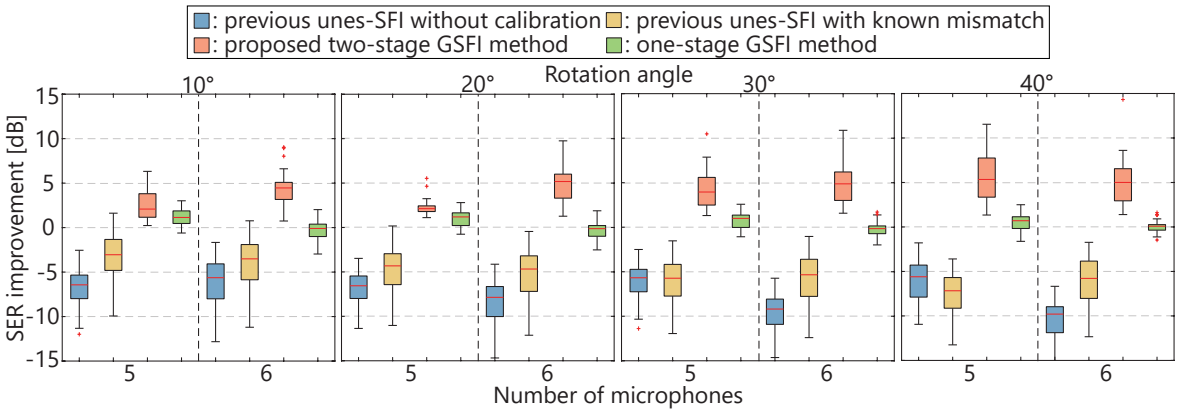

Figure 5.5: *Boxplots of mean SER improvement on NCMAs at frequencies up to 1 kHz across various scenarios with different M and ϕ values.*

5.4.2 Channelwise SER improvements

This section presents the experimental results for channelwise SER improvements across various scenarios involving different values of M and ϕ, achieved using multiple methods on diverse NCMAs. The sound source and NCMA positions were varied as illustrated in Figure 5.4. SER improvement served as the evaluation metric to measure the enhancement achieved through signal processing, comparing the interpolated signal's SER with the baseline SER, which was determined by comparing the uninterpolated signal after rotation with the target signal before rotation. The mean SER improvements relative to cases without interpolation are summarized in Figure 5.5. Each box plot includes 280 samples, representing the mean SER improvement for seven sound sources at four NCMA locations, with 10 unique NCMAs per location.

The results clearly demonstrate that the proposed two-stage GSFI method consistently achieves significant SER improvements across all tested scenarios on NCMAs. Using our proposed two-stage GSFI method, we achieved an average SER improvement of up to 15 dB higher compared to the previous unes-SFI method. Although the achieved SER improvement score is only about 5 dB and slightly smaller than

that obtained with unes-CMAs in Chapter 4, the enhancement is still substantial and demonstrates the robustness of the method. In contrast, directly applying the unes-SFI method, developed for unes-CMAs, fails to deliver satisfactory results for NCMAs, even when the error angles of each microphone are known. This discrepancy arises because unes-SFI assumes the NCMA behaves like a CMA, leading to a fundamental mismatch between the sound signals used for interpolation and the error angle information used to compute the compensation matrix. Such a mismatch significantly degrades the interpolation performance.

The proposed method addresses this issue through unsupervised calibration, which identifies corresponding positions on a pCMA that align with the received sound signals. This alignment minimizes errors in the calculation of the compensation matrix, resulting in superior interpolation performance. Additionally, we evaluated a one-stage GSFI method, which utilized only one pCMA to directly estimate the before-rotation signal and treated it as the NCMA's before-rotation signal. However, the one-stage method demonstrated smaller improvements compared to the two-stage approach with a decrease up to 5 dB, as it failed to account for how each rotation alters the correspondence between the NCMA and its associated pCMA. This oversight limits its effectiveness, further underscoring the robustness of the proposed two-stage GSFI method.

5.4.3 Results of source enhancement with batch processing

In this experiment, we evaluated the source enhancement performance of the MPDR beamformer under varying conditions. The number of microphones was fixed at $M = 6$, while the rotation angle ϕ was varied across 10°, 20°, 30°, and 40°. To establish a benchmark, we first computed the MPDR beamformer filter weight $\boldsymbol{w}$ using the RTF and the multichannel STFT spectrogram from the NCMA in its original, unrotated posi-

Table 5.1: *Abbreviations for spectrograms from different evaluation settings*

Spectrograms from different evaluation settings	Status of evaluation conditions					
	B	R	I	K	C	T
B0: No beamforming	0	0	0	0	0	0
R0: No rotation	1	0	0	0	0	0
I0: No interpolation	1	1	0	0	0	0
C0: unes-SFI without calibration	1	1	1	0	0	0
K1: unes-SFI with known mismatches	1	1	1	1	0	0
T0: one-stage GSFI	1	1	1	0	1	0
T1: proposed two-stage GSFI	1	1	1	0	1	1

tion. This baseline scenario, denoted as R0, utilized true, uninterpolated signals and represented the optimal performance case. For simplicity and clarity, we use the same abbreviations as in Chapter 4 to denote the evaluation conditions for spectrograms obtained from different methods, where B represents Beamforming, R represents Rotation, I represents Interpolation, K represents Known Mismatches, C represents Calibration, T represents Two-stage processing, Index 0 represents off, and Index 1 represents On.

The spectrograms obtained from different evaluation settings were subsequently processed using the precomputed MPDR beamformer weight $\boldsymbol{w}$ to generate the estimated target signal. A summary of these spectrograms is provided in Table 5.1. For additional comparison, an unprocessed scenario (B0), where the raw microphone signals were treated as the target signal, was included alongside R0.

Figure 5.6 presents the SDR and SIR results for the different methods across various scenarios. The proposed two-stage GSFI method (T1) consistently delivers a 10 dB higher performance than previous unes-SFI methods (C0 and K1) and a 5 dB higher performance than the one-stage GSFI method (T0) across all simulated conditions. This demonstrates the robustness of the proposed method in mitigating the effects of rotation for source enhancement when applied to NCMAs.

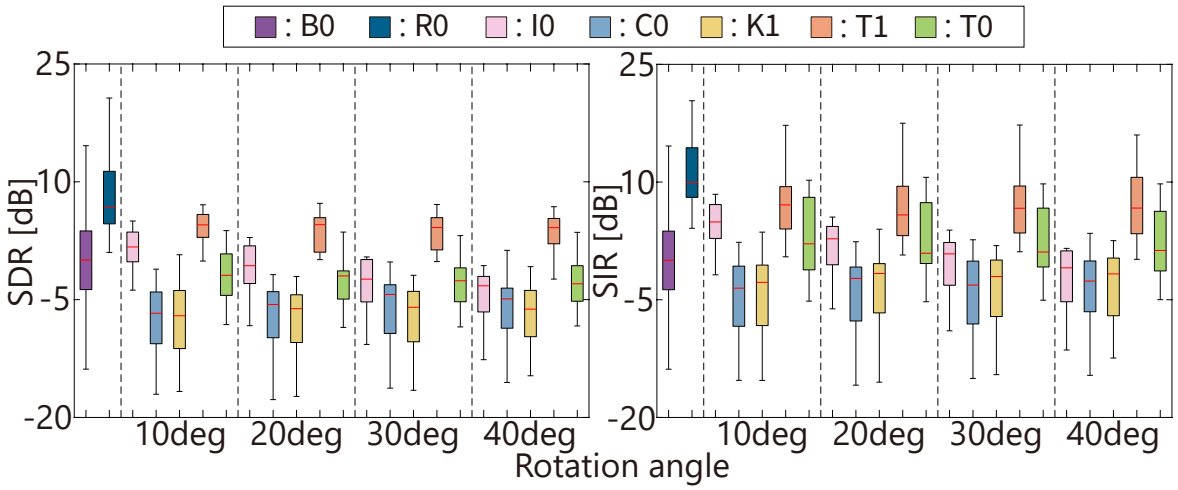

Figure 5.6: *Boxplots of SDR and SIR obtained by MPDR beamformer in seven situations: unprocessed* (B0)*, no rotation of the NCMA* (R0)*, without interpolation when the NCMA rotates* (I0)*, with previous unes-SFI without calibration when the NCMA rotates* (C0)*, previous unes-SFI with known mismatches when the NCMA rotates* (K1)*, with one-stage GSFI method when the NCMA rotates* (T0) *and our proposed two-stage method when NCMA rotates* (T1)*.*

Notably, K1 did not achieve better performance than C0, despite having access to the actual error angles. Furthermore, both C0 and K1 produced worse SDR and SIR results than I0, highlighting the limitations of directly treating an NCMA as a CMA and applying unes-SFI without the proposed two-stage calibration. These findings underscore the importance of the proposed method in effectively addressing the unique challenges posed by NCMAs.

5.4.4 Results of source enhancement with online processing

In this experiment, we evaluated the application of the GSFI technique to online beamforming on NCMAs, focusing on scenarios where the acoustic transfer system (ATS) underwent continuous, dynamic changes. To adapt to these conditions, we

employed a smoothing factor α for online updating of the spatial covariance matrix (SCM), as described in Section 2.3.3, and used the Sherman–Morrison formula to reduce the computational complexity of inverting the covariance matrix within the MPDR beamforming framework.

The algorithm and experimental setup followed the methodology outlined in Section 3.4.4. Two sound sources, each lasting 40 s, were positioned at 30° and 210°, while the microphone array was placed at 180°, as depicted in Figure 5.4. The frame length was set to 256 ms, and a segmental SDR was calculated at intervals of 1 s to evaluate source enhancement performance. The smoothing factor α was empirically set to 0.99, which provided the highest segmental SDR. The inversion of the covariance matrix $\hat{\boldsymbol{V}}_f^{-1}$ was initialized using the first 10 frames. During the experiment, the NCMA underwent two gradual rotations: the first, from 0° to 20°, began at 10 s, and the second, from 20° to 40°, began at 30 s. Both rotations occurred at a constant speed of 0.01° per time sample (equivalent to 160° per second), simulating realistic human or humanoid robot rotation speeds. Observations were generated by concatenating data from the NCMA at various angles.

Figure 5.7 shows the segmental SDRs obtained for different methods with $M = 6$. The R0 scenario continues to achieve the highest source enhancement performance (5 dB), serving as the optimal benchmark. Both C0 and K1 showed inferior performance compared to I0, echoing the trends observed in Figure 5.6. In contrast, the proposed method (T1) demonstrated significant performance improvements, with an average performance increase of 15 dB and 5 dB over previous interpolation methods (C0 and K1) and the one-stage GSFI method (T0), respectively. However, in certain cases, T1 failed to consistently achieve high SDRs, with performance occasionally falling below that of the B0 scenario.

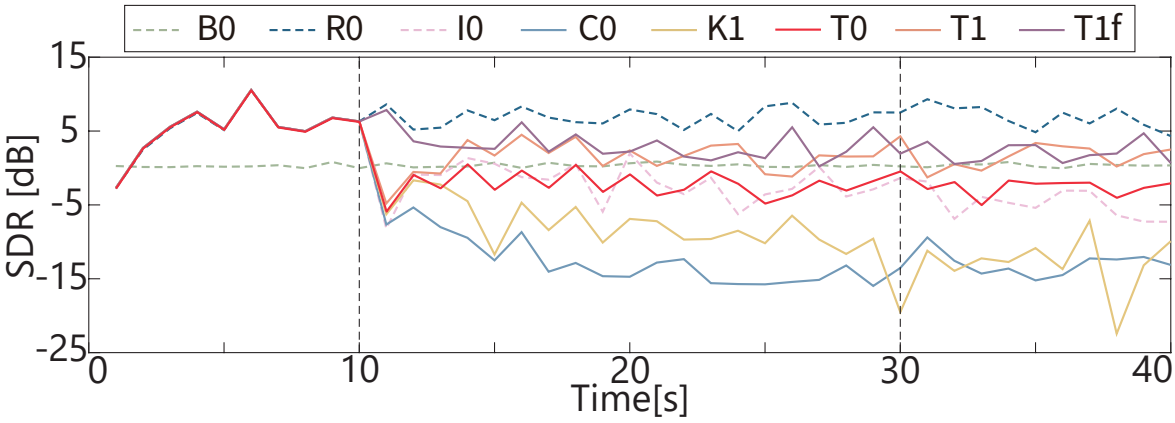

Figure 5.7: *Segmental SDR every* 1 s *with* $M = 6$ *on an NCMA, where the two vertical dashed lines indicate the time points when the rotation started:* $0° \Rightarrow 20° \Rightarrow 40°$. B0 *shows the mixture itself,* R0 *shows the case where no rotation occurs,* I0 *shows online processing without interpolation,* C0 *and* K1 *respectively show online processing using previous unes-SFI without calibration and with known mismatches,* T0 *shows online processing with the one-stage GSFI method,* T1 *shows online processing with the proposed two-stage GSFI method, and* T1f *shows online processing with the proposed two-stage GSFI method and the frozen SCM.*

As discussed in Chapter 4, during online beamforming, the interpolated signal was used to update the SCM in real-time. This process introduced inaccuracies in the SCM, likely contributing to the degradation in beamforming performance. To address this issue, we introduced the T1f method, which involved using the two-stage GSFI method but halting SCM updates during NCMA rotations. The SCM was iteratively updated only before the rotation, and no further updates were made once rotation commenced.

Interestingly, in contrast to the results presented in Figure 4.12 of Chapter 4, T1f achieved SDRs similar to those of T1, with both showing an approximate SDR of 3 dB. However, by preventing errors from propagating during SCM updates, T1f consistently outperformed the B0 scenario, a result not achieved by T1. This finding suggests that the performance degradation in T1 was likely due to inaccuracies introduced during SCM updates. Despite this improvement, the SDRs achieved by T1f remained close

to those of B0, indicating that there are still inherent limitations in applying GSFI to online beamforming with NCMAs.

5.5 Conclusion

In this chapter, we extended the GSFI framework to address the challenges posed by NCMAs and evaluated its performance across diverse experimental scenarios. By employing unsupervised calibration, we constructed a pCMA that closely replicates the received signal of the NCMA, effectively serving as a surrogate. This approach mitigates the spatial complexity introduced by the NCMA's non-circular geometry, enabling the application of methods originally designed for circular arrays.

Furthermore, using a two-stage strategy, we resolved the limitation that a pCMA can only act as a substitute for an NCMA at a specific position. As the NCMA rotates to different positions, it corresponds to different pCMAs respectively. The two-stage GSFI method effectively bridges this gap, facilitating accurate interpolation of sound signals on an NCMA.

Comprehensive simulation experiments proved the proposed method's effectiveness in accurately estimating before-rotation signals. The results of channelwise SER evaluations showed that the two-stage GSFI method consistently outperforms previous methods, such as unes-SFI without calibration and the one-stage GSFI method, across various experimental conditions. Furthermore, in downstream array signal processing tasks, the proposed method exhibits a certain level of capability to enhance source separation performance, overcoming the challenges associated with the irregular geometries of NCMAs.

Despite these advancements, residual challenges remain, particularly in the context of online beamforming. Future research could focus on developing alternative interpo-

lation frameworks to further enhance performance in complex and dynamic acoustic environments.

6 Conclusions

6.1 Summary of This Thesis

This thesis has presented a systematic study and development of techniques for rotation-robust microphone array signal processing, with a particular focus on wearable auditory systems equipped with circular microphone arrays (CMAs). These systems, designed for applications like augmented hearing and humanoid robots, face unique challenges due to the dynamic nature of their operation. In such scenarios, the rotation of the microphone array induces time-variant acoustic transfer systems (ATS), significantly complicating real-time signal processing and necessitating innovative solutions to maintain robust signal processing performance.

The core contribution of this thesis is the development of a generalized sound field interpolation (GSFI) framework, designed to overcome the limitations of existing methods. Specifically, GSFI overcomes the reliance on equally spaced CMAs (es-CMAs) and progressively extends its applicability to handle more complex scenarios, including unequally spaced CMAs (unes-CMAs), unknown microphone distributions, and nearly-circular microphone arrays (NCMAs).

Chapter 2 focused on establishing the theoretical background and foundational concepts underlying the methods proposed in this thesis. The theory of the beamforming, with a particular emphasis on Minimum Power Distortionless Response (MPDR) beamforming, was briefly introduced as a representative example of array signal pro-

cessing that this research aims to make rotation-robust. The chapter further detailed the formulation of sound field interpolation (SFI), which serves as the cornerstone of the proposed methods. A key contribution of this chapter was the analysis of the periodicity and singularity properties of SFI, along with the influence of the Nyquist component. These aspects, which are critical to understanding the framework, were introduced and are further discussed in subsequent chapters. Additionally, the chapter outlined the application of SFI in both batchwise and online beamforming, setting the stage for its integration into source enhancement techniques.

The first major contribution of this thesis was the development of the unequally spaced sound field interpolation (unes-SFI) method. In Chapter 3, the focus was on breaking through the limitation of existing SFI by enabling their application on unes-CMAs. By compensating for positional deviations, unes-SFI effectively transforms the time-variant ATS on an unesCMA into a time-invariant ATS on an esCMA. This innovation enables the estimation of before-rotation signals of an unesCMA, achieving rotation-robust beamforming. Furthermore, building upon the analysis presented in the previous chapter, a more comprehensive examination of the properties of the unes-SFI and the influence of the Nyquist component was conducted. This in-depth analysis culminated in the derivation of a more generalized conclusion regarding the behavior and applicability of unes-SFI. Through a series of simulation experiments, this method demonstrated substantial improvements in reconstructing before-rotation signals and maintaining beamforming accuracy in dynamic conditions, establishing a foundation for robust signal processing in wearable CMA applications.

The second major contribution of this thesis addresses the practical challenge of unknown microphone configurations in wearable CMAs. To tackle this issue, Chapter 4 introduces the GSFI framework, which combines unes-SFI with an iterative optimiza-

tion method known as unsupervised calibration. This calibration technique estimates the positional errors of individual microphones, accurately determining their distribution on the array without requiring prior knowledge of their locations. With the estimated microphone positions, unes-SFI is employed to reconstruct the before-rotation signal of the microphone array. This enables the use of pre-estimated spatial filters without the need for re-estimation, thus achieving rotation-robust beamforming. By integrating unsupervised calibration with unes-SFI, the GSFI framework significantly enhances its practicality, making it suitable for real-world scenarios where microphone positions cannot be pre-defined or controlled. Simulation experiments results validate the efficacy of GSFI, demonstrating its ability to maintain high interpolation accuracy and achieve effective source enhancement performance on unes-CMAs with unknown microphone distributions.

Recognizing that wearable arrays often deviate from a perfect circular geometry due to the shape of the robot's head, the final contribution of this thesis, presented in Chapter 5, focuses on extending the GSFI framework to accommodate NCMAs. To address this challenge, a virtual pseudo-CMA (pCMA) is constructed through unsupervised calibration, which effectively reduces the spatial complexity inherent in NCMAs. Furthermore, a two-stage strategy is introduced to tackle the dynamic changes in the correspondence between the pCMA and the NCMA during rotation. This approach enables accurate signal reconstruction and ensures rotation-robust beamforming. This extension not only demonstrates the feasibility of GSFI for handling non-ideal array geometries but also highlighted its adaptability to real-world conditions, where deviations from ideal circular geometries are unavoidable.

6.2 Future Work

Despite the significant advancements achieved in this thesis, there are some limitations to the current GSFI framework that merit further investigation. Addressing these limitations offers promising avenues for future research:

1. **Performance with a Small Number of Microphones**: The GSFI framework does not perform well when the number of microphones in the array is small. This limitation arises from the reduced spatial information available for interpolation and beamforming, which compromises the framework's ability to accurately reconstruct signals. Future work could focus on exploring alternative methods tailored for sparse arrays, ensuring robust performance even in low-sensor-count scenarios.

2. **Sensitivity to Microphone Distribution**: The performance of GSFI is highly sensitive to the spatial distribution of microphones. In extreme cases, where microphones are clustered closely together, the interpolation process suffers due to the lack of diversity in spatial information. Future studies could investigate ways to enhance the robustness of GSFI against such irregular distributions, possibly by incorporating adaptive weighting mechanisms or distribution-aware optimization techniques.

3. **Challenges with NCMAs**: While GSFI has been extended to accommodate NCMAs, the framework still struggles to achieve significantly high performance on these non-ideal geometries. This limitation stems from the inherent complexity of interpolating signals on arrays that deviate substantially from a circular structure. Further research could focus on refining the GSFI framework for NCMAs, possibly by incorporating geometry-specific adjustments or leveraging advanced

modeling techniques.

4. **Exploration of Alternative Models and Algorithms**: To address the aforementioned limitations and enhance overall performance, future work could explore alternative models and algorithms. For example:

   - **Regression Models**: Methods such as Gaussian Process Regression (GPR) could provide a flexible approach for modeling sound fields, enabling improved interpolation and handling of irregular microphone distributions.
   - **Deep Learning Strategies**: Data-driven approaches, such as neural networks, could be leveraged to learn complex sound field representations and address challenges posed by non-ideal microphone distributions and geometries. These strategies might also offer greater adaptability to diverse acoustic environments and dynamic conditions.

5. **More Complex Acoustic Environments**: In this thesis, we focus solely on rotational motion. However, in real-world applications, robots often rotate their heads while simultaneously undergoing translational motion. Eliminating the impact of translational motion on array signal processing is therefore a crucial challenge. While our proposed method has the potential to be extended to such scenarios, further research and experimental validation are required to fully assess its effectiveness. Future work could focus on exploring the application of GSFI in real-world environments, where exists the rotational and translational motion simultaneously.

By addressing these limitations and exploring innovative methodologies, future research can further improve the robustness, adaptability, and practical applicability

of the GSFI framework, enabling more effective solutions for dynamic and complex acoustic scenarios.

Acknowledgments

This thesis was completed under the meticulous guidance of Professor Tomoki Toda, to whom I owe my deepest gratitude. Toda-sensei provided me with the invaluable opportunity to explore the fascinating field of array signal processing. His profound expertise and rigorous academic approach have significantly contributed to my academic and personal growth. Moreover, his gentle, modest, diligent, and conscientious demeanor has profoundly influenced me, both professionally and personally. Throughout my journey at Nagoya University, Toda-sensei has consistently offered encouragement and support with patience, guiding me through challenges and setbacks in both life and study. His wisdom, urging me to make bold hypotheses and verify them cautiously and to go to the ends of the earth to seek answers with hands and feet, has been a guiding light. Through his mentorship, I was able to overcome my initial lack of knowledge in array signal processing and successfully complete my studies within a limited timeframe. His guidance has not only deepened my understanding of this remarkable field but also kindled my passion for it, inspiring me to pursue it as a lifelong career.

I also extend my heartfelt gratitude to Professor Yukoh Wakabayashi for his substantial guidance and assistance. When I began my PhD journey, Wakabayashi-sensei's patient mentorship and exemplary teaching helped me become familiar with my research focus and set me on the right track. His rigorous and pragmatic academic attitude, combined with his perseverance, deeply inspired me, and his gentle demeanor remains

a model for me to emulate. Without his support, the successful completion of this thesis would not have been possible.

I am profoundly grateful to Professor Kazuya Takeda, whose insightful guidance during seminars significantly contributed to the completion of my research. I also owe special thanks to Professor Nobutaka Ono for his invaluable feedback on my work and to Professor Wen-chin Huang for his support during the writing of my thesis.

My sincere appreciation extends to the staff of the Toda Laboratory, Nami Noro and Mayuko Hayashi, for their kind assistance. Their tremendous help, particularly during my initial transition to life in Japan, enabled me to focus fully on my research without distraction, for which I am deeply grateful.

I am equally thankful to my friends and colleagues in the laboratory: Fengji Li, Xiaohan Shi, Shaowen Chen, Rui Wang, Ding Ma, and Jiajun He, for their companionship and collaboration during long hours of work. Special thanks are due to Fengji Li and Xiaohan Shi for their valuable advice regarding my future career. I look forward to sharing more memorable moments with all of you, whether traveling, cycling, or enjoying good food together.

I also wish to express my heartfelt gratitude to all members of the Toda Laboratory for their continuous support. Their efforts in maintaining the laboratory's infrastructure, including servers and computers, and their contributions to the daily operations and collaborative environment of the lab, have been invaluable to my research and academic journey.

Finally, I express my deepest gratitude to my family and my girlfriend for their unwavering care and support over the years. My family has always respected my decisions without imposing any demands, providing me with a safe harbor to face life's storms. Without their supervision and companionship, I would not have been able to

complete this thesis or find the motivation to move forward. I am especially grateful to my girlfriend. Thank you for standing by my side, tolerating my shortcomings. Thank you for your constant guidance and encouragement, and even more so for your support and reassurance during times when I felt lost or discouraged. You have always cared deeply for me and consistently kept our shared future in mind. Having you by my side is the greatest blessing in my life.

Throughout the past three years of my doctoral studies, I have received immense support from countless individuals, far too many to name here. I am profoundly grateful to each and every one of them. Although this thesis is now nearing completion, it remains imperfect. I will continue to strive for improvement, ensuring that I live up to the generosity and kindness of my mentors, friends, and family.

References

[1] K. Yamaoka, N. Ono, S. Makino, and T. Yamada, "Time-frequency-bin-wise switching of minimum variance distortionless response beamformer for underdetermined situations," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7908–7912.

[2] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6855–6859.

[3] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3461–3475, 2021.

[4] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 165–172.

[5] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5–8, 2006. Proceedings 6*. Springer, 2006, pp. 601–608.

[6] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[7] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2011, pp. 189–192.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2009.

[9] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 236–240.

[10] L. Pandey, A. Kumar, and V. Namboodiri, "Monoaural audio source separation using variational autoencoders." in *Interspeech*, 2018, pp. 3489–3493.

[11] E. Karamatlı, A. T. Cemgil, and S. Kırbız, "Audio source separation using variational autoencoders and weak class supervision," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1349–1353, 2019.

[12] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[13] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for de-

termined audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1601–1615, 2019.

[14] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv preprint arXiv:1808.00892*, 2018.

[15] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[16] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural Computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[17] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[18] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 5. IEEE, 2003, pp. V–469.

[19] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ica and residual crosstalk subtraction," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 87, no. 8, pp. 1941–1948, 2004.

[20] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Sound source separation of moving speakers for robot audition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3685–3688.

[21] K. Nakadai, H. Nakajima, G. Ince, and Y. Hasegawa, "Sound source separation and automatic speech recognition for moving sources," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 976–981.

[22] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "Multimodal blind source separation for moving sources," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 125–128.

[23] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

[24] S. M. Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 201–204.

[25] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel nmf and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2017.

[26] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational em algorithm for the separation of moving sound sources," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[27] Y. Wang, A. Politis, and T. Virtanen, "Attention-driven multichannel speech enhancement in moving sound source scenarios," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 221–11 225.

[28] Z. Yermeche, N. Grbic, and I. Claesson, "Beamforming for moving source speech enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE, 2005, pp. 25–28.

[29] Z. Yermeche, N. Grbic, and I. Claesson, "Moving source speech enhancement using time-delay estimation," *Int. Worshop on Acoust. Echo and Noise Control*, 2005.

[30] Z. Yermeche, N. Grbic, and I. Claesson, "Blind subband beamforming with time-delay constraints for moving source speech enhancement," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 8, pp. 2360–2372, 2007.

[31] S. M. Naqvi, M. Yu, and J. A. Chambers, "Multimodal blind source separation for moving sources based on robust beamforming," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 241–244.

[32] M. Barnard, P. Koniusz, W. Wang, J. Kittler, S. M. Naqvi, and J. Chambers, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 864–880, 2014.

[33] M. Taseska and E. A. Habets, "Blind source separation of moving sources using sparsity-based source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 657–670, 2017.

[34] W. Xu, T.-C. Liu, and H. Schmidt, "Beamforming based on spatial-wavelet decomposition," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002.* IEEE, 2002, pp. 480–484.

[35] W. Chen and X. Huang, "Wavelet-based beamforming for high-speed rotating acoustic source," *IEEE Access*, vol. 6, pp. 10 231–10 239, 2018.

[36] K. Nakadai, T. Matsui, H. G. Okuno, and H. Kitano, "Active audition system and humanoid exterior design," in *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, vol. 2. IEEE, 2000, pp. 1453–1461.

[37] V. Tourbabin and B. Rafaely, "Direction of arrival estimation using microphone array processing for moving humanoid robots," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2046–2058, 2015.

[38] V. Khersonskii, A. Moskalev, and D. Varshalovich, *Quantum Theory Of Angular Momemtum.* World Scientific Publishing Company, 1988. [Online]. Available: https://books.google.co.jp/books?id=nXcGCwAAQBAJ

[39] P. J. Kostelec and D. N. Rockmore, "Ffts on the rotation group," *Journal of Fourier analysis and applications*, vol. 14, pp. 145–179, 2008.

[40] W. Ma, H. Bao, C. Zhang, and X. Liu, "Beamforming of phased microphone array for rotating sound source localization," *Journal of Sound and Vibration*, vol. 467, p. 115064, 2020.

[41] J. Casebeer, J. Donley, D. Wong, B. Xu, and A. Kumar, "Nice-beam: Neural integrated covariance estimators for time-varying beamformers," *arXiv preprint arXiv:2112.04613*, 2021.

[42] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Rotation-robust beamforming based on sound field interpolation with regularly circular microphone array," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 771–775.

[43] Y. Wakabayashi, K. Yamaoka, and N. Ono, "Sound field interpolation for rotation-invariant multichannel array signal processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2286–2298, 2023.

[44] T. Nakashima, Y. Wakabayashi, N. Ono *et al.*, "Self-rotation-robust online-independent vector analysis with sound field interpolation on circular microphone array," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, 2024.

[45] G. Lian, Y. Wakabayashi, T. Nakashima, and N. Ono, "Self-rotation angle estimation of circular microphone array based on sound field interpolation," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1016–1020.

[46] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[47] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[48] B. G. Ferguson, "Minimum variance distortionless response beamforming of acoustic array data," *The Journal of the Acoustical Society of America*, vol. 104, no. 2, pp. 947–954, 1998.

[49] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2002.

[50] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2009.

[51] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of mvdr and mpdr beamformers," in *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in israel*. IEEE, 2010, pp. 000 416–000 420.

[52] D. Li, Q. Yin, P. Mu, and W. Guo, "Robust mvdr beamforming using the doa matrix decomposition," in *2011 1st International symposium on access spaces (ISAS)*. IEEE, 2011, pp. 105–110.

[53] C. Pan, J. Chen, and J. Benesty, "Performance study of the mvdr beamformer as a function of the source incidence angle," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 67–79, 2013.

[54] Y. Xiao, J. Yin, H. Qi, H. Yin, and G. Hua, "Mvdr algorithm based on estimated diagonal loading for beamforming," *Mathematical Problems in Engineering*, vol. 2017, no. 1, p. 7904356, 2017.

[55] S. Darlington, "Linear least-squares smoothing and prediction, with applications," *Bell System Technical Journal*, vol. 37, no. 5, pp. 1221–1294, 1958.

[56] R. Brown and J. Nilsson, *Introduction to Linear Systems Analysis*. Wiley, 1962. [Online]. Available: https://books.google.co.jp/books?id=GHZRAAAAMAAJ

[57] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[58] A. V. Oppenheim, *Discrete-time signal processing.* Pearson Education India, 1999.

[59] J. G. Proakis, *Digital signal processing: principles algorithms and applications.* Pearson Education India, 2001.

[60] N. Gößling and S. Doclo, "Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC).* IEEE, 2018, pp. 146–150.

[61] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 544–548.

[62] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 1018–1033, 2015.

[63] R. C. Felsheim, A. Brendel, P. A. Naylor, and W. Kellermann, "Head orientation estimation from multiple microphone arrays," in *2020 28th European Signal Processing Conference (EUSIPCO).* IEEE, 2021, pp. 491–495.

[64] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[65] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source sepa-

ration," in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*. IEEE, 2014, pp. 107–111.

[66] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.

[67] W. W. Hager, "Updating the inverse of a matrix," *SIAM review*, vol. 31, no. 2, pp. 221–239, 1989.

[68] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[69] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 216–220.

[70] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*. Springer, 2003, pp. 91–109.

[71] A. Falini, "A review on the selection criteria for the truncated svd in data science applications," *Journal of Computational Mathematics and Data Science*, p. 100064, 2022.

[72] S. Araki, F. Nesta, E. Vincent, Z. Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011):-audio

source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 414–422.

[73] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[74] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[75] H.-L. Wei, S. A. Billings, Y. Zhao, and L. Guo, "An adaptive wavelet neural network for spatio-temporal system identification," *Neural Networks*, vol. 23, no. 10, pp. 1286–1299, 2010.

[76] A. Wright, E.-P. Damskägg, L. Juvela, and V. Välimäki, "Real-time guitar amplifier emulation with deep learning," *Applied Sciences*, vol. 10, no. 3, p. 766, 2020.

[77] R. Abdelmalek, Z. Mnasri, and F. Benzarti, "Signal reconstruction based on the relationship between stft magnitude and phase spectra," in *Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT' 18), Vol. 2.* Springer, 2020, pp. 24–36.

[78] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[79] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity

using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[80] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity.* John Wiley & Sons, 2005.

[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[82] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory.* John Wiley & Sons, 2004.

List of Publications

Journal Papers

1. Shuming Luan, Y. Wakabayashi, T. Toda, "Unequally spaced sound field interpolation for rotation-robust beamforming," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 32, pp. 3185-3199, June 2024.

2. Shuming Luan, Y. Wakabayashi, T. Toda, "Generalized sound field interpolation for freely spaced microphone arrays in rotation-robust beamforming," Applied Acoustics, Vol. 236, Article 110706, pp. 1-15. June 2025.

3. Shuming Luan, L. Cheng, X. Sun, et al, "A two-stage deep learning based method for acoustic echo cancellation and speech dereverberation", Journal of Signal Processing, vol. 36(6), pp. 948-957, 2020.

4. S. Song, L. Cheng, Shuming Luan, et al, "An integrated multi-channel approach for joint noise reduction and dereverberation," Applied Acoustics, Vol. 171, Article 107526, Jan. 2021.

International Conferences

1. Shuming Luan, Y. Wakabayashi, T. Toda, "Modified sound field interpolation method for rotation-robust beamforming with unequally spaced circular microphone array," Proc. EUSIPCO, pp. 344-348, Belgrade, Serbia, Aug. 2022.

2. Shuming Luan, Y. Wakabayashi, T. Toda, "Sound field interpolation with unsupervised calibration for freely spaced circular microphone array in rotation-robust beamforming," Proc. EUSIPCO, pp. 21-25, Helsinki, Finland, Sep. 2023.

Awards

1. EUSIPCO 2022 Student Travel Grant

2. EUSIPCO 2023 Student Travel Grant