Multi-Microphone Signal Parameter Estimation in Various Acoustic Scenarios

Li, C.

DOI
10.4233/uuid:0d6f592e-42d8-40ab-8a9f-47ae7b5eb6fe

Publication date
2025

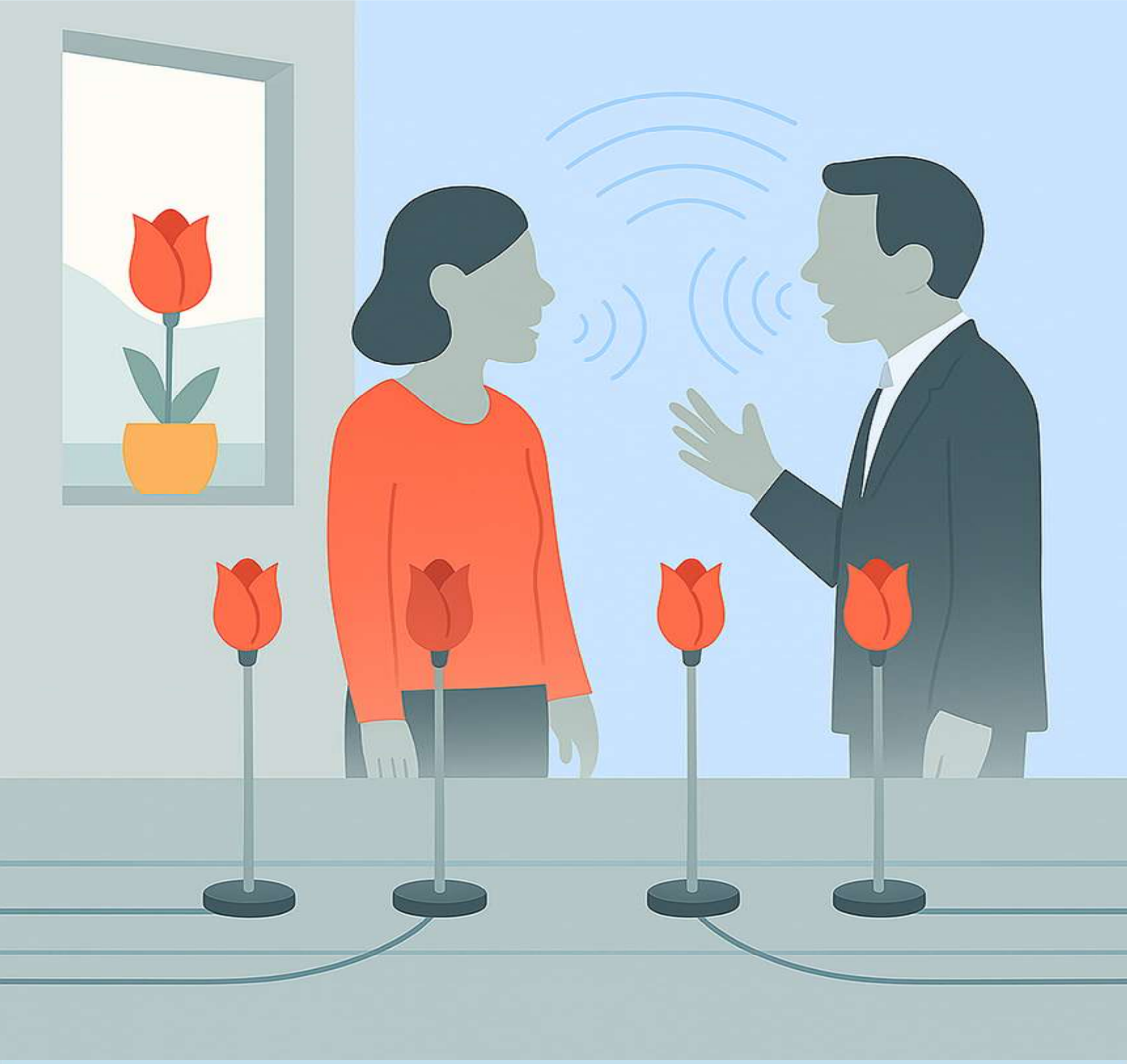
Document Version
Final published version

Citation (APA)
Li, C. (2025). *Multi-Microphone Signal Parameter Estimation in Various Acoustic Scenarios*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:0d6f592e-42d8-40ab-8a9f-47ae7b5eb6fe

Multi-Microphone Signal Parameter Estimation in Various Acoustic Scenarios

Low Complexity Approaches Utilizing Temporal Information

Changheng Li



Multi-Microphone Signal Parameter Estimation in Various Acoustic Scenarios

Low Complexity Approaches Utilizing Temporal Information

Multi-Microphone Signal Parameter Estimation in Various Acoustic Scenarios

Low Complexity Approaches Utilizing Temporal Information

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Thursday 18 September 2025 at 10:00 o'clock

by

Changheng LI

Master of Science in Information and Communication Engineering, University of Science and Technology of China, China

born in Henan, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. A.J. van der Veen,	Delft University of Technology, *promotor*
Dr. ir. R.C. Hendriks,	Delft University of Technology, *promotor*
Independent members:	
Prof. dr. ir. R. Heusdens,	Delft University of Technology
Prof. dr. ir. G.J.T. Leus,	Delft University of Technology
Prof. dr. Z.H. Tan,	Aalborg University, Denmark
Prof. dr. S. Gannot,	Bar-Ilan University, Israel
Prof. dr. ir. T. van Waterschoot,	Katholieke Universiteit Leuven, Belgium
Prof. dr. ir. M.J.T. Reinders	Delft University of Technology, reserve member

Keywords: Microphone array signal processing, Source separation, Dereverberation, Noise reduction, RTF estimation, PSD estimation.

Acknowledgement: This research was financially supported in part by the China Scholarship Council (CSC) under Grant 202006340031 and in part by the Signal Processing Systems (SPS) group, Delft University of Technology, Delft, The Netherlands.

ISBN 978-94-6518-068-7

An electronic copy of this dissertation is available at https://repository.tudelft.nl/.

To my parents Chuanxin and Shiling,
my brother Peiheng,
my love Karlie.

CONTENTS

SUMMARY

Many modern devices, such as mobile phones, hearing aids and (hands-free) acoustic human-machine interfaces are equipped with microphone arrays that can be used for various applications. These applications include source separation, audio quality enhancement, speech intelligibility improvement and source localization. In an ideal anechoic chamber, the signals received by ideal microphones are just attenuated and delayed version of the original sound. However, in practice, obstacles such as the floor, the ceiling and the surrounding walls will reflect the sound to the microphones. Also, the microphone itself will generate noise, distorting the recorded signals. Lastly, it is possible that multiple point sources are active simultaneously. When we consider one point source as the target signal, the other sources could be considered interfering signals. These distortions make it difficult to get access to the target signal. Therefore, spatial filtering is often applied to the microphone signals.

To achieve satisfying performance, these spatial filters typically need to be adaptive to the (changing) scene. Specifically, the filter coefficients depend on the acoustic-scene related parameters that model the microphone signals. These parameters, such as the relative transfer functions (RTFs) of the sources, the power spectral densities (PSDs) of the sources, the late reverberation and the ambient noise, are typically unknown in practice. Therefore, estimation of these parameters is crucial and thus the main focus of the dissertation. While it is relatively straightforward to estimate these parameters in less complex acoustic scenes, these algorithms are usually not applicable and not extendable to more complex acoustic scenes. Therefore, the complexity of the estimation methods needed depends on the complexity of the acoustic scene.

In Chapter 3, we consider the simplest acoustic scene in this dissertation, where there is only a single source in a reverberant and noiseless environment. The parameters that we aim to estimate are the RTFs, the PSDs of the target signal and the PSDs of the late reverberation. A joint estimator using a single time frame is first proposed, having a closed form. Then, a joint estimator using multiple time frames having the same RTF is proposed, where the solution for each iteration step is in closed form. The parameter estimation accuracy and the additional performance of noise reduction, speech quality and speech intelligibility of the proposed method are compared to various state-of-the-art reference methods. The proposed method reduces computational costs and improves performance as demonstrated by the experiments.

Next, we extend the noiseless signal model in Chapter 3 to the noisy model in Chapters 4 and 5. In Chapter 4, we focus on RTF estimation and propose an estimator that is robust to the late reverberation and noise PSD errors. This is achieved by using only off-diagonal elements of a simplified covariance matrix. The experiments demonstrate the effectiveness

of the proposed method. In Chapter 5, a joint estimator of the RTFs, the PSDs of the source, the PSDs of the late reverberation, and the PSDs of the ambient noise is proposed when using a single time frame as well as when using multiple time frames that share the same RTF.

Beyond the acoustic scene of a single point source, in Chapter 6 and 7, we consider the scenario of multiple point sources. In Chapter 6, we first consider the case where the environment is close to non-reverberant and noiseless. Under this assumption, we propose a method to estimate the RTFs. We obtain satisfying estimates by averaging covariance matrices for as many time frames as possible without suffering too much from model mismatch errors caused by distortion signals. This method is based on a comparison of several estimates from different averaged covariance matrices, which is somewhat heuristically motivated and not satisfying for reverberant and noisy environments. Therefore, in Chapter 7, we propose a robust method that works in reverberant and noisy environment and estimates not only the RTFs but also the PSDs of the sources and the late reverberation.

As in most of the works we have introduced, we use the prior information that several consecutive time frames share the same RTF. However, this is only possible if the source stays at the same position during these time frames. In Chapter 8, we therefore propose a method to adaptively segment signals into segments where the source is considered static. The proposed method is combined with the estimator we proposed in Chapter 3 for estimating the parameters of a single non-static source. It is shown in the experiments that with our proposed adaptive time segmentation, the estimation performance is improved over the use of a fixed time segmentation.

SAMENVATTING

Veel moderne apparaten, zoals mobiele telefoons, gehoorapparaten en (handsfree) akoestische mens-machine-interfaces zijn uitgerust met microfoonarrays die voor verschillende toepassingen kunnen worden gebruikt. Deze toepassingen omvatten bronscheiding, verbetering van de audiokwaliteit, verbetering van de spraakverstaanbaarheid en bronlokalisatie. In een ideale echovrije kamer zijn de signalen die door ideale microfoons worden ontvangen slechts een verzwakte en vertraagde versie van het oorspronkelijke geluid. In de praktijk zullen obstakels zoals de vloer, het plafond en de omringende muren het geluid echter naar de microfoons reflecteren. Ook zal de microfoon zelf ruis genereren, waardoor de opgenomen signalen worden vervormd. Ten slotte is het mogelijk dat meerdere puntbronnen tegelijkertijd actief zijn. Wanneer we één puntbron als het doelsignaal beschouwen, kunnen de andere bronnen als storende signalen worden beschouwd. Deze vervormingen maken het moeilijk om toegang te krijgen tot het doelsignaal. Daarom wordt er vaak ruimtelijke filtering toegepast op de microfoonsignalen.

Om bevredigende prestaties te bereiken, moeten deze ruimtelijke filters doorgaans adaptief zijn aan de (veranderende) scène. De filtercoëfficiënten zijn met name afhankelijk van de akoestische scènegerelateerde parameters die de microfoonsignalen modelleren. Deze parameters, zoals de relatieve overdrachtsfuncties (RTF's) van de bronnen, de vermogensspectraaldichtheden (PSD's) van de bronnen, de late nagalm en de omgevingsruis, zijn in de praktijk doorgaans onbekend. Daarom is het schatten van deze parameters cruciaal en dus de belangrijkste focus van het proefschrift. Hoewel het relatief eenvoudig is om deze parameters te schatten in minder complexe akoestische scènes, zijn deze algoritmen meestal niet toepasbaar en niet uitbreidbaar naar complexere akoestische scènes. Daarom hangt de complexiteit van de benodigde schattingsmethoden af van de complexiteit van de akoestische scène.

In hoofdstuk 3 beschouwen we de eenvoudigste akoestische scène in dit proefschrift, waarbij er slechts één bron is in een galmende en ruisloze omgeving. De parameters die we willen schatten, zijn de RTF's, de PSD's van het doelsignaal en de PSD's van de late nagalm. Eerst wordt een gezamenlijke schatter voorgesteld die gebruikmaakt van één tijdsbestek, met een gesloten vorm. Vervolgens wordt een gezamenlijke schatter voorgesteld die gebruikmaakt van meerdere tijdsbestekken met dezelfde RTF, waarbij de oplossing voor elke iteratiestap in gesloten vorm is. De nauwkeurigheid van de parameterschatting en de extra prestaties van ruisonderdrukking, spraakkwaliteit en spraakverstaanbaarheid van de voorgestelde methode worden vergeleken met verschillende state-of-the-art referentiemethoden. De voorgestelde methode verlaagt de rekenkosten en verbetert de prestaties, zoals blijkt uit de experimenten.

Vervolgens breiden we het ruisloze signaalmodel in hoofdstuk 3 uit naar het ruismodel

in hoofdstukken 4 en 5. In hoofdstuk 4 richten we ons op RTF-schatting en stellen we een schatter voor die robuust is tegen de late nagalm en ruis-PSD-fouten. Dit wordt bereikt door alleen off-diagonale elementen van een vereenvoudigde covariantiematrix te gebruiken. De experimenten tonen de effectiviteit van de voorgestelde methode aan. In Hoofdstuk 5 wordt een gezamenlijke schatter van de RTF's, de PSD's van de bron, de PSD's van de late nagalm, en de PSD's van het omgevingsgeluid voorgesteld bij gebruik van een enkel tijdsbestek en bij gebruik van meerdere tijdsbestekken die dezelfde RTF delen.

Naast de akoestische scène van een enkele puntbron, beschouwen we in Hoofdstuk 6 en 7 het scenario van meerdere puntbronnen. In Hoofdstuk 6 beschouwen we eerst het geval waarin de omgeving bijna niet-nagalmend en ruisloos is. Onder deze aanname stellen we een methode voor om de RTF's te schatten. We verkrijgen bevredigende schattingen door covariantiematrices te middelen voor zoveel mogelijk tijdsbestekken zonder al te veel last te hebben van modelmismatchfouten veroorzaakt door vervormingssignalen. Deze methode is gebaseerd op een vergelijking van verschillende schattingen van verschillende gemiddelde covariantiematrices, wat enigszins heuristisch gemotiveerd is en niet bevredigend voor nagalmende en ruisende omgevingen. Daarom stellen we in hoofdstuk 7 een robuuste methode voor die werkt in galmende en lawaaierige omgevingen en die niet alleen de RTF's schat, maar ook de PSD's van de bronnen en de late galm.

Zoals in de meeste werken die we hebben geïntroduceerd, gebruiken we de voorafgaande informatie dat verschillende opeenvolgende tijdsbestekken dezelfde RTF delen. Dit is echter alleen mogelijk als de bron gedurende deze tijdsbestekken op dezelfde positie blijft. In hoofdstuk 8 stellen we daarom een methode voor om signalen adaptief te segmenteren in segmenten waarbij de bron als statisch wordt beschouwd. De voorgestelde methode wordt gecombineerd met de schatter die we in hoofdstuk 3 hebben voorgesteld voor het schatten van de parameters van een enkele niet-statische bron. In de experimenten wordt aangetoond dat met onze voorgestelde adaptieve tijdssegmentatie de schattingsprestaties worden verbeterd ten opzichte van het gebruik van een vaste tijdssegmentatie.

1

INTRODUCTION

It is not knowledge, but the act of learning, not the possession of but the act of getting there, which grants the greatest enjoyment.

Carl Friedrich Gauss

1.1. MICROPHONE ARRAY

A microphone array is a set of microphones that can be used to simultaneously record sound at multiple locations aiming to improve the quality and intelligibility of a particular target signal. The application of microphone arrays is ubiquitous, being used in hearing aids, mobile phones, teleconferencing, hands-free acoustic human-machine interfaces and acoustic surveillance including security and monitoring. Depending on the intended applications, these devices are equipped with different types of microphone arrays.

For instance, the number of microphones can range from a few microphones to several hundreds. Using a larger number of microphones provides more information, typically leading to better performance, but it also increases the complexity and the cost. Smartphones, for example, typically have 2 to 4 microphones to enhance the audio recording quality. A conferencing system can have 8 microphones, while more sophisticated systems like acoustic cameras have tens or hundreds of microphones.

The microphone array geometry is another factor that can vary among the different microphone arrays. Common array geometries include linear, circular, spherical, and arrays with a random topology. The microphones are placed in a certain geometric structure for reasons such as the complexity, the device structure and the application needs. For instance, uniform linear arrays are commonly used since they can simplify the estimation problem. In smartphones, the microphones are usually arranged in a linear configuration along the top or bottom edges of the devices to fit within their slim profile. Some devices need a three-dimensional localization ability, hence the microphones cannot be placed within the same line. In this dissertation, the microphone geometry is not a limiting factor although in most experiments we will use linear arrays.

The last factor we want to introduce is microphone directivity, which refers to the sensitivity pattern of a microphone. The major types of microphones have a cardioid (uni-directional), bi-directional, omnidirectional or shotgun sensitivity pattern. Cardioid microphones have the greatest sensitivity at the front, only partially at the sides, and little at the back; Bi-directional microphones have the greatest sensitivity at both the front and the back; Omnidirectional microphones have equal sensitivity in all directions; Shotgun microphones have highly focused sensitivity in a single direction only. In this dissertation, we consider only omnidirectional microphones.

In daily life situations, the signals recorded by the microphones are inevitably distorted. The microphone signals are a mixture of the target signal and various distortions like reverberation, interfering sources and diffuse noise. Since the individual signal components in the microphone signals are unknown, we need to use some form of prior information to extract the target signal. For instance, the different components can be assumed uncorrelated across time. Moreover, since different components typically have a distinct spatial distribution, we can use microphone arrays to extract the target signal based on spatial information with significant improvement on the quality and intelligibility compared to using a single microphone.

1.2. ACOUSTIC DISTORTIONS

1.2.1. REVERBERATION

The microphone recordings can be distorted by various acoustic sources. One typical type of distortion is reverberation, which is caused by reflections in the room. When a sound source is transmitted inside a room, the sound signal arrives at the microphone through different paths, including not only a direct path but also numerous reflections via other paths caused by surrounding objects like the walls. These reflections are delayed versions of the direct path signal, whose magnitude attenuates as the delay time increases. For speech signals, the reflections are in general harmful to speech quality (SQ). The early reflections are however beneficial to the speech intelligibility (SI) [1], while the late reverberation is also harmful to SI. Therefore, reducing the reverberation (dereverberation) is an essential problem in the microphone array signal processing area. The reverberant level of a room can be characterized by reverberation time (RT), T_{60}.

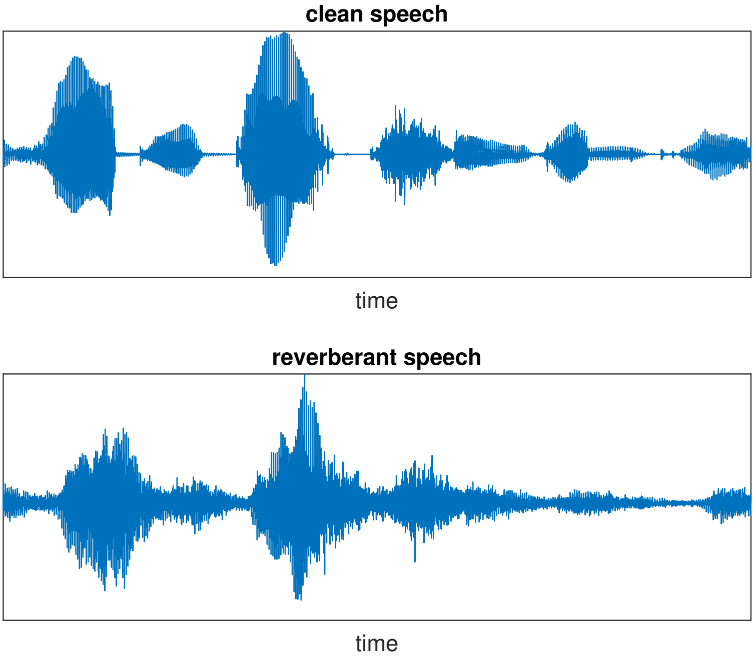


Figure 1.1: A clean speech signal and a reverberant speech signal. The reverberant signal has been scaled larger by a factor of 5 for better visibility.

The RT measures the time it takes for the sound pressure level to decay 60 dB after the sound has been stopped. For outdoor conditions with no reflection objects or an anechoic chamber with perfect absorption, the RT equals zero seconds. Within a car, the

RT is typically less than 0.2 s. In an office room, the RT is typically in the range of 0.2 s to 0.8 s, while for a class room, it has values between 0.4 s and 1 s. For larger rooms like a church or auditorium, it can be several seconds [2].

The RT mainly depends on the room volume and the materials that construct the room. It can also be affected by the room's shape and objects placed within the room. Therefore, for each room, in practice the most accurate way to measure the RT is using real on-site recording data. However, for rooms with regular shapes, an empirical equation can be used, which is known as Sabine's equation [3],

$$T_{60} = \frac{0.161V}{S\alpha}, \tag{1.1}$$

with V the room volume in m^3, S the total surface area and α the average absorption coefficient. Another equation was proposed by Carl F. Eyring of Bell Labs in 1930 [4] to better estimate the RT in "dead" rooms, which means small and very absorptive rooms. The well-known Eyring's equation is

$$T_{60} = -\frac{0.161V}{S\ln(1-\alpha)}. \tag{1.2}$$

Note that if $\alpha = 1$, i.e., the room is perfectly absorbing, the RT should be zero, as calculated by Eyring's equation.

1.2.2. NOISE AND INTERFERERS

Noise sources can be classified into point sources and the more diffuse noise sources. To obtain the signal of interest, we can use microphone arrays to remove the noise. The type of algorithm that is required to reduce the noise depends on the given prior information. In this dissertation, we consider both point interferers and microphone self-noise (as a more diffuse noise type). Notice that for many of the algorithms, the self noise could be replaced by other types of diffuse noise, as long as the coherence matrix of the diffuse noise is known.

In scenarios like meetings or crowded places, where multiple speakers (point sources) are active simultaneously, we have multiple point sources of which one speaker is considered as the target source and the remaining ones as the interferers. Typically, the interfering signals are non-stationary and their locations are unknown. Moreover, there are usually many reflections due to the reverberation. All these facts make it difficult to remove these distortions from the recorded microphone signals.

Other than the interfering noise, the microphone self-noise also needs to be considered. The self-noise exists as long as the microphone starts working, even in a silent environment. It originates from the addition of many noise components such as thermal noise, shot noise and air molecule Brownian motion. For instance, the random motion of electrons within the microphone's electronic components, such as resistors and transistors, can generate thermal noise. Also, shot noise is generated by the gate currents running through semiconductor junctions, such as in field-effect transistors (FETs) or bipolar junction transistors (BJTs). The sound pressure level (SPL) L_p can be used to measure the microphone self-noise level, expressed as

$$L_p = 20\log_{10}\frac{p}{p_0}\ \text{dB}, \tag{1.3}$$

where p is the root-mean-square sound pressure and p_0 is the reference sound pressure. The commonly used reference sound pressure is $p_0 = 20\ \mu\text{Pa}$. The threshold of human hearing has a SPL of 0 dB, which is roughly equal to the sound of a mosquito flying 3 meters away. While a normal conversation at 1 meter away is in the range of 40 to 60 dB SPL. For microphone self-noise, the SPL of expensive high-quality microphones is between 3 to 10 dB SPL. The majority of other less expensive microphones can have a SPL of 10 to 20 dB.

Note that in addition to interfering point sources and microphone self-noise, we will also consider the late reverberation (i.e., the combination of late reflections, say, sums after the early reflections) that can be modelled as a spatially homogeneous sound field characterized by a time-invariant spatial coherence matrix with a time-varying PSD. Note that, by using a different spatial coherence matrix, we can also model other noise signals such as wind noise.

1.3. SPATIAL FILTERING

In the previous section, we have introduced the different signal components in a microphone recording. Since the spatial locations of the microphones in an array are different, each microphone signal consists of different combinations of these contributions as they result from different spatial propagation modelled by different room impulse responses. We can combine these spatial observations (in a linear/ non-linear way) to extract the signal of interest.

To extract the target signal from the reverberant and noisy microphone signals, spatial filters have been widely used. Although non-linear filtering methods have shown performance gain over linear filtering methods [5], the commonly used filtering methods are linear due to their simplicity and low complexity. These linear spatial filters are also known as beamformers, which refers to the fact that under certain condition they form a beam in the direction of arrival (DOA). Initially, beamformers were formulated based on pure geometric information of the scene (sources and microphones' location/direction), which is translated into the DOAs. More generally, they can also be formulated using source-to-microphone acoustic transfer functions (ATFs).

When the DOA of the target source and the microphone positions are known a priori, fixed beamformers (FBFs) can be used to preserve the sources coming from a given direction, while eliminating sources that come from all other directions (including noise sources and reflections of the target source). The delay and sum beamformer (DS) is a commonly used FBF, which is also known as Bartlett beamformer [6]. It averages microphone signals after applying delay compensation in order to preserve the target, since the direct path signals for different microphones are attenuated and delayed versions of each other. When the DOA is estimated, the beamformer is called semi-fixed.

Above mentioned filters are data-independent. For the sake of higher performance, data-dependent spatial filters are widely used. A well-known beamformer in this category is the minimum variance distortionless response beamformer (MVDR), which is also known as Capon's beamformer [7]. It is designed to minimize the output noise power while preserving the target signal after filtering. Unlike other array signals coming from a single direct path, microphone array signals contain reflections. Therefore, the MVDR for microphone array signals depends on acoustic transfer functions (ATFs) or relative transfer functions (RTFs) instead of the steering vector modeled by DOA. The ATF describes the spatial information from the source to the microphones. The RTF describes the relative spatial information between the microphones for a given source. When the true ATF or RTF is given, the MVDR can preserve the target signal perfectly. However, the ATF or RTF is unknown in practice and hence needs to be estimated with high accuracy. An extension of the MVDR filter known as the linearly-constrained minimum variance (LCMV) filter is also widely used.. The LCMV uses multiple linear equality constraints to introduce more control on the filter. For instance, the LCMV can be used for spatial cue preservation in a binaural setting or to cancel interferers with nulling constraints.

In complex scenarios with multiple sound sources, the MVDR may not provide sufficient noise reduction, since it is designed to keep the target source undistorted. To address this limitation, a post filter can be added to adjust the trade-off between noise reduction and signal distortion. When choosing the single-channel Wiener filter as the post filter, the MVDR with this post filter forms the well-known multichannel Wiener filter (MWF) [8]. The MWF is designed to minimize the mean square error between the target signal and the filtered signal. Some variants of the MWF such as the SDW-MWF [9], were proposed to better control the trade-off between noise reduction and signal distortion. Both the MVDR and the MWF can be seen as special cases of the SDW-MWF, where the MVDR obtains the lowest signal distortion and the MWF obtains the best noise reduction performance. When using the MWF, the post-filter needs information of the power spectral densities (PSDs) of the target signal and the noise. These PSDs are also unknown in practice and hence need to be estimated.

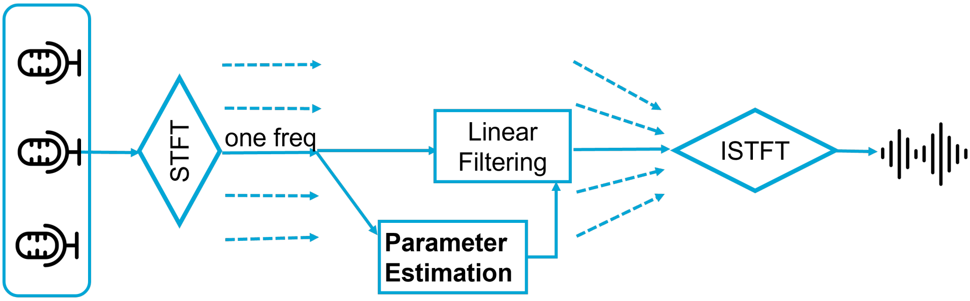


Figure 1.2: Microphone array signal processing diagram.

The typical flowchart of microphone array signal processing based on frequency domain linear filtering is shown in Fig. 1.2. The time domain signals recorded by microphones are transferred to the frequency domain, using the short-time Fourier transform (STFT). For each frequency, we estimate the parameters of interest, given multiple noisy microphone signals. Then we can use linear filtering to extract the target signal, based on the parameters we estimated. Collecting the reconstructed signals from all frequencies, we use the inverse short-time Fourier transform (ISTFT) to obtain the target signal in the time domain. The parameter estimation block in this process is the focus of this thesis.

1.4. RESEARCH QUESTIONS

To extract a target signal from a noisy and reverberant environment, we can thus exploit techniques such as spatial filtering when we have access to multiple microphones. To use such techniques, we need the signal-dependent or acoustic scene-dependent parameters such as the relative transfer functions (RTFs) and the power spectral densities (PSDs) of the point sources. In practice, these parameters are unknown and need to be estimated using the signals recorded by the microphones. However, depending on the scenarios, the estimation problem can be very challenging and sometimes even ill-posed as we can have many unknown parameters, few microphones and non-stationary sources.

Many methods have been proposed in recent years to estimate the parameters, e.g., [2], [10]–[20]. Typically, the unknown parameters include the sources' RTFs, the sources' PSDs, the reverberation PSDs and the noise PSDs. Many estimation methods only consider the estimation of a subset of them by assuming that the remaining parameters are known. In [14], [17], for example, the signal received as the direct path is considered as the target signal and therefore the RTFs can be modelled by the DOA of the source position and the microphone array geometry. By further assuming the DOA is known, the RTFs are considered known. However, not only the direct sound, but also the early reflections can be beneficial to speech intelligibility [1]. They are therefore often considered as part of the target sound, making DOA modelling not suitable for RTFs. It is therefore very challenging to estimate the RTFs. In most works of this dissertation, we will focus on the joint estimation of these typically unknown acoustic parameters.

Meanwhile, most existing methods [10], [11], [13]–[15], [17]–[19] use only a single time frame (related to a single covariance matrix) to estimate the parameters. However, time frames (especially adjacent ones) often share some common information like the RTFs. This could be exploited to obtain better estimates [16].

In [16], the task of joint estimation of the unknown parameters using multiple time frames was proposed, leading to quite good performance. However, the method proposed in [16] suffers from a rather high computational cost, which hinders real world applications of this method. Therefore, there is a need for methods that can achieve the same state-of-the-art estimation performance, while having low complexity.

The main contents of this dissertation addresses the problem of estimating these signal model parameters from observed microphone array signals, aiming at providing low complexity based approaches to estimate the parameters jointly. This problem is still very general as the parameters of interest can vary depending on the scenarios considered. In the following, we will present our specific research questions for various acoustic scenarios of increasing complexity: from a single point source in a noiseless reverberant scenario, to a highly complex scenario with multiple sources, reverberation and noise.

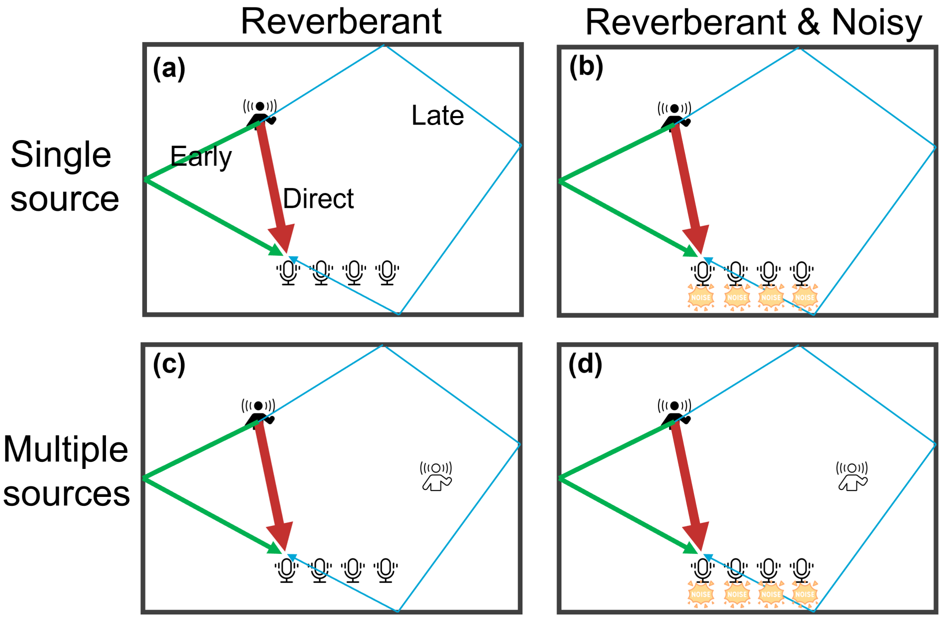


Figure 1.3: Illustration of different acoustic acenarios.

The first scenario we consider is the simple case of a single point source. The research question is then:

> ***(RQ 1): How to estimate the microphone array signal parameters for a single source?***

Depending on the environment and what parameters we aim to estimate, we subdivide RQ 1 into the following three questions.

First, as introduced in Section 1.2.2, when using high-quality microphones or when the sound sources have a high SPL, the signal-to-noise ratio (SNR) of the microphone signals can be about 50 dB. In this case, a noiseless signal model can be assumed. The parameters of interest are then composed of the RTF and the PSDs of the source and the late reverberation, which leads to the question:

> *(RQ 1.1): How to estimate the microphone array signal parameters for a single source in a reverberant but noiseless environment?*

For this scenario as illustrated in Fig. 1.3 (a), we will present the signal model in

Fig. 2.2 (a) and propose a maximum likelihood estimator in Chapter 3.

A natural next step is to consider both a reverberant and noisy environment as illustrated in Fig. 1.3 (b). In such a case, for a single source, we first consider an estimator for the RTF specifically, leading to research question RQ 1.2:

> *(RQ 1.2): How to estimate the RTF for a single source in a reverberant and noisy scenario?*

Then, we continue to investigate possible estimators for a joint estimation of the RTF and the PSDs of the source, the late reverberation and the noise:

> *(RQ 1.3): How to estimate the microphone array signal parameters jointly for a single source reverberant and noisy scenario?*

For both research questions 1.2 and 1.3, we will present the signal model in Fig. 2.2 (b). For RQ 1.2, we will propose a RTF estimator in Chapter 4. For RQ 1.3, we will propose a joint estimator for the RTF of the source and the PSDs of the source, the late reverberation and the noise in Chapter 5.

> ***(RQ 2): How to estimate the microphone array signal parameters for multiple sources?***

For the second research question, we decide to face the more challenging problem of multiple sources. The problem involving multiple sources is closely related to other research areas that have attracted much interest, such as factor analysis [21] and blind source separation [22], [23]. Therefore, we can get inspiration from methods proposed in these research fields to solve our estimation problems.

For this multi-source scenario, we start by assuming the environment is nearly non-reverberant and noiseless. We focus on the estimation of the RTFs of the multiple sources. The acoustic scenario is illustrated in Fig. 1.3 (c).

> *(RQ 2.1): How to estimate the RTFs for multiple non-reverberant sources?*

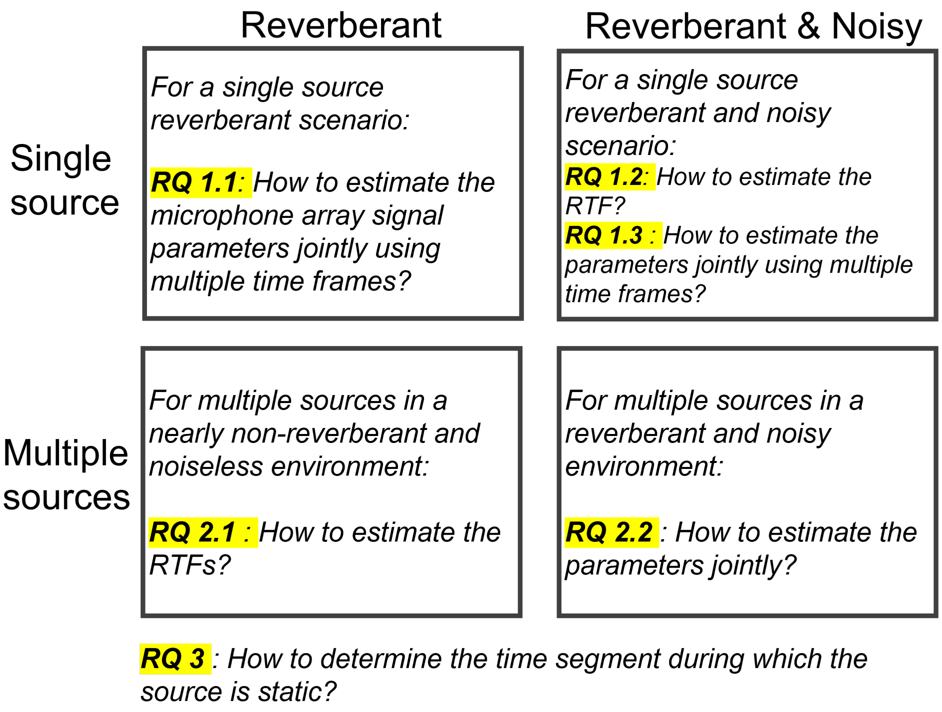
Non-reverberant implies that there are only early and direct reflections, but no late reflections. The signal model related to this question will be presented in Fig. 2.2 (c) and the estimator will be proposed in Chapter 6.

In the next step, we consider a noisy and reverberant environment illustrated by Fig. 1.3 (d), where the noise component is assumed to be stationary (i.e., with constant PSD), leading to the following question:

> *(RQ 2.2): How to estimate the microphone array signal parameters for multiple sources in a reverberant and noisy environment?*

We will show the corresponding signal model in Fig. 2.2 (d) and propose a joint estimator for the RTFs of the sources and the PSDs of the sources and the late reverberation in Chapter 7.

When solving some of the estimation problems listed above, we assume that the source does not move for a duration (longer than the stationary period of speech signals).



	Reverberant	Reverberant & Noisy
Single source	*For a single source reverberant scenario:* ***RQ 1.1**: How to estimate the microphone array signal parameters jointly using multiple time frames?*	*For a single source reverberant and noisy scenario:* ***RQ 1.2**: How to estimate the RTF?* ***RQ 1.3** : How to estimate the parameters jointly using multiple time frames?*
Multiple sources	*For multiple sources in a nearly non-reverberant and noiseless environment:* ***RQ 2.1** : How to estimate the RTFs?*	*For multiple sources in a reverberant and noisy environment:* ***RQ 2.2** : How to estimate the parameters jointly?*

***RQ 3** : How to determine the time segment during which the source is static?*

Figure 1.4: Research questions summary of the dissertation.

With this assumption, we can make use of the prior information that the RTF is constant for this duration and obtain improved estimation performance (mainly on the RTF estimation). In practice, this duration can vary from zero to the whole signal duration if the source keeps moving or is always static, respectively. However, for scenarios in between these two, i.e., the source is static (not moving) only for a short unknown duration, it is crucial to consider the question:

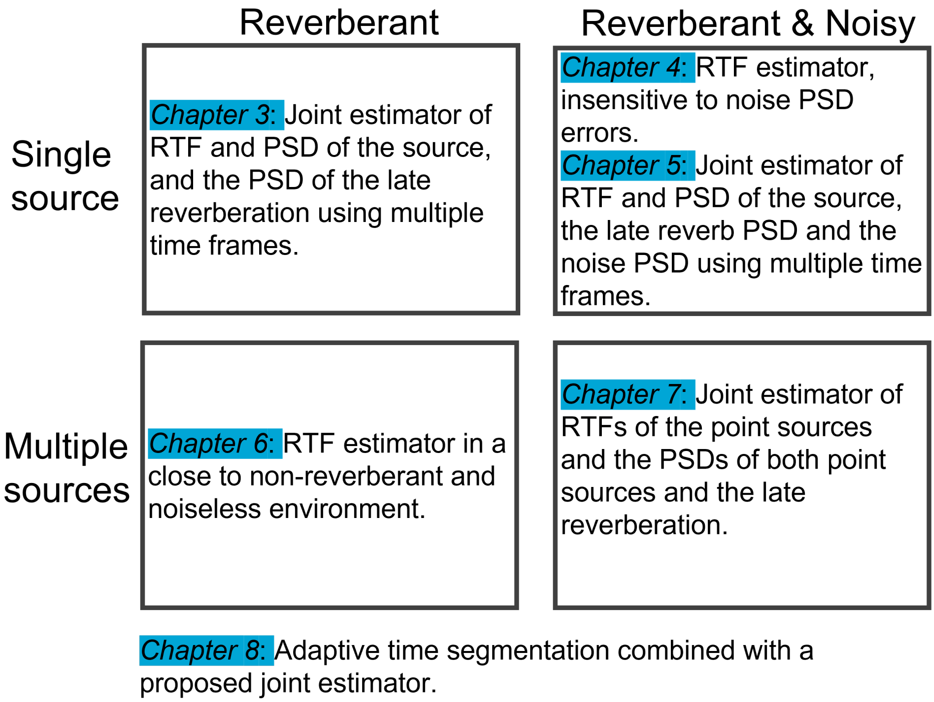
> ***(RQ 3): How to determine the time segment during which the source is static?***

We will propose a method to answer this question in Chapter 8. This method will be combined with the estimator proposed in Chapter 3 for the single source reverberant scenario.

A summary of all research questions is presented in Fig. 1.4.

1.5. DISSERTATION CONTRIBUTIONS AND OUTLINE

In this section, we describe the dissertation outline and summarize the contribution of each chapter. A contribution overview is presented in Fig. 1.5.



	Reverberant	Reverberant & Noisy
Single source	*Chapter 3*: Joint estimator of RTF and PSD of the source, and the PSD of the late reverberation using multiple time frames.	*Chapter 4*: RTF estimator, insensitive to noise PSD errors. *Chapter 5*: Joint estimator of RTF and PSD of the source, the late reverb PSD and the noise PSD using multiple time frames.
Multiple sources	*Chapter 6*: RTF estimator in a close to non-reverberant and noiseless environment.	*Chapter 7*: Joint estimator of RTFs of the point sources and the PSDs of both point sources and the late reverberation.

Chapter 8: Adaptive time segmentation combined with a proposed joint estimator.

Figure 1.5: Contribution overview of the dissertation.

CHAPTER 2

The background theory for microphone array signal processing are provided in Chapter 2. We first introduce the mathematical signal model and its limitations in the time domain. Then, we introduce the signal model in the frequency domain and explain the reasons behind the assumptions made for different acoustic components. In addition, we show the general framework of linear filtering for source separation, dereverberation and noise reduction. We review some classic filtering techniques and the corresponding parameters. At last, we give definitions to different segment durations of the signals that we will exploit and explain the practical meaning behind these definitions.

CHAPTER 3

When using high-quality microphones inside a room where the speech sources energy is sufficiently large compared to the noise energy (e.g., the speaker standing not far from the microphone array), we can assume a noiseless case. That is the scenario of a single reverberant source we consider in Chapter 3. The parameters need to be estimated include the relative transfer functions (RTFs) from source to microphones, source power spectral densities (PSDs) and PSDs of the late reverberation. We first consider the joint estimation using a single time frame and find the solution that has a closed form. In this joint estimation case, the solution of the RTF estimator turns out to be simple covariance whitening and the solution of the PSDs can also be esaily obtained using the estimated RTF. Similar to [16], we assume the RTF changes slower than the PSDs and multiple consecutive time frames share the same RTF. In this case, we expect performance gain by using multiple time frames to estimate the parameters jointly. Based on the maximum likelihood cost function, we cannot obtain a closed form solution for the multi-time frame case like the single time frame case. We therefore propose to solve the problem in an iterative fashion. In each iteration step, the estimator has a closed form solution. Therefore, the proposed method has a much lower computational complexity compared to the method from [16].

CHAPTER 4

In practice, microphones are noisy with a noise level that depends on the microphone quality. Depending on the microphone-source distance, this influences the SNR. In anyway, we cannot assume a noiseless scenario in practice as we did in Chapter 3. However, we can still assume the noise to be stationary, which implies an estimate of the noise PSD is easily available, e.g. by means of a voice activity detection (VAD). With the estimated noise PSD, we can subtract the noise component from the noisy covariance matrix and use methods that assume noiseless scenarios. The problem is that such a methodology relies on the accuracy of the noise PSD estimate. In other words, such methods are sensitive to the noise PSD estimate. To break the limitations of such methods, we propose a method that does not need noise PSD information and can estimate the RTF directly from the noisy covariance matrix.

CHAPTER 5

For a single source in a reverberant and noisy environment, the problem of joint estimation of the RTF of the target source and the PSDs of the three components (target source, late reverberation and noise) becomes more complex than the single-source, reverberant but noiseless scenario, which we have considered in Chapter 3. To estimate all these parameters in a joint fashion while maintaining low complexity, we consider the least square cost function instead of the maximum likelihood cost function.

In this work, we do not only consider the estimation using a single time frame, but also consider the case of using multiple time frames that share the same RTF. For the single time frame case, these is one existing reference work aiming at solving the joint estimation problem using the least square cost function [15]. Based on [15], we propose an improved algorithm and extend it to the multiple time frames case. Note that, the proposed algorithms are all iterative. Also, although we assumed the PSDs to be positive during the iterations, the estimates we get can become negative. To solve this issue, we propose several ways to both upper bound and lower bound the PSD estimates only for cases where such bad estimates happen.

CHAPTER 6

After solving the joint estimation problem for a single source, we can move on to the next more complex problem: mutliple sources. In Chapter 6, we start from the simplest case that the environment is little reverberant and close to noiseless.

As we mentioned before, the multi-source scenario is related to the blind source separation problem. In this work, we show that one classic blind source separation method, JOINT (also called SOBI), can be modified to estimate the RTFs of the sources. Furthermore, we propose a more robust method that proposes various factorizations in the first step of JOINT and get several estimated RTFs. Among these RTFs, we select a best estimate regarding to the minimum cost function value. Note that this approach of trying various initializations increases the computational complexity, even though the overall complexity is still much better than the state of the art method (refered to as SCFA). In addition, this approach appears to be a bit heuristic. In the next chapter, we therefore find a way to calculate the first step in an optimal way, which makes the algorithm faster and more robust, even in a reverberant and noisy environment.

CHAPTER 7

The last and most complex estimation problem in this dissertation is to estimate the parameters of interest for multiple sources in a reverberant and noisy environment. The parameters include the RTFs of the point sources and the PSDs of both point sources and the late reverberation. To solve this problem at low complexity, we estimate the parameters subsequently, where we first find a late reverberation PSD estimator and then propose a joint estimator for the sources' RTFs and PSDs. Note that we still only consider stationary noise component such as the microphone self-noise. Since the microphone self-noise can be assumed stationary, we can estimate the noise covariance matrix when all sources are absent. With the estimated noise covariance matrix, we can subtract it from the noisy covariance matrix to get an estimate of the covariance matrix for the remaining components. In consequence, we can use the method based on the noiseless signal model to solve the estimation problem. However, since we almost surely get estimation errors when subtracting estimated covariance matrices, the method based on a noiseless signal model might not work. For instance, this could lead to a negative estimate of the late reverberation PSD. Therefore, we propose a more robust estimation scheme for the late reverberation PSD estimator. Then, for the RTFs and PSDs of the point sources, we can subtract the covariance matrix of the late reverberation and noise from the noisy covariance matrix to get an improved estimate of the source covariance matrix. Given a set of the source covariance matrices (corresponding to multiple time frames sharing the same RTFs), we show how we can modify an existing blind source separation method called SOBI to solve the estimation problem. Furthermore, we can use a linear combination of these source covariance matrices at the first step of the modified algorithm. Then, by analysing the variances of the error matrix of the sample covariance matrix, we propose an optimal linear combination where the coefficients can be calculated from the estimated parameters: the late reverberation PSD and the noise PSD.

CHAPTER 8

In previous chapters, we either assume the sources are static or assume that we know which time frames share the same RTF. However, in practice, the sources can be moving and we do not know whether any two time frames share the same RTF. Therefore, in this chapter, we present an algorithm to obtain an optimal adaptive time segmentation and combine this with the joint maximum likelihood estimator (JMLE) discussed in Chapter 3 for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment.

CHAPTER 9

In this last chapter, we give a summary of all the works in this dissertation. Other than the conclusions, we discuss some possible directions that we consider interesting and valuable for future investigation. In addition, we also propose some open questions in these research directions and share our thoughts and suggestions on how to address these questions in future research.

1.6. LIST OF PUBLICATIONS

In this section, we present all papers submitted and published in the PhD duration.

JOURNAL PAPERS

1. **C. Li** and R. C. Hendriks, “Multimicrophone Signal Parameter Estimation in A Multi-Source Noisy Reverberant Scenario”. submitted to *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 678-692, 2025.
2. **C. Li** and R. C. Hendriks, "Alternating Least-Squares-Based Microphone Array Parameter Estimation for a Single-Source Reverberant and Noisy Acoustic Scenario," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3922-3934, 2023.
3. **C. Li**, J. Martinez and R. C. Hendriks, "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario," in

IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 695-705, 2023.

CONFERENCE PROCEEDINGS

1. **C. Li** and R. C. Hendriks, "Adaptive Time Segmentation for Improved Signal Model Parameter Estimation for a Single-Source Scenario," 57th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2023.
2. **C. Li** and R. C. Hendriks, "Noise PSD Insensitive RTF Estimation in a Reverberant and Noisy Environment," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023.
3. **C. Li**, J. Martinez and R. C. Hendriks, "Low Complex Accurate Multi-Source RTF Estimation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022.

OTHER CONTRIBUTIONS

These contributions will not be included in the dissertation.

1. L. Luo, **C. Li** and Q. Li,"Deterministic Algorithms to Solve the (n,k)-Complete Hidden Subset Sum Problem", submitted to *Theoretical Computer Science*.
2. Z. Li, **C. Li** and R. Rajan, "'On the Stability of Consensus Control under Rotational Ambiguities ," in *IEEE Control Systems Letters*, vol. 8, pp. 3273-3278, 2024.
3. J. Zhang and **C. Li**, "Quantization-Aware Binaural MWF Based Noise Reduction Incorporating External Wireless Devices," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3118-3131, 2021.

REFERENCES

[1] J. S. Bradley, H. Sato, and M. Picard, “On the importance of early reflections for speech in rooms”, *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

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

[3] W. C. Sabine, *Collected Papers on Acoustics*. Cambridge, MA: Harvard University Press, 1922.

[4] C. F. Eyring, “Reverberation time in "dead" rooms”, *J. Acoust. Soc. Amer.*, vol. 1, pp. 217–241, 1930.

[5] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 563–575, 2023.

[6] M. S. Bartlett, “Periodogram analysis and continuous spectra”, *Biometrika*, vol. 37, no. 1/2, pp. 1–16, 1950, ISSN: 00063444, 14643510.

[7] T. Marzetta, “A new interpretation of capon’s maximum likelihood method of frequency-wavenumber spectral estimation”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 2, pp. 445–449, 1983.

[8] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[9] J. Benesty, S. Makino, J. Chen, S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Speech distortion weighted multichannel wiener filtering techniques for noise reduction”, *Speech enhancement*, pp. 199–228, 2005.

[10] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[11] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[12] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[13] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[14] I. Kodrasi and S. Doclo, "Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[15] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[16] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[17] Y. Laufer and S. Gannot, "Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[18] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[19] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, 2019.

[20] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square Root-Based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 755–769, 2020.

[21] D. N. Lawley and A. E. Maxwell, "Factor analysis as a statistical method", *J. R. Stat. Soc.*, vol. 12, no. 3, pp. 209–229, 1962.

[22] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[23] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound", *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

2 BACKGROUND

All models are wrong, but some are useful.

George Edward Pelham Box

This chapter aims to provide the necessary and sufficient background knowledge for the main contents of this dissertation. The introduction of the microphone signals and their processing techniques will be continued, mainly using mathematical formulations instead of high-level descriptions as in the previous chapter.

This chapter is organized as follows. In Section 2.1, we introduce the signal model in the time domain. Time-domain microphone signals are often processed in the frequency domain by applying the short-time Fourier transform (STFT) procedure, as we will also do in this dissertation. Therefore, in Section 2.2, we express the signal model in the frequency domain.

2.1. TIME DOMAIN SIGNAL MODEL

We assume that R point sources $s_r(t)$ for $r = 1, \cdots, R$ are active inside a room and generate sound waves. These sound waves travel through the air and arrive at each microphone after being attenuated, reflected and delayed. The received waveform not only includes a direct path from the original position to the microphone position, but also an infinite number of reflections, due to the existence of the surrounding objects such as walls. These reflections and the direct path form only one part of the received microphone signals, since there exists also ambient noise originating from other point sources or more diffuse sources. Typically, this combination of signals is assumed to be additive and sampled in time. We consider an array of M microphones and move from the physical domain to the discrete time signal domain. Let the discrete microphone signals be denoted by $y_m(t)$ for $m = 1, \cdots, M$ with m the microphone index and where t now represents the discrete time index.

At the m-th microphone, the summation of the reflections for the r-th source $s_r(t)$ can be modelled as the convolution between the source $s_r(t)$ and the time-varying acoustic impulse response (AIR) $a_{r,m}(t, \tau)$, where τ is the discrete delay time index. That is

$$\sum_{\tau=0}^{\infty} a_{r,m}(t,\tau)\, s_r(t-\tau). \tag{2.1}$$

The AIRs can be seen as the microphone recording of the impulse that was emitted at the source position. Therefore, it can be time-varying when the source, the microphones or the other objects move in the room. When they are all static, we can consider the

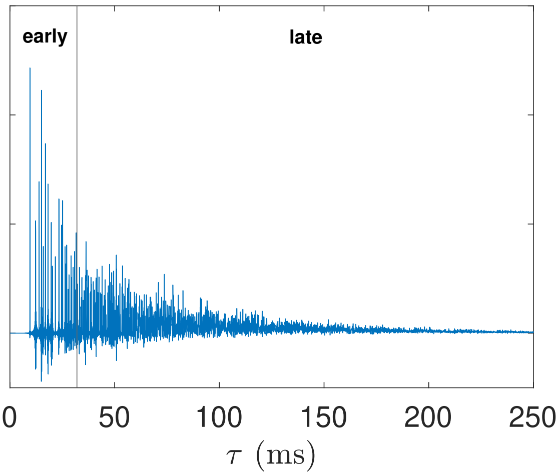


Figure 2.1: Acoustic impulse response split into two parts: the early part and the late part. This AIR is simulated using the Image method [1] with the reverberation time 0.4 s and source-to-microphone distance 3.27 m.

AIRs to be time-invariant. In this dissertation, the AIRs are assumed either always static (i.e., $a_{r,m}(\tau)$) or static for a short duration.

Fig. 2.1 shows an AIR generated using the Image method [1] with the reverberation time of 0.4 s and source-to-microphone distance of 3.27 m. Note that it has both positive and negative values. We used a vertical line at $\tau = 32$ ms to divide the AIR into two parts. The left part is called the early part, which includes the direct path and early reflections. The largest value corresponds to the direct path and the first few sparse peaks correspond to the first few reflections. The right part corresponds to the late reflections or late reverberation. It can be seen that the late reflections are densely distributed and cannot be distinguished from each other. Also, their magnitude decays nearly exponentially.

We use $v_m(t)$ to denote the noise component at the m-th microphone. Note that in most works of this dissertation, we only consider the microphone self-noise. The noisy and reverberant signal y_m thus can be modelled as

$$y_m(t) = \sum_{r=1}^{R} \left\{ \underbrace{\sum_{\tau=0}^{\tau_e} a_{r,m}(t,\tau) s_r(t-\tau)}_{x_{r,m}(t)} + \underbrace{\sum_{\tau=\tau_e+1}^{\infty} a_{r,m}(t,\tau) s_r(t-\tau)}_{l_{r,m}(t)} \right\} + v_m(t), \tag{2.2}$$

where $x_{r,m}(t)$ denotes the early reflections and $l_{r,m}(t)$ denotes the late reflections. Note that the time domain signal model Eq. (2.2) consists of an infinite number of parameters due to the infinite impulse response. In practice, a finite impulse response is used by considering $l_{r,m}(t) = \sum_{\tau=\tau_e+1}^{\tau_l} a_{r,m}(t,\tau) s_r(t-\tau)$, where τ_l generally takes value of several hundreds. Note that the large number of parameters leads to high computational cost [2]. Therefore, the frequency domain signal model is often used, which we will introduce in Section 2.2.

2.2. STFT DOMAIN SIGNAL MODEL

Since the Fourier transform of a convolution of two signals is equivalent to a multiplication of the Fourier transforms of the two signals and the Fourier transform can efficiently be implemented using an FFT, we can significantly reduce the complexity by using frequency domain processing instead of time domain convolution. However, the microphone signals (in particular the reverberant speech signals) are non-stationary. A full-length Fourier transform of the microphone signal is therefore not appropriate. However, it is typically assumed that the vocal tract has a similar shape for time intervals of about 20-30 ms [3]. Speech signals are therefore often considered to be short-time stationary for a short duration (20-30 ms) referred to as a time frame. Therefore, the short-time Fourier transform (STFT) is commonly used to process the microphone signals.

The STFT coefficients of a signal $s(t)$ are given by

$$s(l,k) = \sum_t s(t)\,\psi(t-lT)\,e^{-j\frac{2\pi}{K}k(t-lT)}, \tag{2.3}$$

where $\psi(t)$ denotes an analysis window of length K, k is the frequency bin index and T represents the time shift of the window. Given a specific analysis window, and given that we can find a synthesis window $\bar{\psi}(t)$ such that

$$\sum_l \psi(t-lT)\,\bar{\psi}(t-lT) = \frac{1}{K}, \tag{2.4}$$

for all t, then we can resconstruct the time domain signal $s(t)$ perfectly using

$$s(t) = \sum_l \sum_{k=0}^{K-1} s(l,k)\,\bar{\psi}(t-lT)\,e^{j\frac{2\pi}{K}k(t-lT)}. \tag{2.5}$$

Note that, when $T \leq K$, there might be multiple analysis windows satisfying the condition in Eq. (2.4) for a given synthesis window. Considering the STFT coefficients of the early reflections $x(t) = \sum_{\tau=0}^{\tau_e} a(\tau)\,s(t-\tau)$, we should note that strictly speaking, they are not just the multiplication of the STFT coefficients of the AIR and the source signal $s(t)$, but the following inter-frame and inter-band convolution between the AIR and the source signal [4] due to the window function applied in the STFT procedure, that is,

$$x(l,k) = \sum_{k'=0}^{K-1} \sum_{l'} s\left(l-l',k'\right) a\left(l',k,k'\right), \tag{2.6}$$

where $a(l',k,k')$ is the impulse response in the time-frequency domain given by [4]

$$a\left(l',k,k'\right) = \sum_{t'} \sum_t a(t)\,\bar{\psi}\left(t'-t+l'T\right) e^{j\frac{2\pi}{K}k'\left(t'-t+l'T\right)}\,\psi\left(t'\right) e^{-j\frac{2\pi}{K}kt'}. \tag{2.7}$$

This representation is an accurate model but involves multiple filtering steps. This can be simplified by only considering the subband filtering ($k'=k$) [5], i.e.,

$$x(l,k) \approx \sum_{l'} s\left(l-l',k\right) a\left(l',k\right). \tag{2.8}$$

The above approximated signal model is known as the Convolutive transfer function (CTF) model, which has been considered in some works such as [6], [7]. Note that the CTF model usually needs less taps compared to the time domain convolution, but it has not often been used in practice, since it still relies on quite a number of unknown parameters to be estimated [8].

If the frame-length is sufficiently large and the analysis window $\psi(t)$ is smooth compared to the AIR, we can approximate $\psi(t-\tau)\,a(\tau)$ by $\psi(t)\,a(\tau)$. The convolution in Eq. (2.6) can be approximated by a multiplication, i.e.,

$$x(l,k) \approx s(l,k)\,a(l,k), \tag{2.9}$$

which is known as the multiplicative transfer function (MTF) model or the narrowband approximation. The MTF model has been widely assumed in many works such as [9]–[11]. Despite its simplicity and widely usage, the MTF model has other problems, of which the most notable one is the gain ambiguity for each single source and the permutation ambiguity for multiple sources.

The gain ambiguity comes from the fact that we will have the same signal $x(l,k)$ even if we use a scaled $s(l,k)$, say $cs(l,k)$ and the inverse scaled $a(l,k)$, say $\frac{1}{c}a(l,k)$ for any constant $c \neq 0$. This gain ambiguity has been avoided by considering the relative transfer function (RTF) between microphones for each source. In this case, the target signal is no longer the sound source generated at the source location but the direct and early reflections at a reference microphone. For instance, when selecting the first microphone as the reference microphone, the RTF for the first microphone is 1. For the other microphones, the RTF is the ratio between the ATF of that microphone and the ATF of the first microphone. The choice of the reference microphone can be determined by, for instance, aiming at improving the SNRs [12]. In this dissertation, we always select the first microphone as the reference.

When considering multiple sources, the STFT domain signal model of the early reflections at the m-th microphone then becomes

$$x_m = \sum_r s_r(l,k)\, a_{r,m}(l,k). \tag{2.10}$$

We also have to deal with the permutation ambiguity, which means that we do not know which of the STFT coefficients $s_r(l,k)$ (with $r = 1,\cdots,R$) across frequency belong to the same source r. This problem is beyond the scope of this dissertation and methods on this topic, to name a few, were investigated in [13], [14]. In this dissertation, we assume that the permutation has been perfectly aligned such that for all $s_r(l,k)$ it is known for every frequency bin to which source r it belongs.

Altogether, in vector form, the STFT coefficients of the microphone signals can be represented by

$$\mathbf{y}(l,k) = \underbrace{\sum_{r=1}^{R} \mathbf{a}_r(l,k)\, s_r(l,k)}_{\mathbf{x}(l,k)} + \mathbf{d}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1}\ , \tag{2.11}$$

where each column vector is stacked with M elements such as $\mathbf{y}(l,k) = [y_1(l,k),\cdots,y_M(l,k)]^T$. Vector $\mathbf{d}(l,k)$ denotes the STFT coefficients of the late reverberation component and $\mathbf{v}(l,k)$ denotes the STFT coefficients of the noise component.

Typically, $\mathbf{y}(l,k)$ can be assumed to follow a circularly-symmetric complex Gaussian distribution with zero mean and cross power spectral density (CPSD) matrix

$$\begin{aligned}\mathbf{P_y}(l,k) &= \mathrm{E}\left[\mathbf{y}(l,k)\,\mathbf{y}^H(l,k)\right] \\ &= \mathbf{P_x}(l,k) + \mathbf{P_l}(l,k) + \mathbf{P_v}(l,k) \in \mathbb{C}^{M\times M},\end{aligned} \tag{2.12}$$

where we have assumed that $\mathbf{x}(l,k)$, $\mathbf{d}(l,k)$ and $\mathbf{v}(l,k)$ are statistically mutually uncorrelated. If the sources are also assumed uncorrelated, we have

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(l,k) &= \mathrm{E}\left[\mathbf{y}(l,k)\,\mathbf{y}^{H}(l,k)\right] \\ &= \sum_{r=1}^{R}\phi_r(l,k)\,\mathbf{a}_r(l,k)\,\mathbf{a}_r^{H}(l,k)+\mathbf{P}_{\mathbf{l}}(l,k)+\mathbf{P}_{\mathbf{v}}(l,k)\in\mathbb{C}^{M\times M},\end{aligned} \tag{2.13}$$

where $\phi_r(l,k)=\mathrm{E}\left[|s_r(l,k)|^2\right]$ is the PSD of the r-th source. We can also write Eq. (2.13) in the more compact form of

$$\mathbf{P}_{\mathbf{y}}(l,k)=\mathbf{A}(l,k)\,\mathbf{P}(l,k)\,\mathbf{A}^{H}(l,k)+\mathbf{P}_{\mathbf{l}}(l,k)+\mathbf{P}_{\mathbf{v}}(l,k)\in\mathbb{C}^{M\times M}, \tag{2.14}$$

where $\mathbf{P}(l,k)$ is diagonal with diagonal elements $\phi_r(l,k)$ for $r=1,\cdots,R$ and $\mathbf{A}(l,k)=[\mathbf{a}_1(l,k),\cdots,\mathbf{a}_R(l,k)]$.

Typically, the covariance matrix of the late reverberation $\mathbf{P}_{\mathbf{l}}(l,k)$ can be assumed to be the product of a time-invariant full rank spatial coherence matrix $\boldsymbol{\Gamma}(k)$ and a time-varying PSD $\phi_\gamma(l,k)$ [15], [16], that is,

$$\mathbf{P}_{\mathbf{l}}(l,k)=\phi_\gamma(l,k)\,\boldsymbol{\Gamma}(k)\ . \tag{2.15}$$

Since the late reverberation is a sum of many late reflections, we can use the law of large numbers to get many useful properties. For instance, the time domain late reverberation signal can be assumed to follow a zero-mean Gaussian distribution with its amplitude decaying exponentially (according to the room's reverberation time T_{60}) as the delay time [8], [17]. In most experiments of this dissertation, we assume that the reverberant sound field is diffuse, homogeneous and isotropic. Under this assumption we can calculate the normalized correlation (interchannel coherence) between every two different microphones analytically [18]. Assuming a spherical diffuse noise field for the late reverberation, we can use the following expression to calculate the interchannel coherence [18],

$$\boldsymbol{\Gamma}_{i,j}(k)=\mathrm{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right), \tag{2.16}$$

where $\mathrm{sinc}(\cdot)=\frac{\sin(\cdot)}{(\cdot)}$, $d_{i,j}$ is the inter-distance between microphones i and j, f_s is the sampling frequency, c is the speed of sound and K is the FFT length. For a cylindrical diffuse sound field, another similar expression exists to calculate the interchannel coherence, where the sinc function is replaced by the Bessel function [18]. For more complex situations where the spatial coherence matrix is difficult to describe, we assume it is time-invariant and can be measured. In summary, $\boldsymbol{\Gamma}_{i,j}(k)$ is always assumed given and the parameter that needs to be estimated is the time-varying PSD $\phi_\gamma(l,k)$.

2.3. DETAILED PROBLEM FORMULATION

Since there is a lot of variety among real world acoustic scenarios, we consider in this dissertation various scenarios ranging from a single-source and reverberant scenario to multi-source reverberant and noisy scenario. The different scenarios were presented in Fig. 1.3. We also illustrated the contribution of each Chapter in Fig. 1.5 and summarized the research questions in Fig. 1.4. These figures are linked to the scenarios in Fig. 1.3, with corresponding sub-figure indices indicating the relationship among contributions, research questions, and scenarios. In this section, we present different signal models (in particular the STFT domain signal model and its corresponding covariance matrix) in Fig. 2.2, corresponding to the scenarios from Fig. 1.3.

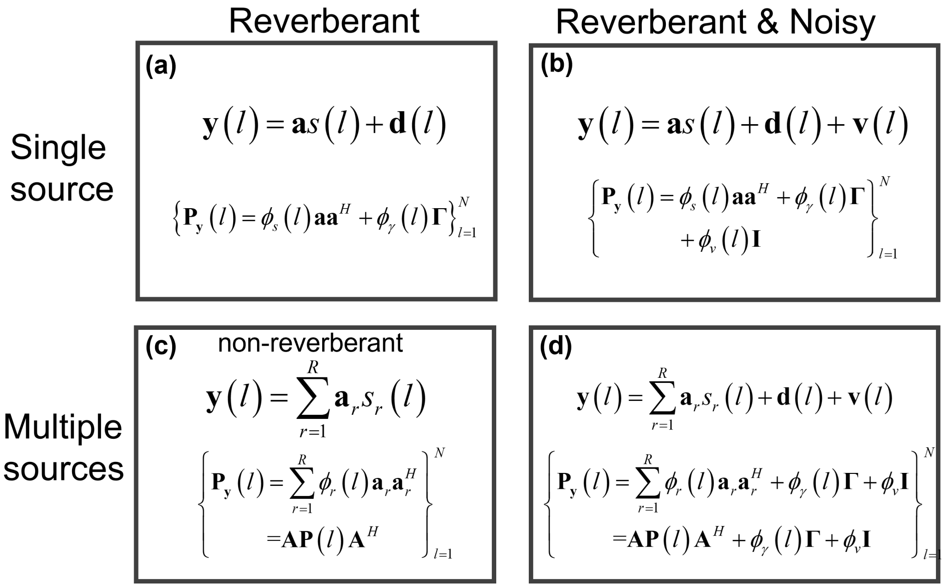

Reverberant
Reverberant & Noisy
Single source
Multiple sources
(a)
$\mathbf{y}(l)=\mathbf{a}s(l)+\mathbf{d}(l)$
$\left\{\mathbf{P}_{\mathrm{y}}(l)=\phi_s(l)\mathbf{a}\mathbf{a}^H+\phi_\gamma(l)\mathbf{\Gamma}\right\}_{l=1}^N$
(b)
$\mathbf{y}(l)=\mathbf{a}s(l)+\mathbf{d}(l)+\mathbf{v}(l)$
$\left\{\begin{array}{c}\mathbf{P}_{\mathrm{y}}(l)=\phi_s(l)\mathbf{a}\mathbf{a}^H+\phi_\gamma(l)\mathbf{\Gamma}\\+\phi_v(l)\mathbf{I}\end{array}\right\}_{l=1}^N$
(c)
non-reverberant
$\mathbf{y}(l)=\sum_{r=1}^R\mathbf{a}_r s_r(l)$
$\left\{\begin{array}{c}\mathbf{P}_{\mathrm{y}}(l)=\sum_{r=1}^R\phi_r(l)\mathbf{a}_r\mathbf{a}_r^H\\=\mathbf{A}\mathbf{P}(l)\mathbf{A}^H\end{array}\right\}_{l=1}^N$
(d)
$\mathbf{y}(l)=\sum_{r=1}^R\mathbf{a}_r s_r(l)+\mathbf{d}(l)+\mathbf{v}(l)$
$\left\{\begin{array}{c}\mathbf{P}_{\mathrm{y}}(l)=\sum_{r=1}^R\phi_r(l)\mathbf{a}_r\mathbf{a}_r^H+\phi_\gamma(l)\mathbf{\Gamma}+\phi_v\mathbf{I}\\=\mathbf{A}\mathbf{P}(l)\mathbf{A}^H+\phi_\gamma(l)\mathbf{\Gamma}+\phi_v\mathbf{I}\end{array}\right\}_{l=1}^N$

Figure 2.2: Different signal models for different scenarios (for each frequency).

- Fig. 2.2(a) depicts a single-source reverberant signal model, which addresses Research Question 1.1 (see Fig. 1.4) and will be explored in Chapter 3 (see Fig. 1.5).
- Fig. 2.2(b) presents a single-source reverberant and noisy signal model, which addresses Research Questions 1.2 and 1.3 (see Fig. 1.4) and will be investigated in Chapter 4 and Chapter 5 (see Fig. 1.5).
- Fig. 2.2(c) illustrates a multi-source signal model in a non-reverberant and noiseless environment, which addresses Research Question 2.1 (see Fig. 1.4) and will be explored in Chapter 6 (see Fig. 1.5). Note that in the experiments of Chapter 6 we will apply this method to a low-reverberant, near-noiseless environment.
- Fig. 2.2(d) depicts a multi-source reverberant and noisy signal model, which addresses Research Question 2.2 (see Fig. 1.4) and will be investigated in Chapter 7 (see Fig. 1.5).
- We will also present an adaptive time segmentation method as mentioned in

Chapter 1. This addresses Research Question 3 (see Fig. 1.4) and will be investigated in Chapter 8 (see Fig. 1.5). While this method can be combined with almost any of the methods presented in this thesis, we only assume a single-source reverberant signal model, which is the same as the model shown in Fig. 2.2(a).

Note that we omitted frequency indices in Fig. 2.2 since we will process each frequency bin separately. We use the time frame index to indicate that some parameters are time varying (with the index) and some are time invariant (without the index). We also use time frame index from 1 to N to indicate that we will use the signals within a time segment.

2.4. FILTERING

Given the multimicrophone observation $\mathbf{y} \in \mathbb{C}^M$, we can use a filtering function $f(\cdot)$ to reconstruct the target signal s:

$$\hat{s} = f(\mathbf{y}). \tag{2.17}$$

The filtering function can be non-linear but the corresponding filter design is challenging. In recent years, researchers use neural networks to learn non-linear filtering operations [19]. While neural networks are powerful tools for non-linear filtering of microphone signals, they have many limitations such as data dependency, computational demands, and generalization and interpretability issues, which must be carefully considered in real-life audio processing systems. In contrast, linear filtering methods have been widely used in real applications mainly due to their simplicity. The signal estimated using linear filters can be written as

$$\hat{s} = \mathbf{w}^H\mathbf{y} = \sum_{m=1}^{M} w_m^* y_m. \tag{2.18}$$

To demonstrate the effectiveness of our proposed methods for signal model parameter estimation, we will constrain ourselves to linear filters in the following chapters.

2.4.1. MWF

Let the error between the target signal s and the reconstructed signal $\hat{s}$ be

$$e = s - \hat{s} = s - \mathbf{w}^H\mathbf{y}. \tag{2.19}$$

Typically, we consider the signals to be random. We can therefore find the optimal filter coefficients by minimizing the variance of the error or find the minimum mean square error (MMSE), that is

$$\mathrm{E}\left[|e|^2\right] = \mathrm{E}\left[\left|s - \mathbf{w}^H\mathbf{y}\right|^2\right] = \mathbf{w}^H\mathbf{P}_\mathbf{y}\mathbf{w} - 2\Re\left\{\mathbf{w}^H\mathrm{E}\left[s^*\mathbf{y}\right]\right\} + \mathrm{E}\left[|s|^2\right], \tag{2.20}$$

where the signals s and $\mathbf{y}$ have been assumed zero mean.

In this dissertation, we consider the combination of the direct and early reflections of the r-th signal at the reference microphone to be the target signal, i.e., s_r. The remaining signal components including the interfering signals, the late reverberation and the noise are considered as distortion. Note that in general we consider a multi-source scenario. The filters related to a single source scenario can be derived similarly, by setting $R = 1$. Given $\mathbf{P_y}$ in Eq. (2.13) and the cross-correlation term $\mathrm{E}[s_r^*\mathbf{y}] = \phi_r\mathbf{a}_r$, the minimizer of the cost function in Eq. (2.20) is given by

$$\mathbf{w}_{MWF} = \left(\phi_r\mathbf{a}_r\mathbf{a}_r^H + \mathbf{P_n}\right)^{-1}\phi_r\mathbf{a}_r = \frac{\phi_r\mathbf{P_n}^{-1}\mathbf{a}_r}{1+\phi_r\mathbf{a}_r^H\mathbf{P_n}^{-1}\mathbf{a}_r}, \tag{2.21}$$

with

$$\mathbf{P_n} = \sum_{r_0=1, r_0\neq r}^{R} \phi_{r_0}\mathbf{a}_{r_0}\mathbf{a}_{r_0}^H + \mathbf{P_l} + \mathbf{P_v}. \tag{2.22}$$

The above filter is the well-known multichannel Wiener filter (MWF). As we set the r-th signal as the target, the noise covariance matrix $\mathbf{P_n}$ includes the interfering signals (i.e., the remaining point sources), the late reverberation and the ambient noise.

We can see that, before we can use the MWF filter, we first need to estimate the filter parameters, which include the PSDs and the RTFs of all the source signals $\{\phi_r, \mathbf{a}_r\}_{r=1}^R$, the covariance matrix of the late reverberation $\mathbf{P_l}$ and the covariance matrix of the ambient noise $\mathbf{P_v}$. If $\mathbf{P_l}$ and $\mathbf{P_v}$ are further modelled as $\mathbf{P_l} = \phi_\gamma\mathbf{\Gamma}$ and $\mathbf{P_v} = \phi_v\mathbf{I}$ (only considering microphone self-noise) with $\mathbf{\Gamma}$ given, then the parameters that need to be estimated include the PSD of the late reverberation ϕ_γ and the PSD of the microphone self-noise ϕ_v.

2.4.2. MVDR

Depending on the properties that we want for the reconstructed target signal, we need different filtering methods. For instance, the MWF filter aims at reducing the noise component in the reconstructed signal. We can also reduce the noise energy while keeping the reconstructed signal undistorted. That leads to another well-known and widely-used linear filtering method, the minimum variance distortionless response (MVDR) filter. Specifically, instead of finding the MMSE as with the MWF, the MVDR is the solution to the following optimization problem

$$\min_{\mathbf{w}} \mathbf{w}^H\mathbf{P_n}\mathbf{w} \text{ s.t. } \mathbf{w}^H\mathbf{a}_r = 1. \tag{2.23}$$

The above problem has a closed-form solution, which is

$$\mathbf{w}_{MVDR} = \frac{\mathbf{P_n}^{-1}\mathbf{a}_r}{\mathbf{a}_r^H\mathbf{P_n}^{-1}\mathbf{a}_r}. \tag{2.24}$$

The MVDR requires knowledge on $\mathbf{a}_r$ and $\mathbf{P_n}$. Estimating these can still be very challenging especially if noise or interfering components have time-varying PSDs.

Applying the MVDR to the noisy microphone signal $\mathbf{y}$, we have

$$\mathbf{w}_{MVDR}^H\mathbf{y} = \frac{\mathbf{a}_r^H\mathbf{P_n}^{-1}\mathbf{y}}{\mathbf{a}_r^H\mathbf{P_n}^{-1}\mathbf{a}_r}, \tag{2.25}$$

which is known to be a sufficient statistic for the target source s_r under a Gaussian noise assumption [20], [21]. That means there is no information loss on s_r by using $\mathbf{w}_{MVDR}^{H}\mathbf{y}$ instead of $\mathbf{y}$. Therefore, further processing such as the MMSE estimation can be applied directly to the MVDR output instead of being applied to $\mathbf{y}$. Moreover, it has been shown [22] that the MWF filter can be factorized into an MVDR filter and a post filter (single channel Wiener filter), that is

$$\mathbf{w}_{MWF} = \frac{\phi_r}{\phi_r + \left(\mathbf{a}_r^H {\mathbf{P}_\mathbf{n}}^{-1}\mathbf{a}_r\right)^{-1}} \mathbf{w}_{MVDR} \tag{2.26}$$

or

$$\mathbf{w}_{MWF} = \frac{\phi_r}{\phi_r + \mathbf{w}_{MVDR}^H \mathbf{P}_\mathbf{n} \mathbf{w}_{MVDR}} \mathbf{w}_{MVDR}. \tag{2.27}$$

Moreover, both the MVDR and the MWF can be seen as special cases of the speech distortion weighted MWF (SD-MWF) [23]. The optimization problem of SD-MWF is

$$\min_{\mathbf{w}} \left|1 - \mathbf{w}^H \mathbf{a}_r\right|^2 \phi_r + \mu \mathbf{w}^H \mathbf{P}_\mathbf{n} \mathbf{w}, \tag{2.28}$$

where μ is a parameter used to manipulate the tradeoff between speech distortion (the first term) and noise reduction (the second term). The SD-MWF also has a closed-form solution, which is

$$\mathbf{w}_{SD-MWF} = \frac{\phi_r {\mathbf{P}_\mathbf{n}}^{-1}\mathbf{a}_r}{\mu + \phi_r \mathbf{a}_r^H {\mathbf{P}_\mathbf{n}}^{-1}\mathbf{a}_r} = \frac{\phi_r}{\phi_r + \mu\left(\mathbf{a}_r^H {\mathbf{P}_\mathbf{n}}^{-1}\mathbf{a}_r\right)^{-1}} \mathbf{w}_{MVDR}. \tag{2.29}$$

We can choose the value of μ in the interval $(0, \infty)$ to get different levels of speech distortion compared to noise reduction. The two special cases are $\mu = 1$, which gives us the MWF filter, and $\mu \to 0$, which identifies the MVDR filter.

2.4.3. MPDR

The last filter we want to introduce is the minimum power distortionless response (MPDR). The optimization problem and the solution for the MPDR are

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{P}_\mathbf{y} \mathbf{w} \text{ s.t. } \mathbf{w}^H \mathbf{a}_r = 1. \tag{2.30}$$

and

$$\mathbf{w}_{MPDR} = \frac{{\mathbf{P}_\mathbf{y}}^{-1}\mathbf{a}_r}{\mathbf{a}_r^H {\mathbf{P}_\mathbf{y}}^{-1}\mathbf{a}_r}. \tag{2.31}$$

It has been proven that the MVDR and the MPDR are equivalent [24], [25], if $\mathbf{P}_\mathbf{y} = \phi_r \mathbf{a}\mathbf{a}^H + \mathbf{P}_\mathbf{n}$. We can see that to calculate the MPDR filter, we only need to estimate the RTF vector and $\mathbf{P}_\mathbf{y}$. However, the MPDR is less robust to RTF errors compared to other filters [25] like the MVDR.

2.5. TIME SEGMENTATION

As mentioned in previous sections, we often use the STFT procedure to transform time-domain microphone signals $\mathbf{y}(t)$ into the frequency domain signals $\mathbf{y}(l,k)$, that is,

$$\mathbf{y}(l,k) = \sum_t \mathbf{y}(t)\,\psi(t-lT)\,e^{-j\frac{2\pi}{N}k(t-lT)}. \tag{2.32}$$

With this procedure, we apply an analysis window $\psi(t)$ to the microphone signal $\mathbf{y}(t)$ before an N-length FFT is applied to the signal. The window length is usually very short in duration (about 30 ms), which means that $\mathbf{y}(t)\,\psi(t-lT)$ only captures information within a short duration. We define this duration as a sub-time frame indexed by l.

We also showed that in order to use the spatial filters such as the MVDR and the MWF, we need to estimate the acoustic scene related parameters, among which the covariance matrix of $y(l,k)$, $\mathbf{P}_{\mathbf{y}}$. A common way to estimate $\mathbf{P}_{\mathbf{y}}$ is using the sample covariance matrix

$$\hat{\mathbf{P}}_{\mathbf{y}}(i,k) = \frac{1}{L_{sf}} \sum_{l=1+(i-1)L_{sf}}^{iL_{sf}} \mathbf{y}(l,k)\,\mathbf{y}(l,k)^H. \tag{2.33}$$

However, this requires the assumption that the microphone signals are stationary and ergodic over the time indices $l = 1+(i-1)L_{sf},\cdots,iL_{sf}$. On the other hand, speech signals can be assumed to be stationary for a short duration of at most 50 ms. It means that when using 50% overlap and 30 ms frames to obey the stationarity assumption, L_{sf} should be less than 2 windows. Such a small value of L_{sf} leads to inaccurate estimated covariance matrices. Also, with only 2 samples, the estimated covariance matrix has a rank of less than 2, which hinders the way for further parameter estimation. We therefore select values of L_{sf} in the range of 20 and 40. Although the microphone signals in this duration are not stationary, the sample covariance matrix can be seen as the estimate of the average of the different ground truth covariance matrices within this duration. We define this duration of L_{sf} windows as a time frame.

Lastly, we define the duration that the sound source position does not change as a time segment, i.e., the duration over which $\mathbf{a}$ is assumed not to change. Since the time segment duration is usually much longer than the time frame duration, we are in this dissertation interested in the number of time frames per time segment. With that knowledge, we can use the prior information that the RTF $\mathbf{a}$ is constant for multiple time frames that belong to the same time segment. For a static sound source, we can choose the number of time frames very large. However, in practice, we need to consider the computation time and the latency. The number of time frames per time segment, can therefore not be too large. For a moving source, we might need to detect the point in time that the source position changes and adapt the time segment accordingly.

REFERENCES

[1] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[2] J. Chen, J. Benesty, and Y. Huang, “A minimum distortion noise reduction algorithm with multiple microphones”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 481–493, 2008.

[3] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2002.

[4] Y. Avargel and I. Cohen, “System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[5] R. Talmon, I. Cohen, and S. Gannot, “Relative Transfer Function Identification Using Convolutive Transfer Function Approximation”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[6] X. Li, L. Girin, S. Gannot, and R. Horaud, “Multichannel speech separation and enhancement using the convolutive transfer function”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 645–659, 2019.

[7] F. Feng and M. Kowalski, “Underdetermined reverberant blind source separation: Sparse approaches for multiplicative and convolutive narrowband approximation”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 442–456, 2019.

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

[9] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech”, *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[10] D. Cherkassky and S. Gannot, “Successive Relative Transfer Function Identification Using Blind Oblique Projection”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 474–486, 2020.

[11] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[12] J. Zhang, H. Chen, L.-R. Dai, and R. C. Hendriks, “A study on reference microphone selection for multi-microphone speech enhancement”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 671–683, 2021.

[13] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Frequency-domain blind source separation of many speech signals using near-field and far-field models”, *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[14] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, “Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207, 2009.

[15] S. Braun and E. A. Habets, “Dereverberation in noisy environments using reference signals and a maximum likelihood estimator”, in *Proc. EURASIP Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[16] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[17] J.-D. Polack, “Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics”, *Appl. Acoust.*, vol. 38, no. 2, pp. 235–244, 1993, ISSN: 0003-682X.

[18] B. F. Cron and C. H. Sherman, “Spatial-correlation functions for various noise models”, *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[19] K. Tesch and T. Gerkmann, “Insights into deep non-linear filters for improved multi-channel speech enhancement”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 563–575, 2023.

[20] M. J. Schervish, *Theory of statistics*. Springer Science & Business Media, 2012.

[21] K. Tesch and T. Gerkmann, “Nonlinear spatial filtering in multichannel speech enhancement”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1795–1805, 2021.

[22] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013.

[23] J. Benesty, S. Makino, J. Chen, S. Doclo, A. Spriet, J. Wouters, and M. Moonen, “Speech distortion weighted multichannel wiener filtering techniques for noise reduction”, *Speech enhancement*, pp. 199–228, 2005.

[24] H. L. Van Trees, *Detection, estimation, and modulation theory, optimum array processing*. John Wiley & Sons, 2004.

[25] H. Cox, “Resolving power and sensitivity to mismatch of optimum array processors”, *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.

3

JOINT MAXIMUM LIKELIHOOD ESTIMATION OF MICROPHONE ARRAY PARAMETERS FOR A REVERBERANT SINGLE SOURCE SCENARIO

This chapter is based on the article published as: C. Li, J. Martinez, and R. C. Hendriks. "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario". In: IEEE/ACM Trans. Audio, Speech, Language Process. 31 (2023), pp. 695– 705.

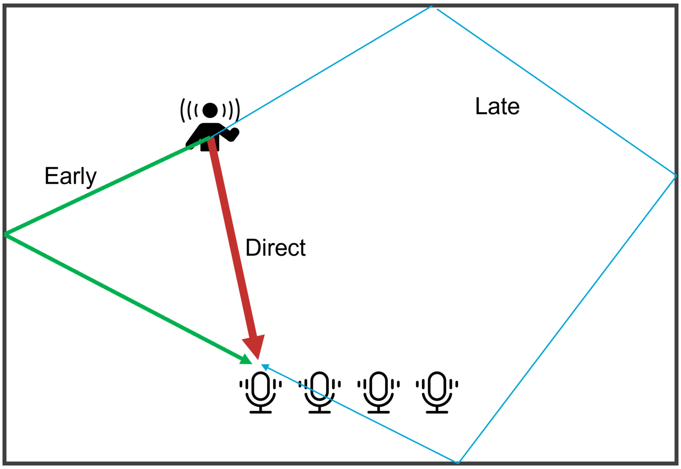


Figure 3.1: Illustration of a single reverberant source.

In this chapter, we will consider the single-source reverberant scenario as illustrated in Fig. 3.1, with which we will address research question 1.1 shown in Fig. 1.4 using the signal model presented in Fig. 2.2 (a).

Estimation of the acoustic-scene related parameters such as relative transfer functions (RTFs) from source to microphones, source power spectral densities (PSDs) and PSDs of the late reverberation is essential and also challenging. Existing maximum likelihood estimators typically consider only subsets of these parameters and use each time frame separately. In this chapter we explicitly focus on the single source scenario and first propose a joint maximum likelihood estimator (MLE) to estimate all parameters jointly using a single time frame. Since the RTFs are typically invariant for a number of consecutive time frames we also propose a joint maximum likelihood estimator (MLE) using multiple time frames, which has similar estimation performance compared to a recently proposed reference algorithm called simultaneously confirmatory factor analysis (SCFA), but at a much lower complexity. Moreover, we present experimental results which demonstrate that in terms of the estimation accuracy, together with the performance of noise reduction, speech quality and speech intelligibility, our proposed joint MLE outperforms those of existing MLE based approaches that use only a single time frame.

3.1. INTRODUCTION

Microphone array signal processing has ubiquitous applications like source dereverberation [1]–[4], noise reduction [5]–[8], source separation [9]–[11] and source localization [12]. These applications heavily depend on acoustic-scene related parameters such as relative transfer functions (RTFs), power spectral densities (PSDs) of the source, PSDs of the late reverberation and PSDs of the microphone self noise. These parameters are typically unknown in practical scenarios. Therefore, estimation of these parameters is an essential problem for microphone array signal processing applications.

As speech sources are typically non-stationary, their PSD changes over time. Moreover, the source might be moving, resulting in changes in the RTF as well. The estimation of the RTF and the PSDs of the source and the late reverberation is therefore rather challenging, especially when considering to estimate them simultaneously at low complexity. To get a full understanding of the problem, we constrain ourselves in this chapter to the single source reverberant scenario and focus on the joint estimation of the source's RTF, PSD of the early reflections and the PSD of the late reverberation. In future work, we will extend this towards the multi-source scenario.

There are many existing methods that consider maximum likelihood estimation of these parameters [1], [13]–[16]. However, most of these methods do not estimate the parameters in a joint manner. In [1], [13], the RTFs are assumed to be known and the MLE for the PSDs of the source and the late reverberation is proposed. In [2], the estimate of the late reverberation is obtained without estimating the RTFs or the PSDs of the source. In [14], the RTFs are estimated given that the PSDs of the late reverberation are assumed to be known or have been estimated. In [15], by assuming the late reverberation is stationary, the expectation maximization (EM) method [17] was used to estimate the RTFs and the PSD of the source. However, in practice, the late reverberation is non-stationary and the PSDs of the late reverberation can change from time-frame to time-frame, which limits the scenarios to which the method in [15] can be applied.

Apart from the fact that most reference methods only estimate a subset of these parameters, all these methods, i.e., [1], [13]–[16], use each time frame separately. However, in most practical scenes, the RTFs change slower than the PSDs of the source and the late reverberation, and can be assumed invariant for a number of consecutive time frames. Therefore, better estimates of these parameters can be obtained by using the time frames that share the same RTFs jointly. A recently proposed method referred to as the simultaneous confirmatory factor analysis (SCFA) method considers the joint estimation of these parameters using multiple time frames [18] and has a much better estimation performance compared to methods using each time frame separately. However, since the problem formulated in [18] is non-convex, this method suffers from a rather high computational cost, which makes it difficult to be applied when dealing with practical problems.

To estimate all the aforementioned parameters of interest jointly and accurately with low computational complexity, we first propose a joint maximum likelihood estimator (MLE) using a single time frame. This has a closed form solution and can be solved efficiently. Note that recently the joint MLE using a single time frame is also proposed in [16], but we provide an alternative proof. More importantly, we propose an extension, which is a joint MLE using multiple time frames. This extension uses the rough estimates obtained by the MLE for a single time frame as initialisation and estimates all the parameters in an iterative manner. Since the computational cost for each step in the proposed method mainly comes from an eigenvalue decomposition, it has similar computational complexity as the MLE approach for a single time frame. Experimental results demonstrate that our proposed MLE for multiple time frames has similar estimation performance compared to the recently proposed SCFA method from [18], but, at a much lower computational complexity. Moreover, both the proposed and SCFA methods outperform two other reference methods that consist of combining several existing state-of-the-art methods.

The remaining parts of the chapter are structured as follows. We present the notation, the signal model and the main goal of this chapter in Section 3.2. In Section 3.3, we propose the joint maximum likelihood estimator using a single time frame in Section 3.3.1 and using multiple time frames in Section 3.3.2. In Section 3.4, we first introduce some reference methods and compare them to our proposed joint MLE in different acoustic scenarios. In the last section, Section 3.5, conclusions will be drawn.

The matlab code of the joint MLE can be downloaded from:
http://sps.ewi.tudelft.nl/Repository/

3.2. PRELIMINARIES

3.2.1. NOTATION

In this chapter, we denote scalars using lower-case letters, vectors using bold-face lower-case letters and matrices using bold-face upper-case letters (in some cases with subscripts using bold-face lower-case letters, e.g. $\mathbf{P}_{\mathbf{y}}$). Matrix notation with subscripts using two lower-case letters (e.g. $\mathbf{P}_{\mathbf{y}_{i,j}}$) denotes the element of the matrix. $\Re(\cdot)$ and $\Im(\cdot)$ represents the real part and the imaginary part of a complex-valued variable, respectively. Further, $\mathrm{E}(\cdot)$ denotes the expected value of a random variable, $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, and if not further specified, $|\cdot|$ denotes the determinant of a matrix. Finally, $\mathrm{diag}[a_1, \cdots, a_M]$ denotes a diagonal matrix with diagonal elements $a_1, \cdots, a_M$ and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

3.2.2. SIGNAL MODEL

We consider a single acoustic point source observed by a microphone array consisting of M microphones with an arbitrary geometric structure in a reverberant and noisy environment. Decomposing the signal into its direct component with its early reflections, and the late reverberant components, we can write the signal received at the m_{th} microphone in the short-time Fourier transform (STFT) domain as

$$y_m(i,k) = x_m(i,k) + l_m(i,k) + v_m(i,k), \tag{3.1}$$

where i is the time-frame index and k is the frequency bin index, $x_m(i,k)$ is the sum of the direct components and the early reflections, $l_m(i,k)$ is the sum of all late reflections and $v_m(i,k)$ is the microphone self-noise. The direct components and early reflections are beneficial for speech intelligibility [19]. The combination of these components, denoted by $x_m(i,k)$ in Eq. (3.1), forms our target signal. In this work, we differentiate between time segments (indexed by β) and time frames (indexed by i). Each time segment consists of N time frames, i.e., for each β, $i=(\beta-1)N+1,\cdots,\beta N$. The target signal at the m_{th} microphone is given by

$$x_m(i,k)=a_m(\beta,k)\,s(i,k), \tag{3.2}$$

where $a_m(\beta,k)$ is the relative transfer function (RTF) for source s from the reference location to the m_{th} microphone in time segment β and s is the target source including direct and early reflections at the reference microphone. Note that, for ease of analyzing, we use the multiplicative transfer function (MTF) approximation instead of the convolutive transfer function (CTF) approximation in Eq. (3.2). CTF can be more accurate than MTF but has a more complicated signal model [20], [21]. We assume that the RTFs are constant during a time segment (thus during multiple time frames that fall in one segment) and $a_1=1$, which means that the first microphone is selected as the reference microphone. Stacking the M microphone STFT coefficients into a column vector, we have

$$\mathbf{y}(i,k)=\mathbf{a}(\beta,k)\,s(i,k)+\mathbf{l}(i,k)+\mathbf{v}(i,k)\in\mathbb{C}^{M\times 1}. \tag{3.3}$$

3.2.3. CROSS POWER SPECTRAL DENSITY MATRICES

We assume the STFT coefficients of the microphone signal have a circularly-symmetric complex Gaussian distribution[1], i.e.: $\mathbf{y}(i,k)\sim\mathcal{N}_C(\mathbf{0},\mathbf{P_y}(i,k))$, where $\mathbf{P_y}(i,k)$ is the noisy cross power spectral density (CPSD) matrix, expressing the covariance across microphones. Assuming that all components in Eq. (3.3) are Gaussian distributed with zero mean and mutually uncorrelated, we have

$$\mathbf{P_y}(i,k)=\mathbf{P_x}(i,k)+\mathbf{P_l}(i,k)+\mathbf{P_v}(i,k)\in\mathbb{C}^{M\times M}, \tag{3.4}$$

where $\mathbf{P_x}$ is given by

$$\mathbf{P_x}(i,k)=p(i,k)\mathbf{a}(\beta,k)\mathbf{a}^H(\beta,k), \tag{3.5}$$

and where $p(i,k)=\mathrm{E}\left[|s(i,k)|^2\right]$ is the power spectral density (PSD) of the source at the reference microphone with $|\cdot|$ the absolute value. Note that although the mutual uncorrelation assumption is commonly used, these components are not perfectly uncorrelated in practice.

The CPSD matrix of the late reverberation component is commonly modelled as [1], [26]

$$\mathbf{P_l}(i,k)=\gamma(i,k)\boldsymbol{\Gamma}(k), \tag{3.6}$$

[1] Although a super-Gaussian distribution can better model the coefficients [22]–[24], the estimators based on it are much more cumbersome than that based on the Gaussian distribution [25] and hence are not considered in this chapter.

where the time-varying coefficient $\gamma(i,k)$ is the PSD of the late reverberation and the time-invariant matrix $\mathbf{\Gamma}(k)$ is the spatial coherence matrix of the late reverberation. $\mathbf{\Gamma}(k)$ is assumed to be non-singular and known in this chapter. Several methods have been proposed to measure $\mathbf{\Gamma}(k)$ by using pre-calculated room impulse responses [27] or by using knowledge on the microphone array geometry [28], [29]. We use the latter one and model the coherence matrix as a spherically isotropic noise field [30]

$$\mathbf{\Gamma}(k) = \mathrm{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right), \tag{3.7}$$

where $\mathrm{sinc}(x) = \frac{\sin x}{x}$, $d_{i,j}$ is the inter-distance between microphones i and j, f_s is the sampling frequency, c denotes the speed of sound and K is the number of frequency bins.

Lastly, the microphone self-noise component is assumed to have slow varying statistics and its CPSD matrix $\mathbf{P_v}(i,k)$ can be modelled as a time-invariant diagonal matrix with its M diagonal elements being the PSD of the self noise corresponding to the M microphones

$$\mathbf{P_v}(k) = \mathrm{diag}\left[n_1(k), \cdots, n_M(k)\right]. \tag{3.8}$$

Due to its time-invariant property, a voice activity detector (VAD) can be used to detect the noise-only segments of the signal such that the covariance matrix of the noise can be estimated [31]. Moreover, the power of the microphone self-noise is usually very small compared to the other components. Therefore, we assume in this chapter that $\mathbf{P_v}(k)$ is negligible or can be subtracted from the noisy covariance matrix.

3.2.4. PROBLEM FORMULATION

Based on the assumptions made in the previous subsection and Eqs. (3.5) and (3.6), we can rewrite the noisy CPSD matrix for each time frame i as

$$\mathbf{P_y}(i,k) = p(i,k)\mathbf{a}(\beta,k)\mathbf{a}^H(\beta,k) + \gamma(i,k)\mathbf{\Gamma}(k). \tag{3.9}$$

Each time frame i consists of T_{sf} overlapping sub frames indexed by t_s, each with equal length N_s. For a visual interpretation of time segments, frames and sub frames see Fig. 3.2. Assuming the noisy signal is stationary within a time frame, we can estimate the CPSD matrix per time frame i based on a sampled covariance matrix using the sub-time frames, that is,

$$\hat{\mathbf{P}}_{\mathbf{y}}(i,k) = \frac{1+iT_{sf}}{iT_{sf}}\sum_{t_s=1}^{T_{sf}} \mathbf{y}(t_s,k)\mathbf{y}(t_s,k)^H, \tag{3.10}$$

where $\mathbf{y}(t_s,k)$ denotes the STFT coefficients vector, where $\lceil\cdot\rceil$ denotes taking the next highest integer. Note that each time frame contains multiple sub-time frames as illustrated in Fig. 3.2 and these sub-time frames are used to estimate the covariance matrix of a single time frame. Notice that across the time frames of one time segment,

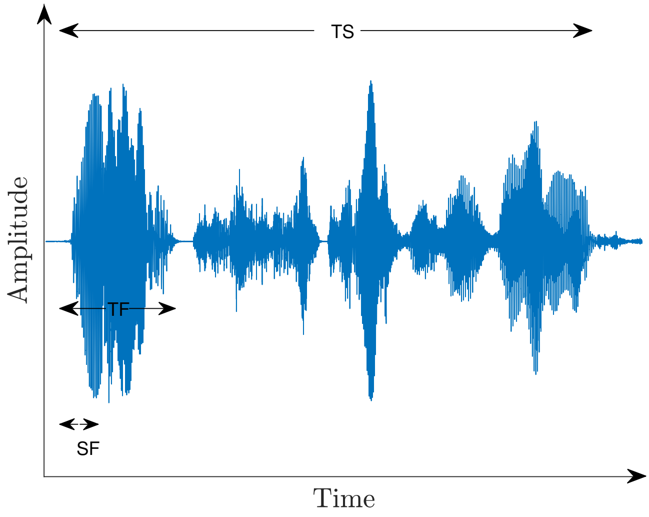


Figure 3.2: Illustration of the durations of Time segment (TS), time frames (TF) and subframes (SF).

the RTF vector is assumed to be constant and the PSDs of the source and late reverberation power $\gamma(i,k)$ are assumed to be time-variant.

Accurate estimation of the parameters from the signal model in Eq. (3.9) is very important for speech enhancement and intelligibility improvement algorithms. However, this is also very challenging when the source is only stationary for a short time and microphone and source positions are time varying. The main goal of this chapter therefore is to estimate the RTF vector, the PSD of the source and the PSD of the late reverberation simultaneously using N estimated CPSD matrices $\hat{\mathbf{P}}_{\mathbf{y}}(i,k)$ for $i=1,\cdots,N$, while the source is only stationary within a time frame and the RTF changes from segment to segment. Since we process the signal for each frequency bin independently, we omit the frequency bin index k in the following sections for notational convenience.

3.3. JOINT MLE

In this work, we present a novel maximum likelihood estimator (MLE) to jointly estimate the parameters from the signal model in Eq. (3.9). Note that MLEs have been proposed before in this context [1], [13], [15], but typically they assume that the RTF vector $\mathbf{a}$ is known and only determine the MLEs of $p(i)$ and $\gamma(i)$ for each time frame i separately. We will first in Section 3.3.1 propose the joint MLE estimator of $p(i)$, $\mathbf{a}$ and $\gamma(i)$ using the estimated noisy CPSD matrix for a single time frame. Since the CPSD matrices for multiple time frames in a single time segment share the same RTF vector, we can use these matrices jointly to obtain a better estimate of $\mathbf{a}$. Therefore, we will also propose in Section 3.3.2 the joint MLE estimator of $p(i), \mathbf{a}$ and $\gamma(i)$ using the CPSD matrices for multiple time frames.

3.3.1. JOINT MLE FOR A SINGLE TIME FRAME

Assuming that the T_{sf} sub-time frames in a single time frame i per frequency band k are independent and identically distributed (i.i.d.), we can write the joint PDF $f\left(\mathbf{y}(1), \cdots, \mathbf{y}\left(T_{sf}\right)\right)$ as

$$f\left(\mathbf{y}(1), \cdots, \mathbf{y}\left(T_{sf}\right)\right) = \left(\frac{\exp\left[-\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right)\right]}{\pi^M \left|\mathbf{P}_{\mathbf{y}}\right|}\right)^{T_{sf}}, \tag{3.11}$$

where $\hat{\mathbf{P}}_{\mathbf{y}}$ is given in Eq. (3.10) and $\mathbf{P}_{\mathbf{y}}$ in Eq. (3.9). The negative log-likelihood function with respect to (w.r.t.) $p, \mathbf{a}$ and γ is given by

$$-L(p, \mathbf{a}, \gamma) = T_{sf}\left[\log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right)\right], \tag{3.12}$$

where the additive constant term $T_{sf}M\log\pi$ has been omitted as it is irrelevant for the parameters of interest. The MLEs of $p, \mathbf{a}$ and γ are given by minimizing the cost function in Eq. (3.12), i.e.,

$$\underset{p,\mathbf{a},\gamma}{\arg\min} \log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right). \tag{3.13}$$

To solve this problem, we reparameterize the signal model in Eq. (3.9) as

$$\begin{aligned}\mathbf{P}_{\mathbf{y}} &= p\mathbf{a}\mathbf{a}^H + \gamma\mathbf{\Gamma} \\ &= \mathbf{L}\left(p\mathbf{L}^{-1}\mathbf{a}\mathbf{a}^H\mathbf{L}^{-H} + \gamma\mathbf{I}\right)\mathbf{L}^H \\ &= \mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\mathbf{I}\right)\mathbf{L}^H,\end{aligned} \tag{3.14}$$

where $\mathbf{L}$ is the Cholesky factor of $\mathbf{\Gamma}$ (i.e. $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H$), $\tilde{\mathbf{a}} = \frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{p} = p\mathbf{a}^H\mathbf{\Gamma}^{-1}\mathbf{a}$. Therefore, the optimization problem in Eq. (3.13) can be cast as

$$\underset{\tilde{p},\tilde{\mathbf{a}},\gamma}{\arg\min} \log\left|\mathbf{P}_{\mathbf{y}}\right| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{P}_{\mathbf{y}}^{-1}\right). \tag{3.15}$$

By using this reparameterization, we can make the estimation of $\tilde{\mathbf{a}}$ independent of the estimation of $\tilde{p}$ and γ. Therefore, the joint estimation of these parameters can be decomposed into two simpler estimation steps, as we will show below.

The first term in Eq. (3.15) can be rewritten as

$$\begin{aligned}\log\left|\mathbf{P}_{\mathbf{y}}\right| &= \log\left|\mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\mathbf{I}\right)\mathbf{L}^H\right| \\ &= \log\left(\left|\mathbf{L}\right|\left(\tilde{p}\tilde{\mathbf{a}}^H\tilde{\mathbf{a}} + \gamma\right)\gamma^{M-1}\left|\mathbf{L}^H\right|\right) \\ &= \log\left(\left|\mathbf{\Gamma}\right|\right) + \log\left(\tilde{p} + \gamma\right) + (M-1)\log(\gamma),\end{aligned} \tag{3.16}$$

where we have used the fact that $\tilde{\mathbf{a}}^H\tilde{\mathbf{a}} = 1$. The second term in Eq. (3.15) can be rewritten as

$$\begin{aligned}
\operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{y}\mathbf{P}_\mathbf{y}^{-1}\right) &= \operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{y}\left[\mathbf{L}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\mathbf{I}\right)\mathbf{L}^H\right]^{-1}\right) \\
&= \operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{w}\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\mathbf{I}\right)^{-1}\right) \\
&= \operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{w}\left(\gamma^{-1}\mathbf{I} - \frac{\gamma^{-2}\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H}{1+\gamma^{-1}\tilde{p}\tilde{\mathbf{a}}^H\tilde{\mathbf{a}}}\right)\right) \\
&= \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_\mathbf{w}\right) - \operatorname{tr}\left(\frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\hat{\mathbf{P}}_\mathbf{w}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H\right) \\
&= \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_\mathbf{w}\right) - \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_\mathbf{w}\tilde{\mathbf{a}},
\end{aligned} \tag{3.17}$$

where $\hat{\mathbf{P}}_\mathbf{w} = \mathbf{L}^{-1}\hat{\mathbf{P}}_\mathbf{y}\mathbf{L}^{-H}$ and the Sherman–Morrison formula [32] is used to calculate $\left(\tilde{p}\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\mathbf{I}\right)^{-1}$.

Substituting Eq. (3.16) and Eq. (3.17) in Eq. (3.15) and omitting the constant irrelevant term $\log(|\mathbf{\Gamma}|)$, the cost function from Eq. (3.13) can eventually thus be expressed in the following useful form,

$$\begin{aligned}
\underset{\tilde{p},\tilde{\mathbf{a}},\gamma}{\arg\min} \log\left(\tilde{p}+\gamma\right)\left(\gamma^{M-1}\right) + \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_\mathbf{w}\right) \\
- \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_\mathbf{w}\tilde{\mathbf{a}}.
\end{aligned} \tag{3.18}$$

Since only the last term in Eq. (3.18) depends on $\tilde{\mathbf{a}}$ and $\frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}} > 0$, the estimate of $\tilde{\mathbf{a}}$ can be obtained by solving

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\, \tilde{\mathbf{a}}^H\hat{\mathbf{P}}_\mathbf{w}\tilde{\mathbf{a}}. \tag{3.19}$$

The solution of Eq. (3.19) is known as the principal eigenvector of $\hat{\mathbf{P}}_\mathbf{w}$ and the optimum value of $\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_\mathbf{w}\tilde{\mathbf{a}}$ is the principal eigenvalue λ_{max} of $\hat{\mathbf{P}}_\mathbf{w}$.

Substituting the optimal $\tilde{\mathbf{a}}$ from Eq. (3.19) in Eq. (3.18), we can find the estimates of $\tilde{p}$ and γ by solving

$$\begin{aligned}
\underset{\tilde{p},\gamma}{\arg\min}\, f = \log\left[\left(\tilde{p}+\gamma\right)\gamma^{M-1}\right] + \operatorname{tr}\left(\gamma^{-1}\hat{\mathbf{P}}_\mathbf{w}\right) \\
- \frac{\gamma^{-2}\tilde{p}}{1+\gamma^{-1}\tilde{p}}\lambda_{\text{max}}.
\end{aligned} \tag{3.20}$$

Taking the partial derivatives of the cost function in Eq. (3.20) w.r.t. $\tilde{p}$ and γ and setting them equal to zero, respectively, we obtain

$$\begin{aligned}
\frac{\partial f}{\partial \gamma} = \frac{1}{\tilde{p}+\gamma} + \frac{M-1}{\gamma} - \frac{\operatorname{tr}\left(\mathbf{L}^{-1}\hat{\mathbf{P}}_\mathbf{y}\mathbf{L}^{-H}\right)}{\gamma^2} \\
+ \frac{\tilde{p}\left(2\gamma+\tilde{p}\right)}{\left(\gamma^2+\gamma\tilde{p}\right)^2}\lambda_{\text{max}} = 0
\end{aligned} \tag{3.21}$$

and

$$\frac{\partial f}{\partial \tilde{p}} = \frac{1}{\tilde{p}+\gamma} - \frac{\lambda_{\max}}{(\gamma+\tilde{p})^2} = 0. \tag{3.22}$$

Solving Eq. (3.21) and Eq. (3.22) for $\tilde{p}$ and γ, we obtain

$$\hat{\tilde{p}} = \frac{M\lambda_{\max} - \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{M-1}, \tag{3.23}$$

$$\hat{\gamma} = \frac{\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right) - \lambda_{\max}}{M-1}. \tag{3.24}$$

To show that $\left(\hat{\tilde{p}}, \hat{\gamma}\right)$ is the minimum point of function f, we derive its second order derivatives

$$\begin{aligned}\frac{\partial^2 f}{\partial \gamma^2} = &- \frac{1}{(\tilde{p}+\gamma)^2} - \frac{M-1}{\gamma^2} + \frac{2\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{\gamma^3} \\ &+ \frac{2\lambda_{\max}\left(-3\gamma^2\tilde{p} - 3\gamma\tilde{p}^2 - \tilde{p}^3\right)}{\gamma^3(\gamma+\tilde{p})^3},\end{aligned} \tag{3.25}$$

$$\frac{\partial^2 f}{\partial \gamma \partial \tilde{p}} = -\frac{1}{(\tilde{p}+\gamma)^2} + \frac{2\lambda_{\max}}{(\gamma+\tilde{p})^3}, \tag{3.26}$$

$$\frac{\partial^2 f}{\partial \tilde{p}^2} = -\frac{1}{(\tilde{p}+\gamma)^2} + \frac{2\lambda_{\max}}{(\gamma+\tilde{p})^3}. \tag{3.27}$$

At point $\left(\hat{\tilde{p}}, \hat{\gamma}\right)$, we have

$$\left.\frac{\partial^2 f}{\partial \gamma^2}\right|_{\gamma=\hat{\gamma}} = \frac{(M-1)^3}{\left(\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right) - \lambda_{\max}\right)^2} + \frac{1}{(\lambda_{\max})^2} > 0, \tag{3.28}$$

$$\left.\frac{\partial^2 f}{\partial \tilde{p}^2}\right|_{\tilde{p}=\hat{\tilde{p}}} = \frac{1}{(\lambda_{\max})^2} > 0, \tag{3.29}$$

$$\left.\frac{\partial^2 f}{\partial \gamma^2}\frac{\partial^2 f}{\partial \tilde{p}^2} - \left(\frac{\partial^2 f}{\partial \gamma \partial \tilde{p}}\right)^2\right|_{\substack{\gamma=\hat{\gamma} \\ \tilde{p}=\hat{\tilde{p}}}} = \frac{(M-1)^3/(\lambda_{\max})^2}{\left(\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right) - \lambda_{\max}\right)^2} > 0. \tag{3.30}$$

Therefore, $\left(\hat{\tilde{p}}, \hat{\gamma}\right)$ is the minimum point of function f. Furthermore, we can show that $\hat{\tilde{p}}, \hat{\gamma}$ are both positive such that they can be used as the estimates of $\tilde{p}$ and γ. Since $\hat{\mathbf{P}}_{\mathbf{w}}$ is a positive definite matrix and it's typically not scaled identity matrix, we have $\frac{\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)}{M} < \lambda_{\max} < \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}\right)$. Hence from Eqs. (3.23) and (3.24) it follows that $\hat{\tilde{p}} > 0$ and $\hat{\gamma} > 0$. Note that this examination of the Hessian matrix and $\hat{\tilde{p}}, \hat{\gamma}$ being positive is absent in [16].

Finally, we obtain the optimal estimates of p and $\mathbf{a}$ using the estimated $\hat{\tilde{p}}$ and $\hat{\tilde{\mathbf{a}}}$ by setting

$$\hat{\mathbf{a}} = \mathrm{No}\left(\mathbf{L}\hat{\tilde{\mathbf{a}}}\right) \tag{3.31}$$

and

$$\hat{p} = \frac{\hat{\tilde{p}}}{\hat{\mathbf{a}}^H \boldsymbol{\Gamma}^{-1} \hat{\mathbf{a}}}, \tag{3.32}$$

where $\mathrm{N}o\left(\mathbf{x}\right)$ means taking normalization w.r.t. the first element of $\mathbf{x}$.

As mentioned in [16], the estimation of $\mathbf{a}$ is consistent with the covariance whitening method [5], [14], while we provided an alternative proof that this estimate equals the MLE of $\mathbf{a}$. More specifically, the proof in [16] with respect to the estimation of the PSDs does not include the examination of the Hessian matrix and the estimates of the PSDs being positive. This examination of the Hessian matrix being positive definite is necessary, since setting the partial derivative to zero does not give us the optimal estimate when the Hessian matrix is not positive definite. Also, the examination of estimates of the PSDs being positive is necessary, since the PSDs should always be positive. Moreover, the proof in [16] is based on the proportion of the likelihood function, which makes it difficult to analyze the cost function for multiple time frames. While, in this work, our proof is based on the likelihood function itself and the extension to multiple time frames is straightforward.

3.3.2. JOINT MLE FOR MULTIPLE TIME FRAMES

In the previous subsection we considered the joint MLE for p, γ and $\mathbf{a}$ given a single time frame. As $\mathbf{a}$ is assumed to stay fixed across multiple frames in a segment, we consider in this subsection the joint ML optimal estimates of $p\left(i\right)$, $\gamma\left(i\right)$ for $i = 1, \cdots, N$ and $\mathbf{a}$ using all time-frames in a segment.

Assuming that the N time frames are independent, we can write the negative log-likelihood function of the STFT coefficients as

$$L = -\sum_{i=1}^{N} T_{sf}\left[\log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right| + \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right)\right], \tag{3.33}$$

where non-essential constant terms have been omitted. The joint MLEs for $p\left(i\right)$, $\gamma\left(i\right) \forall i = 1, \cdots, N$ and $\mathbf{a}$ are the solution to the optimization problem

$$\underset{p(i),\mathbf{a},\gamma(i)}{\arg\min} \sum_{i=1}^{N} \log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right| + \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right). \tag{3.34}$$

By reparameterizing the signal model in a similar way as in the previous subsection, i.e., using $\tilde{\mathbf{a}} = \frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^H \boldsymbol{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{p} = p\mathbf{a}^H \boldsymbol{\Gamma}^{-1}\mathbf{a}$, the CPSD matrix for each time frame i has the form

$$\mathbf{P}_{\mathbf{y}}\left(i\right) = \mathbf{L}\left(\tilde{p}\left(i\right)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \gamma\left(i\right)\mathbf{I}\right)\mathbf{L}^H, \tag{3.35}$$

and the optimization problem in Eq. (3.34) can be cast as

$$\underset{\tilde{p}(i),\tilde{\mathbf{a}},\gamma(i)}{\arg\min} \sum_{i=1}^{N} \log\left|\mathbf{P}_{\mathbf{y}}\left(i\right)\right| + \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}\left(i\right)\mathbf{P}_{\mathbf{y}}^{-1}\left(i\right)\right). \tag{3.36}$$

Substituting Eq. (3.16) and Eq. (3.17) in Eq. (3.36) and omitting the irrelevant constant terms, the cost function can be expressed as

$$\begin{aligned}\underset{\tilde{p}(i),\tilde{\mathbf{a}},\gamma(i)}{\arg\min}\sum_{i=1}^{N}\log&\left[\left(\tilde{p}(i)+\gamma(i)\right)\left(\gamma(i)^{M-1}\right)\right]\\ &+\mathrm{tr}\left(\gamma(i)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}(i)\right)\\ &-\frac{\gamma(i)^{-2}\tilde{p}(i)}{1+\gamma(i)^{-1}\tilde{p}(i)}\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}(i)\,\tilde{\mathbf{a}},\end{aligned}\tag{3.37}$$

where similar manipulations have been carried out as in Eq. (3.18).

To estimate $\tilde{\mathbf{a}}$, we can focus on the last term of Eq. (3.37). Hence, the estimation of $\tilde{\mathbf{a}}$ is the solution of the following optimization problem

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\sum_{i=1}^{N}\left(\frac{\tilde{p}(i)}{\gamma(i)+\tilde{p}(i)}\frac{1}{\gamma(i)}\tilde{\mathbf{a}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}(i)\,\tilde{\mathbf{a}}\right),\tag{3.38}$$

which is the principal eigenvector of the matrix

$$\sum_{i=1}^{N}\frac{\tilde{p}(i)}{\gamma(i)+\tilde{p}(i)}\frac{1}{\gamma(i)}\hat{\mathbf{P}}_{\mathbf{w}}(i)\,.\tag{3.39}$$

Note that unlike the estimation of $\tilde{\mathbf{a}}$ in a single time frame case where the estimate is the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{w}}$, the estimate is now the principal eigenvector of a weighted sum of the whitened CPSD matrices for all time frames and the weights depend on the estimation of $\tilde{p}(i)$ and $\gamma(i)$ for $i=1,\cdots,N$. Therefore, a closed form solution to Eq. (3.38) does not exist and we propose a recursive estimation approach.

For the first step, we estimate the parameters for each time frame independently using the method proposed in Section 3.3.1. In this case, we will obtain N different estimates of the RTF vector, say, $\hat{\tilde{\mathbf{a}}}(i)$, which is the principal eigenvector of $\mathbf{L}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}(i)\,\mathbf{L}^{-H}$ per frame i. Given $\hat{\tilde{\mathbf{a}}}(i)$ for a single frame i, the estimates of $\tilde{p}(i)$ and $\gamma(i)$ are obviously identical to expressions in Eq. (3.23) and Eq. (3.24), that is,

$$\hat{\tilde{p}}(i)=\frac{M\lambda_{\max}(i)-\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(i)\right)}{M-1},\tag{3.40}$$

$$\hat{\gamma}(i)=\frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(i)\right)-\lambda_{\max}(i)}{M-1},\tag{3.41}$$

where $\lambda_{\max}(i)$ is the principal eigenvalue of $\hat{\mathbf{P}}_{\mathbf{w}}(i)$.

For the second step, we use the initial estimates of $\tilde{p}(i)$ and $\gamma(i)$ to calculate the matrix in Eq. (3.39) and then use its principal eigenvector as the estimate of the RTF vector $\hat{\tilde{\mathbf{a}}}$. Next, we use the estimated $\hat{\tilde{\mathbf{a}}}$ in Eq. (3.37) and find new update estimates of $\tilde{p}(i)$ and $\gamma(i)$ based on the estimate $\hat{\tilde{\mathbf{a}}}$ which was found using the joint information across all time frames in a segment. That is,

$$\hat{\tilde{p}}(i)=\frac{M\hat{\tilde{\mathbf{a}}}^{H}\hat{\mathbf{P}}_{\mathbf{w}}(i)\,\hat{\tilde{\mathbf{a}}}-\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(i)\right)}{M-1}\tag{3.42}$$

and

$$\hat{\gamma}(i) = \frac{\mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(i)\right) - \hat{\hat{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{w}}(i)\,\hat{\hat{\mathbf{a}}}}{M-1}. \tag{3.43}$$

Note that $\hat{\hat{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{w}}(i)\,\hat{\hat{\mathbf{a}}} \leq \lambda_{\max}(i) < \mathrm{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(i)\right)$, hence $\hat{\gamma}(i) > 0$. But $\hat{\hat{p}}(i)$ in Eq. (3.42) can become negative. We replace these negative values using the initial estimates from Eq. (3.40) and store their corresponding time frame indices as index set G, which will not be included when calculating the weighted sum in Eq. (3.39) to estimate the RTF vector in the next step.

In the remaining steps, we repeat the second step until the relative change of $\hat{\hat{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{w}}(i)\,\hat{\hat{\mathbf{a}}}$ between the current iteration and the last iteration does not exceed a certain number ε, or a certain number of iterations has been executed.

3.3.3. ROBUST PARAMETER ESTIMATION

In [18], it has been shown that linear inequality constraints on the parameters of interest can be used to improve the robustness of the estimation. Herein, we introduce these constraints on the RTF, the PSD of source and the PSD of the late reverberation. Note that, after obtaining estimates in each step of our proposed method, we can project the estimates into the constraint intervals introduced below. These constraints can effectively avoid large underestimation or overestimation errors and therefore can improve the robustness of our proposed joint MLE for multiple time frames.

CONSTRAINTS FOR THE RTFS

Considering only the direct path component, the anechoic acoustic transfer function (ATF) has the following equation [33]

$$\bar{a}_i = \frac{1}{4\pi d_i} \exp\left(-\frac{j2\pi k d_i}{Kc}\right), \tag{3.44}$$

where c denotes the sound speed, K is the FFT length and d_i is the distance between the source and the i_{th} microphone ($d_i > 0$). The RTF in the k_{th} frequency bin is then given by (with the first microphone selected as the reference microphone)

$$a_i(k) = \frac{d_1}{d_i} \exp\left(-\frac{j2\pi k(d_i - d_1)}{Kc}\right). \tag{3.45}$$

Using Eq. (3.45), for any frequency bin, a tight bound for both the real and imaginary parts of a_i is given by

$$-\frac{d_1}{d_i} \leq \Re(a_i), \Im(a_i) \leq \frac{d_1}{d_i}. \tag{3.46}$$

When not only the direct path component but also the early reflections are considered, the RTF value might exceed the tight bound above and we need to use a looser bound. Observing $d_1 \leq d_{1,i} + d_i$ ($d_{1,i}$ is the distance between the first microphone and the i_{th}

microphone) and assuming $d_i \geq d_{\max}$ (i.e. the distance between the source and each microphone is not smaller than a given small value $d_{\max}$), a looser bound for RTFs is

$$-\frac{d_{1,i}+d_{\max}}{d_{\max}} \leq \Re(a_i), \Im(a_i) \leq \frac{d_{1,i}+d_{\max}}{d_{\max}}. \tag{3.47}$$

Note that after obtaining $\hat{\hat{\mathbf{a}}}$ at each step in our proposed method, we first normalize it with its first element to estimate the RTF vector $\hat{\mathbf{a}}$ and then project the estimated RTF vector into the interval $\left[-\frac{d_{1,i}+d_{\max}}{d_{\max}}, \frac{d_{1,i}+d_{\max}}{d_{\max}}\right]$. Finally, we calculate the reparameterized vector using $\hat{\hat{\mathbf{a}}} = \frac{\mathbf{L}^{-1}\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{a}}^H\boldsymbol{\Gamma}^{-1}\hat{\mathbf{a}}}}$.

CONSTRAINTS FOR THE SOURCE PSDS

In Eq. (3.9), using the fact that $a_1 = 1$ and $\boldsymbol{\Gamma}_{1,1} = 1$, we have

$$\mathbf{P}_{y_{1,1}}(i) = p(i) + \gamma(i). \tag{3.48}$$

Hence, an upper bound for $p(i)$, by using a prefixed constant δ (with $\delta \geq 1$), is found as

$$p(i) \leq \delta \mathbf{P}_{y_{1,1}}(i) - \gamma(i), \tag{3.49}$$

and the upper bound for the reparametrized parameter $\tilde{p}(i,k)$ is

$$\tilde{p}(i) \leq \left[\delta \mathbf{P}_{y_{1,1}}(i) - \gamma(i)\right] \mathbf{a}^H \boldsymbol{\Gamma}^{-1} \mathbf{a}. \tag{3.50}$$

CONSTRAINTS FOR THE LATE REVERBERATION PSDS

As shown in [18], the following constraints can be applied to ensure better speech intelligibility performance by reducing overestimation errors on the PSD of the late reverberation [3], [34]

$$\gamma \leq \min\left[\operatorname{diag}\left(\mathbf{P}_y(i)\right)\right]. \tag{3.51}$$

Since $\boldsymbol{\Gamma}_{m,m} = 1$ for $m = 1, \cdots, M$, we have

$$\mathbf{P}_{y_{m,m}}(i) = p(i)\, a_m a_m^H + \gamma(i), \tag{3.52}$$

where $p(i)\, a_m {a_m}^H$ is positive. Hence we have $\mathbf{P}_{y_{m,m}}(i) \geq \gamma(i)$ for all m and Eq. (3.51) holds.

3.4. EXPERIMENTS

In this section, we evaluate the estimation performance of the proposed methods as well as the performance on noise reduction, speech quality and speech intelligibility. We will first introduce the reference methods in Section 3.4.1 and the evaluation metrics in Section 3.4.2. Then, in Section 3.4.3, we consider a static source scenario and use the simulated room impulse responses (RIRs) to construct the microphone signals. At last, in Section 3.4.4, we consider both the static source scenario and the source-moving scenario and use the RIRs recorded in real life from [35].

3.4.1. Reference methods

Combination of existing methods

The first reference method we consider utilizes several existing methods [2], [13], [14] to estimate the PSD of the late reverberation, the RTF vector and the PSD of the source successively. First, by assuming a noiseless or high SNR scenario, we use the eigenvalue decomposition-based method proposed in [2] to estimate the PSD of the late reverberation. With this estimate, we use the covariance whitening method in [14] to estimate the RTF vector. Finally, we use the method proposed in [13] to estimate the PSD of the source. Note that although this reference method is a combination of existing state-of-the-art methods, this combination has the same estimation steps as the joint MLE estimator for a single time frame presented in Section 3.3.1. Note also that this reference method only considers using the CPSD matrix for a single time frame. Therefore, when dealing with multiple time frames in one time segment, we can either use it to estimate parameters for all time frames independently or averaging the CPSD matrices for all time frames in a time segment and use it to estimate parameters with this averaged CPSD matrix. For convenience, we refer to this first case as ‘Ref1’ and the second case as ‘Ref2’ in each figure.

Simultaneous confirmatory factor analysis

The recently published method in [18] is also used for comparison in all the experiments. This method is based on confirmatory factor analysis (CFA) and non-orthogonal joint diagonalization principles and, hence, is called the simultaneous confirmatory factor analysis (SCFA) method. Note that the SCFA method is very accurate and can estimate the RTF matrix, the PSDs of the early components of the sources, the PSD of the late reverberation, and the PSDs of the microphone-self noise jointly, but, also has high computational complexity. With the SCFA method, the parameters estimation problem is modelled as the following optimization problem

$$\begin{aligned} \begin{matrix} \hat{p}(i), \hat{\mathbf{a}} \\ \hat{\gamma}(i), \hat{\mathbf{P}}_{\mathbf{v}} \end{matrix} = & \underset{\begin{matrix} p(i), \mathbf{a} \\ \gamma(i), \mathbf{P}_{\mathbf{v}} \end{matrix}}{\arg\min} \sum_{i=1}^{N} \log|\mathbf{P}_{\mathbf{y}}(i)| + \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{y}}(i)\mathbf{P}_{\mathbf{y}}^{-1}(i)\right) \\ & \text{s.t. } \mathbf{P}_{\mathbf{y}}(i) = \mathbf{P}_{\mathbf{x}}(i) + \mathbf{P}_{\mathbf{l}}(i) + \mathbf{P}_{\mathbf{v}} \end{aligned} \tag{3.53}$$

where $\mathbf{P}_{\mathbf{x}}(i)$, $\mathbf{P}_{\mathbf{l}}(i)$ and $\mathbf{P}_{\mathbf{v}}$ are defined in Eqs. (3.5), (3.6) and (3.8), respectively. This problem is not a convex problem and the computational complexity is high. In [18], the problem is solved iteratively and the *fmincon* procedure in the standard MATLAB optimization toolbox is used to decrease the value of the cost function in Eq. (3.53) for each iteration. The iteration terminates if a given estimation accuracy is achieved or the iteration number exceeds a certain number.

Although the SCFA method can estimate the RTF matrix and the PSDs jointly, it is computationally not efficient and sometimes may have a wrong estimate because it deals with a non-convex problem and does not assure a global optimal solution. Therefore, a set of “box constraints" is proposed in [18] to improve the robustness of the SCFA method. In our experiments, we used the same constraints as in Eqs. $(27),(38),(39)$ and (40) in [18].

3.4.2. EVALUATION METRICS

In all the experiments, three types of performance comparison between the proposed method and the reference methods are presented. We first compare the estimation error of the parameters of interest. For the RTF vector, we use the Hermitian angle measure (in rad) [36] which is averaged over all frequency bins and time segments

$$E_{\mathbf{a}} = \frac{\sum_{\beta=1}^{B}\sum_{k=1}^{K/2+1} \arccos\left(\frac{|\mathbf{a}(\beta,k)^H \hat{\mathbf{a}}(\beta,k)|}{\|\mathbf{a}(\beta,k)\|_F \|\hat{\mathbf{a}}(\beta,k)\|_F}\right)}{B\,(K/2+1)}. \tag{3.54}$$

For the PSDs of the source and the late reverberation, we use the averaged error (in dB)

$$E_s = \frac{10\sum_{\beta=1}^{B}\sum_{i=1}^{N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{p(i,k)}{\hat{p}(i,k)}\right)\right|}{BN\,(K/2+1)} \tag{3.55}$$

and

$$E_\gamma = \frac{10\sum_{\beta=1}^{B}\sum_{i=1}^{N}\sum_{k=1}^{K/2+1}\left|\log\left(\frac{\gamma(i,k)}{\hat{\gamma}(i,k)}\right)\right|}{BN\,(K/2+1)}, \tag{3.56}$$

where $|\cdot|$ denotes taking the absolute value in Eqs. (3.54) to (3.56).

Then, we provide the speech intelligibility and quality comparison among the estimated sources constructed using parameters that are obtained by different methods. That is, we use estimated parameters to calculate the following multi-channel Wiener filter (MWF)

$$\hat{\mathbf{w}} = \frac{\hat{p}}{\hat{p} + \hat{\mathbf{w}}_{\text{MVDR}}^H \hat{\mathbf{R}}_{nn} \hat{\mathbf{w}}_{\text{MVDR}}} \hat{\mathbf{w}}_{\text{MVDR}}, \tag{3.57}$$

where $\mathbf{w}_{\text{MVDR}}$ is the minimum variance distortionless response (MVDR) beamformer [37]

$$\hat{\mathbf{w}}_{\text{MVDR}} = \frac{\hat{\mathbf{R}}_{nn}^{-1}\hat{\mathbf{a}}}{\hat{\mathbf{a}}^H \hat{\mathbf{R}}_{nn}^{-1}\hat{\mathbf{a}}}, \tag{3.58}$$

and

$$\hat{\mathbf{R}}_{nn} = \hat{\gamma}\hat{\mathbf{\Gamma}}. \tag{3.59}$$

Note that $\hat{\mathbf{\Gamma}}$ is calculated by Eq. (3.7) for all methods by assuming the distance between each microphone pair is known. For the SCFA method, we set $\hat{\mathbf{R}}_{nn} = \hat{\gamma}\hat{\mathbf{\Gamma}} + \hat{\mathbf{P}}_v$, since SCFA can provide an estimate of the PSD of the microphone self noise.

After reconstructing the estimated sources, we use the segmental signal-to-noise-ratio (SSNR) [38] to measure the noise reduction performance. In addition, we compare the speech intelligibility performance using the speech intelligibility in bits (SIIB) measure [39], [40]. The speech-to-reverberation modulation energy ratio (SRMR) measure [41] is also calculated in each scenario to demonstrate the speech quality and intelligibility of all reconstructed sources.

Finally, we compare the computation time between our proposed method and the reference methods.

3.4.3. EXPERIMENTS WITH SIMULATED RIRS

SETUP

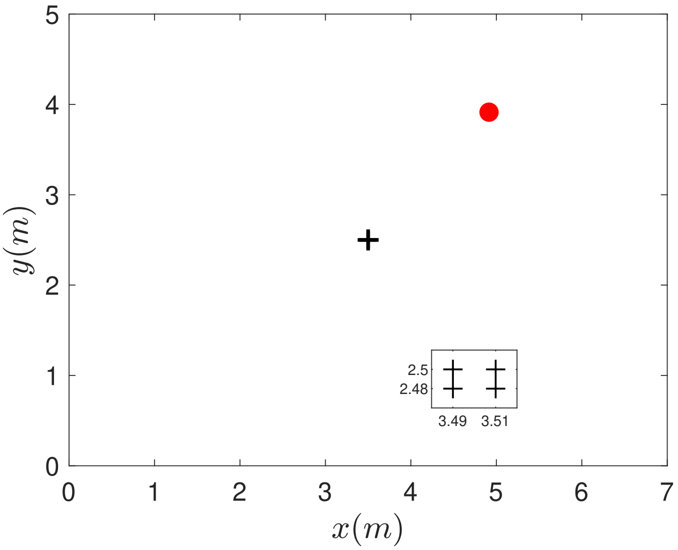
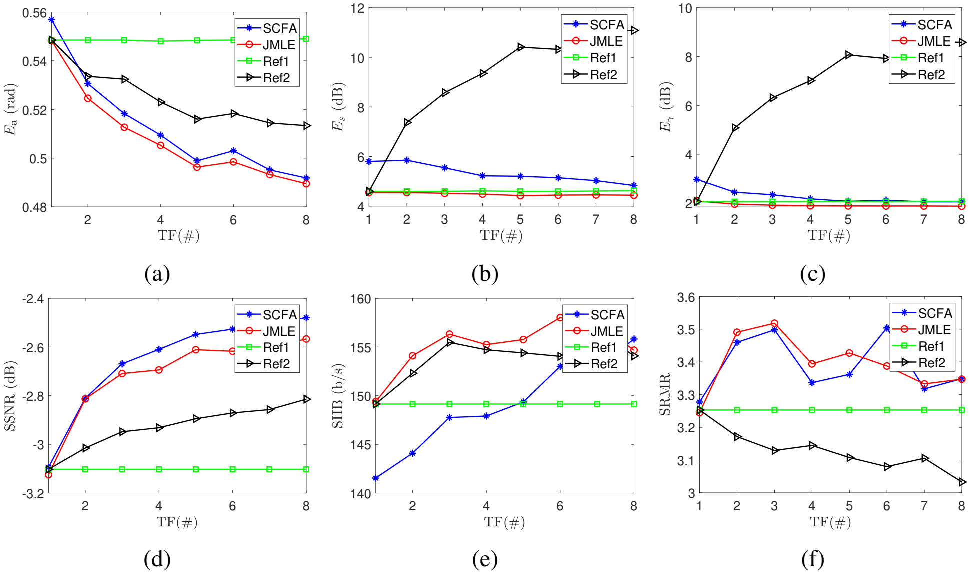


Figure 3.3: Top view of the acoustic scene. The red circle denotes the source. The cross in the center denotes the set of microphones. A zoom-in of that set of four microphones is provided in the little square.

To simulate room impulse responses from source to microphones, we use the image source method [33]. The four microphone signals are then constructed by convolving the speech source (with a duration of 35 s) with each of the four room impulse responses corresponding to each microphone. The positions of four microphones and the position of the source are shown in Fig. 3.3, and the dimensions of the simulated room are set to $7 \times 5 \times 4$ m. Since we used the SCFA method as a reference method, the parameters used in the experiments are similar to those used in [18]. Subsequently, microphone self-noise is simulated by adding realizations of a zero-mean uncorrelated Gaussian process with variance σ_v^2, such that the SNR per microphone due to the self-noise is equal to the values as specified in each figure. Note that since we consider only the microphone self-noise, the noise energy is relatively low resulting in large SNR values of about 50 dB. The sampling frequency is $f_s = 16$ kHz. Per sub-time frame, the sampled noisy microphone signals are converted to the frequency domain using the STFT procedure, where the sub-time frames are windowed with a square-root Hann window with a length of 512 samples (i.e. 32 ms) and an overlap of 50% between sub-time frames. The true RTF is set to the early reflections of the room impulse response, which is set here as the 512-length FFT of the first 512 samples of the room impulse responses, as this equals the early part (first 32 ms) of the impulse response that falls within a single sub-frame. Each time frame consists of $N_s = 40$ overlapped sub frames. The prefixed parameters are $\delta = 1.1$ and $d_{\max} = 0.02$ (i.e. the distance between each microphone and the source is larger than 0.02 m).

RESULTS

In Fig. 3.4, we fix the reverberation time T_{60} at 1 s and obtain noisy speech with the SNR fixed at 50 dB. We change the number of time frames in a time segment from 1 to 8. The CPSD matrix of the microphone self noise is subtracted from the noisy CPSD matrix for JMLE, Ref1 and Ref2 in this scenario. The performance comparison among JMLE and the other three reference methods is shown in Fig. 3.4 as the number of time frames used in each time segment changes from 1 to 8. When using only one time frame, JMLE, Ref1 and Ref2 have exactly the same estimates of the RTF and the PSDs of the source and the late reverberation as expected and their estimation performance is better than SCFA. When the number of time frames in a time segment increases, the



(a)
(b)
(c)
(d)
(e)
(f)

Figure 3.4: Performance vs the number of time frames.

RTF estimation performance for Ref1 nearly does not change since this method always uses each time frame independently and does not use the prior information that the RTF is constant for all time frames in a time segment. However, for JMLE, SCFA and Ref2, the estimation error of the RTF decreases with the increase of the number of time frames in a time segment. For a larger number of time frames, i.e. a longer segment, among these three methods, JMLE and SCFA have similar performance, and both notably outperform Ref2. The PSD estimation performance for JMLE, SCFA and Ref1 does not change much since the PSDs can differ time-frame by time-frame. However, the PSD estimation performance for Ref2 decreases when the number of time frames increases because Ref2 assumes the source is stationary during a time segment, which is mostly not true in a practical scene. For the noise reduction performance and the speech quality and intelligibility performance, we can see that JMLE and SCFA have larger SSNR, SIIB and SRMR values compared to the other two reference methods in most cases.

3.4.4. EXPERIMENTS WITH RECORDED RIRS

The performance of all methods is now compared using recorded room impulse responses from [35]. The reverberation time of the RIRs include 0.36 s and 0.61 s. The positions of the microphones and the position of the source used to record the impulse responses are shown in Fig. 3.5. The source is placed at a distance of 2 m from the center of the uniform linear microphone array of 8 microphones which have inter-distances of 8 cm. Although the angles of the source include $\{-90°,-75°,\cdots,90°\}$ in [35], we use only $\{0°,15°,30°,45°,60°\}$ in this work. We will first consider a static source scenario and evaluate the performance for various SNR values. Then, we will show the influence on the estimation performance of all methods when the source position changes at specific moments.

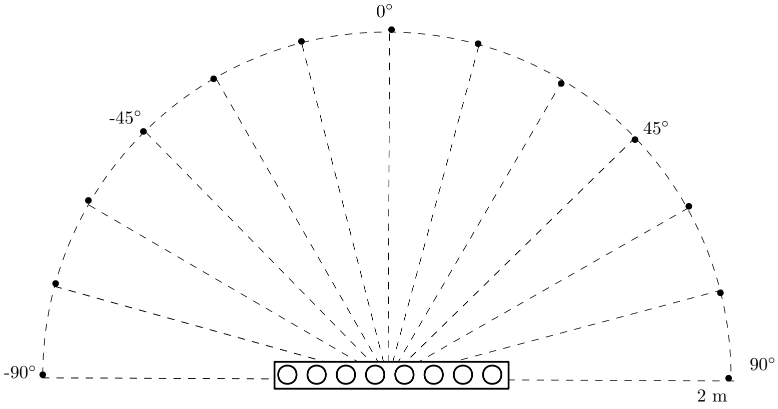
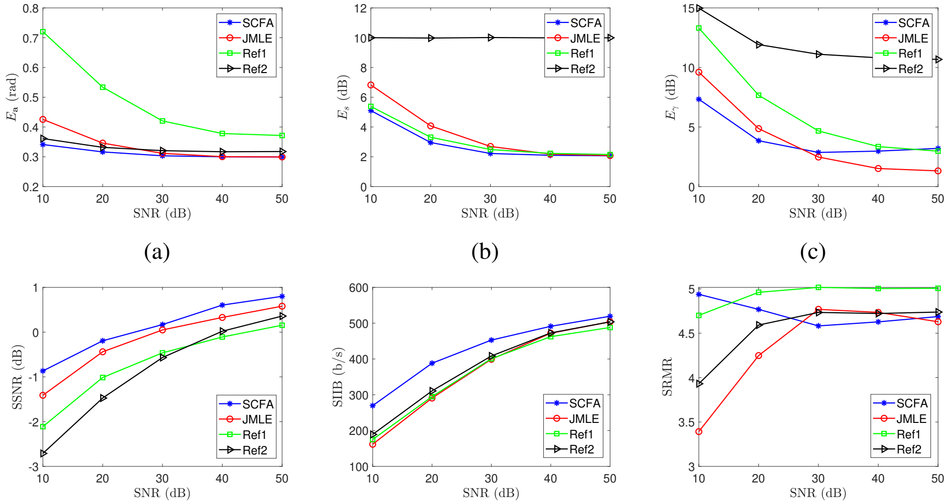


Figure 3.5: Setup for the real RIRs.



(a) (b) (c)

(d) (e) (f)

Figure 3.6: Performance vs SNR.

STATIC SOURCE

For the static source scenario, we use the RIRs for the source position fixed at 0° and the reverberation time of 0.61 s. We obtain noisy speech with the SNR simulating the microphone self noise ranging from 10 dB to 50 dB. Notice that realistic values for microphone self noise are in the order of 40 to 50 dB. Each time segment contains 8 time frames. Note that in this scenario, the prior information of the microphone self noise is used by none of the methods and for JMLE, Ref1 and Ref2, we simply ignore the microphone self noise and use the CPSD matrix of the noisy signal directly.

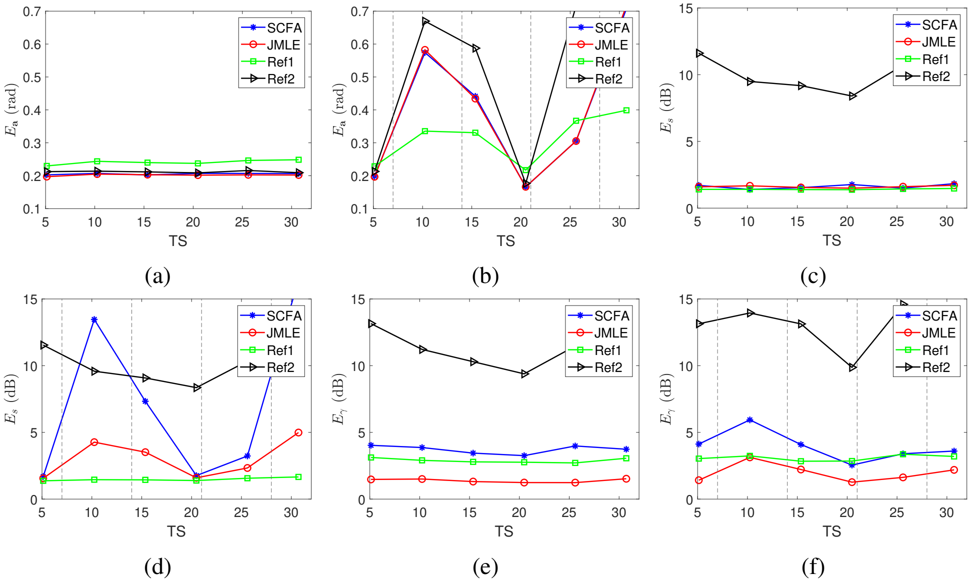
The performance comparison among JMLE and the other three reference methods is shown in Fig. 3.6 as the SNR increases from 10 dB to 50 dB. As shown in Fig. 3.6, JMLE and SCFA outperform Ref1 in the RTF estimation performance and outperform Ref2 in the PSDs estimation performance (of the source and the late reverberation). As the SNR becomes larger, all methods have both better RTF and PSD estimation performance. However, JMLE shows the most significant improvement compared to the other methods. For the noise reduction performance and the speech quality and intelligibility performance, JMLE and SCFA still outperform the other two reference methods.

Table 3.1: Computation time comparison.

method	SCFA	JMLE	Ref1	Ref2
Normalized run time	1310	19	6	1

In Table 3.1 we show the normalized computation time comparison among all methods, where we have averaged the run time over all cases for each method. As expected, SCFA needs significantly more time compared to the other three methods. The computational cost of the proposed method using multiple time frames mainly comes from the calculation of the eigenvalue decomposition of an $M \times M$ matrix in each iteration, which has a complexity of order M^3. The total complexity order is thus $(N+N_i)M^3$ with one initial step and N_i iterative steps. Similarly, for Ref1, its complexity order is NM^3 with N the number of time frames in a time segment. For Ref2, its complexity order is M^3. Therefore, the time cost ratio among JMLE, Ref1 and Ref2 is $N_i+N : N : 1 = 18 : 8 : 1$, which is similar to the real averaged run time ratio in Table 3.1. Note that the proposed method using multiple time frames can be initialized by either Ref1 or Ref2. In this work, we present only using Ref1 as the initialization step. If the Ref2 is used as the initialization, the complexity order of JMLE will be $(N_i+1)M^3$.

MOVING SOURCE



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.7: Performance vs time segments (TS) with the reverberation time 0.36 s.

For the moving source scenario, we place the source at 0° and change the position to 60° in steps of 15° every 7 s. Since each time frame contains 40 sub-time frames of 32 ms taken with 50% overlap and each time segment contains 8 time frames, the time segment duration is about 5.12 s. The 35 s speech is divided into 6 complete time segments (the last incomplete time segment is not used). Only the microphone signals during the first and the fourth time segments are received from a single source position.

In all other segments, the source position changes during the segment. We evaluate the estimation performance of all methods for per time segment.

In Fig. 3.7, the reverberation time is 0.36 s. For comparison, we show the estimation performance of all methods when the source position is fixed at 0° in Figs. 3.7a, 3.7c and 3.7e. As shown, the estimation performance of all methods does not change much for different time segments, except the poor PSDs estimation performance of the Ref2 method. In Figs. 3.7b, 3.7d and 3.7f, we show the estimation performance of all methods when the source position is moved from 0° to 60° by 15° every 7 s. The vertical dashed lines in these figures denote the time point when the source position is changed. As shown, the estimation performance during the first and the fourth time segments is best among others for the methods using multi-time frames in their estimation as during these time segments, the source position is fixed while during other time segments the source position is changed. The RTF estimation performance is influenced the most while the late reverberation PSD estimation performance is influenced the least by source position change. The reason is that the RTF contains information on the source position, while the late reverberation can be considered as a diffuse noise field. For the Ref1 method, its estimation performance is not affected much since it estimates the parameters frame by frame instead of segment by segment and only four time frames are affected by source position change.

3.5. CONCLUDING REMARKS

We considered the problem of estimating the RTFs, the PSDs of the source and the PSDs of the late reverberation jointly for a single source scenario. We first proposed a joint maximum likelihood estimator (JMLE) using a single time frame, which has a closed form solution and can be solved efficiently. Then, we proposed a joint MLE using multiple time frames that share the same RTF and achieved similar estimation accuracy, together with the performance of noise reduction, speech quality and speech intelligibility, compared to the SCFA method, which both outperform the other reference methods combining several existing state-of-the-art methods. Moreover, it is also shown that the proposed JMLE for multiple time frames has a much lower computational complexity than that of the SCFA method.

REFERENCES

[1] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[2] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[3] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[4] O. Schwartz, S. Gannot, and E. A. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm", in *Proc. EURASIP Eur. Signal Process. Conf.*, 2016, pp. 1123–1127.

[5] S. Markovich, S. Gannot, and I. Cohen, "Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals", *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[6] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 137–152, 2017.

[7] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, "Microphone subset selection for MVDR beamformer based noise reduction", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 550–563, 2018.

[8] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks With Arbitrary Topology", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1434–1448, 2018.

[9] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics", *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.

[10] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources", *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[11] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Frequency-domain blind source separation of many speech signals using near-field and far-field models”, *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[12] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, “Informed sound source localization using relative transfer functions for hearing aid applications”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[13] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids”, in *Proc. EURASIP Eur. Signal Process. Conf.*, 2014, pp. 61–65.

[14] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[15] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[16] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[18] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust joint estimation of multimicrophone signal model parameters”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[19] J. S. Bradley, H. Sato, and M. Picard, “On the importance of early reflections for speech in rooms”, *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[20] R. Talmon, I. Cohen, and S. Gannot, “Relative Transfer Function Identification Using Convolutive Transfer Function Approximation”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[21] Y. Avargel and I. Cohen, “System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[22] S. Gazor and W. Zhang, “Speech probability distribution”, *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.

[23] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors”, *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, 2005.

[24] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, “A study of the distribution of time-domain speech samples and discrete Fourier coefficients”, in *Proc. SPS-DARTS*, vol. 1, 2005, pp. 155–158.

[25] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[26] S. Braun and E. A. Habets, “Dereverberation in noisy environments using reference signals and a maximum likelihood estimator”, in *Proc. EURASIP Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[27] T. Lotter and P. Vary, “Dual-channel speech enhancement by superdirective beamforming”, *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–14, 2006.

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

[29] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[30] B. F. Cron and C. H. Sherman, “Spatial-correlation functions for various noise models”, *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[31] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection”, *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[32] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*, Version 20121115, Nov. 2012.

[33] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[34] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[35] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments”, in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sep. 2014.

[36] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”, in *Proc. IEEE Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 11–15.

[37] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

[38] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[39] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility metric based on information theory”, *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, 2017.

[40] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An evaluation of intrusive instrumental intelligibility metrics”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[41] T. H. Falk, C. Zheng, and W.-Y. Chan, “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

4

NOISE PSD INSENSITIVE RTF ESTIMATION IN A REVERBERANT AND NOISY ENVIRONMENT

This chapter is based on the article published as: C. Li and R. C. Hendriks. "Noise PSD Insensitive RTF Estimation in a Reverberant and Noisy Environment". In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023, pp. 1–5.

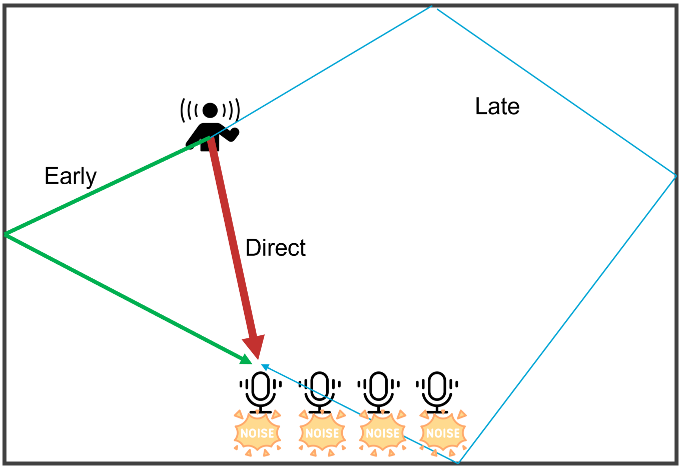


Figure 4.1: Illustration of a single source, reverberant and noisy scenario.

In the previous chapter, we considered a reverberant but noiseless environment. In this work, we include the noise component and consider a single source reverberant and noisy scenario as illustrated in Fig. 4.1. With this chapter, we will answer research question 1.2 shown in Fig. 1.4 using the signal model presented in Fig. 2.2 (b).

Spatial filtering techniques typically rely on estimates of the target relative transfer function (RTF). However, the target speech signal is typically corrupted by late reverberation and ambient noise, which complicates RTF estimation. Existing methods subtract the noise covariance matrix to obtain the target-plus-late reverberation covariance matrix, from where the RTF can be estimated. However, the noise covariance matrix is typically unknown. More specifically, the noise power spectral density (PSD) is typically unknown, while the spatial coherence matrix can be assumed known as it might remain time-invariant for a longer time. Using the spatial coherence matrices we simplify the signal model such that the off-diagonal elements are not affected by the PSDs of the late reverberation and the ambient noise. Then we use these elements to estimate the target covariance matrix, from where the RTF can be obtained. Hence, the resulting estimate of the RTF is insensitive to the noise PSD. Experiments demonstrate the estimation performance of our proposed method.

4.1. INTRODUCTION

Microphone arrays are widely used for hands-free speech communication applications such as mobile phones and hearing aids. Spatial filtering techniques like the minimum variance distortionless response (MVDR) beamformer [1], [2] and the multichannel Wiener filter (MWF) [2], [3] are often used to extract target signals from the noisy microphone recordings typically corrupted by reverberation and ambient noise. However, these filters critically rely on knowing the relative transfer functions (RTFs) from source to microphones. In the previous chapter, we assumed a noiseless environment. However, in practice, the environment is usually noisy. Therefore, in this chapter, we address the RTF estimation problem of a single source in a reverberant and noisy environment.

Several RTF estimation methods have been proposed in recent years [4]–[11], including the covariance subtraction (CS) method [7]–[9] and the covariance whitening (CW) method [8]–[10]. In reverberant and noisy environments, these methods require the noise and late reverberation covariance matrices to be known. The CW method subtracts the noise covariance matrix from the noisy covariance matrix prior to whitening by the late reverberation covariance matrix. However, the noise covariance matrix is usually unknown. In this chapter, we model the noise covariance matrix as a time-varying noise PSD multiplied by a time-invariant spatial coherence matrix. In that case, the noise PSD is assumed unknown, but the spatial coherence matrix can be assumed known as it might remain time-invariant for a longer time. Under this relaxed assumption, we propose a method to estimate the RTF in a reverberant and noisy environment, which avoids using the noise PSD and is insensitive to noise PSD estimation errors.

4.2. PRELIMINARIES

4.2.1. SIGNAL MODEL

We consider the problem of estimating the RTFs of a single acoustic source in a reverberant and noisy environment using an array of M microphones with an arbitrary configuration. In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone is given by

$$y_m(l,k) = x_m(l,k) + r_m(l,k) + v_m(l,k), \tag{4.1}$$

with l the time-frame index, k the frequency bin index, and m the microphone index. Let x_m denote the speech including the direct and early reflections of the source. Let r_m denote the late reverberation including all the late reflections of the source, which can be considered diffuse. Further, v_m denotes the ambient noise component and microphone self-noise. The early speech component can be modelled as

$$x_m(l,k) = a_m(l,k)\, s(l,k), \tag{4.2}$$

with $a_m(l,k)$ the RTF of the source from the reference microphone to the m-th microphone. Without loss of generality, we select in this work the first microphone as the reference microphone, which means $a_1 = 1$. Stacking all M microphone signals $\{y_m\}_{m=1}^{M}$ into a vector, we have

$$\mathbf{y}(l,k) = \mathbf{a}(l,k)\,s(l,k) + \mathbf{r}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1}. \tag{4.3}$$

Assuming the three components in Eq. (4.3) to be mutually uncorrelated, the noisy covariance matrix is given by

$$\mathbf{P}_{\mathbf{y}}(l,k) \triangleq \mathbf{P}_{\mathbf{x}}(l,k) + \mathbf{P}_{\mathbf{r}}(l,k) + \mathbf{P}_{\mathbf{v}}(l,k), \tag{4.4}$$

where $\mathbf{P}_{\mathbf{q}} \triangleq E\left\{\mathbf{q}\mathbf{q}^H\right\}$ for $\mathbf{q} = \mathbf{y}, \mathbf{x}, \mathbf{r}$ or $\mathbf{v}$ with $E\{\cdot\}$ the expectation. From Eq. (4.2), we have

$$\mathbf{P}_{\mathbf{x}}(l,k) = \phi_s(l,k)\,\mathbf{a}(l,k)\,\mathbf{a}^H(l,k), \tag{4.5}$$

with $\phi_s(l,k)$ the PSD of the source at the reference microphone. For the late reverberation, we adopt the commonly used model from [12]

$$\mathbf{P}_{\mathbf{r}}(l,k) = \phi_\gamma(l,k)\,\mathbf{\Gamma}(k), \tag{4.6}$$

where $\phi_\gamma(l,k)$ is the unknown PSD of the late reverberation and $\mathbf{\Gamma}(k)$ is the non-singular and known spatial coherence matrix which can be calculated using the microphone array geometry [13]. For the residual noise, we assume its covariance matrix has a similar form, i.e.,

$$\mathbf{P}_{\mathbf{v}}(l,k) = \phi_v(l,k)\,\mathbf{\Psi}(k), \tag{4.7}$$

where $\phi_v(l,k)$ is the unknown PSD and $\mathbf{\Psi}(k)$ is the known spatial coherence matrix.

4.2.2. PROBLEM FORMULATION

Using Eqs. (4.5) to (4.7), we can formulate the noisy covariance matrix as

$$\mathbf{P}_{\mathbf{y}}(l,k) = \phi_s(l,k)\,\mathbf{a}(l,k)\,\mathbf{a}^H(l,k) + \phi_\gamma(l,k)\mathbf{\Gamma}(k) + \phi_v(l,k)\mathbf{\Psi}(k). \tag{4.8}$$

We assume the microphone signals to be stationary over a frame consisting of L_s sub-time frames, indexed by l_s, and estimate $\mathbf{P}_{\mathbf{y}}(\ell,k)$ for one frame using the sample covariance matrix $\hat{\mathbf{P}}_{\mathbf{y}}(\ell,k) = 1/L_s \sum_{l_s=1+(l-1)L_s}^{lL_s} \mathbf{y}(l_s,k)\,\mathbf{y}^H(l_s,k)$.

The aim of this work is to estimate the RTF vector $\mathbf{a}(\ell,k)$ using the estimated covariance matrix $\hat{\mathbf{P}}_{\mathbf{y}}(\ell,k)$ and the known spatial coherence matrices $\mathbf{\Gamma}(k)$ and $\mathbf{\Psi}(k)$, while the PSDs $\phi_s(\ell,k), \phi_\gamma(\ell,k)$, and $\phi_v(\ell,k)$ are all unknown. Prior to presenting our proposed method in Section 4.4, we summarize in Section 4.3 the CW method from [10] that is meant to estimate $\mathbf{a}(\ell,k)$ assuming the complete $\mathbf{P}_{\mathbf{v}}(\ell,k)$ is known instead of only $\mathbf{\Psi}(k)$. For notational simplicity, we omit the frequency and time indices as all processing will be done per time-frequency bin independently.

4.3. STATE OF THE ART AND MOTIVATION

Existing methods for RTF estimation include the covariance subtraction (CS) method and the covariance whitening (CW) method. The CW method has been shown to outperform the CS method [8], [9]. Therefore, we introduce here only the CW method. To use the CW method, we need to assume the covariance matrix of the noise $\mathbf{P}_{\mathbf{v}}$ is given, and subtract it from the noisy covariance matrix $\mathbf{P}_{\mathbf{y}}$, that is

$$\mathbf{P}_{\mathbf{x}+\mathbf{r}} = \mathbf{P}_{\mathbf{y}} - \mathbf{P}_{\mathbf{v}} = \phi_s \mathbf{a}\mathbf{a}^H + \phi_\gamma \mathbf{\Gamma}. \tag{4.9}$$

With the signal model from Eq. (4.9), the CW method can estimate the RTF vector in three steps. First, it whitens the noisy signal using $\mathbf{\Gamma}^{\frac{1}{2}}$, which is the principal square-root of the spatial coherence matrix $\mathbf{\Gamma}$ satisfying $\mathbf{\Gamma} = \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{\Gamma}^{\frac{H}{2}}$ with $\mathbf{\Gamma}^{\frac{H}{2}}$ the Hermitian transpose of $\mathbf{\Gamma}^{\frac{1}{2}}$. Note that the square-root is not unique and in this work, we use the Cholesky decomposition. The covariance matrix after whitening has the form

$$\mathbf{P}_w = \mathbf{\Gamma}^{-\frac{1}{2}} \mathbf{P}_{\mathbf{x}+\mathbf{r}} \mathbf{\Gamma}^{-\frac{H}{2}} = \phi_s \mathbf{a}_w \mathbf{a}_w{}^H + \phi_\gamma \mathbf{I}, \tag{4.10}$$

where $\mathbf{a}_w = \mathbf{\Gamma}^{-\frac{1}{2}}\mathbf{a}$ is a scaled version of the principal eigenvector of $\mathbf{P}_w$. Hence the second step is to take the eigenvalue decomposition of $\hat{\mathbf{P}}_w$ and find its principal eigenvector $\mathbf{u}$. The last step is to estimate the RTF vector by

$$\hat{\mathbf{a}} = \frac{\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{u}}{\mathbf{e}^T \mathbf{\Gamma}^{\frac{1}{2}}\mathbf{u}}, \tag{4.11}$$

where $\mathbf{e} = [1, 0, \cdots, 0]^T$.

A weakness of the CW method is that it needs to assume the covariance matrix of the ambient noise is known and subtracted. Subtracting an estimated noise covariance matrix $\hat{\mathbf{P}}_{\mathbf{v}} = \hat{\phi}_v \mathbf{\Psi}$, the covariance matrix after whitening becomes

$$\mathbf{P}_w = \phi_s \mathbf{a}_w \mathbf{a}_w{}^H + \phi_\gamma \mathbf{I} + \Delta\phi_v \mathbf{\Gamma}^{-\frac{1}{2}} \mathbf{\Psi} \mathbf{\Gamma}^{-\frac{H}{2}}, \tag{4.12}$$

with $\Delta\phi_v$ the noise PSD estimation error. Here, $\mathbf{a}_w$ is no longer a scaled principal eigenvector of $\mathbf{P}_w$. Hence, inaccuracies in $\hat{\phi}_v$ will lead to significant estimation errors in $\hat{\mathbf{a}}$.

4.4. PROPOSED METHOD

For the case that not the complete covariance matrix of the ambient noise is known, but only $\mathbf{\Psi}$, we propose an alternative way to estimate the RTF vector by using the off-diagonal elements of a simplified covariance matrix. The proposed method will be less sensitive to estimation errors due to variations in the noise PSD ϕ_v. Note that the technique using only off-diagonal elements of a matrix was used before in [14] for the PSDs estimation and in [15] for radio telescope arrays.

4.4.1. PARAMETER IDENTIFIABILITY

Before using any estimation methods, the identifiability condition that the number of equations is equal or larger than the number of unknowns should be satisfied [16]. Since $\mathbf{P}_y$ is a Hermitian matrix, in Eq. (4.8) there are M^2 knowns (taking Hermitian symmetry and complex values of the data into account). Since $a_1 = 1$, there are $2(M-1)$ unknowns due to the complex-valued $\mathbf{a}$ and there are 3 unknown real-valued PSDs. Therefore, we have altogether the necessary condition

$$M^2 \geq 2(M-1)+3, \tag{4.13}$$

which means $M \geq \sqrt{2}+1$. Noticing that M should be an integer value, we have $M \geq 3$.

4.4.2. SIMPLIFICATION

In Eq. (4.8), since the spatial coherence matrices $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ are assumed to be known, we can simplify the signal model by using the square-root decomposition (e.g. the Cholesky decomposition) of $\mathbf{\Psi} = \mathbf{\Psi}^{\frac{1}{2}}\mathbf{\Psi}^{\frac{H}{2}}$

$$\tilde{\mathbf{P}}_{\mathbf{y}} = \mathbf{\Psi}^{-\frac{1}{2}}\mathbf{P}_{\mathbf{y}}\mathbf{\Psi}^{-\frac{H}{2}} = \phi_s\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \phi_\gamma\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Psi}^{-\frac{H}{2}} + \phi_v\mathbf{I}, \tag{4.14}$$

and the eigenvalue decomposition (EVD) of $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{\Psi}^{-\frac{H}{2}} = \mathbf{U}\mathbf{\Lambda}_\gamma\mathbf{U}^H$, such that

$$\bar{\mathbf{P}}_y = \mathbf{U}^H\tilde{\mathbf{P}}\mathbf{U} = \underbrace{\phi_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H}_{\bar{\mathbf{P}}_x} + \phi_\gamma\mathbf{\Lambda}_\gamma + \phi_v\mathbf{I}, \tag{4.15}$$

where $\bar{\mathbf{a}} = \mathbf{U}^H\tilde{\mathbf{a}} = \mathbf{U}^H\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{a}$.

4.4.3. COVARIANCE MATRIX RECONSTRUCTION

The simplified covariance matrix in Eq. (4.15) is now a summation of a rank-1 matrix $\bar{\mathbf{P}}_x$ and a diagonal matrix $\phi_\gamma\mathbf{\Lambda}_\gamma + \phi_v\mathbf{I}$. Hence, the elements of $\bar{\mathbf{P}}_y$ have the form

$$\bar{\mathbf{P}}_{\mathbf{y}\{i,j\}} = \begin{cases} \phi_s|\bar{a}_m|^2 + \phi_\gamma\lambda_m + \phi_v & i=j=m \\ \phi_s\bar{a}_i\bar{a}_j^* & i \neq j \end{cases}, \tag{4.16}$$

where λ_m is the $\{m,m\}$-th element of $\mathbf{\Lambda}_\gamma$. From Eq. (4.16), we know that the off-diagonal elements of $\bar{\mathbf{P}}_x$ are equal to the corresponding off-diagonal elements of $\bar{\mathbf{P}}_y$, i.e.,

$$\bar{\mathbf{P}}_{\mathbf{x}\{i,j\}} = \bar{\mathbf{P}}_{\mathbf{y}\{i,j\}} \text{ for } i \neq j. \tag{4.17}$$

Therefore, in order to estimate $\bar{\mathbf{P}}_x$ by $\hat{\bar{\mathbf{P}}}_x$ prior to calculating $\mathbf{a}$, we first have to estimate the diagonal elements of $\bar{\mathbf{P}}_x$ as the off diagonal elements are already known from $\hat{\bar{\mathbf{P}}}_y$. From now on we will use the estimated covariance matrix $\hat{\bar{\mathbf{P}}}_y$ and show that we can use the off-diagonal elements of $\hat{\bar{\mathbf{P}}}_y$ to estimate the diagonal elements of $\bar{\mathbf{P}}_x$.

For the m_p-th diagonal element, we can select any 2 other microphones m_q, m_r from the remaining $M-1$ microphones and obtain the following estimates

$$\widehat{\phi_s \left|\bar{a}_{m_p}\right|^2} \approx \frac{\hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_p,m_q\}} \hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_r,m_p\}}}{\hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_r,m_q\}}} = \frac{\widehat{\phi_s \bar{a}_{m_p} \bar{a}^*_{m_q}} \widehat{\phi_s \bar{a}_{m_r} \bar{a}^*_{m_p}}}{\widehat{\phi_s \bar{a}_{m_r} \bar{a}^*_{m_q}}}, \tag{4.18}$$

or

$$\widehat{\phi_s \left|\bar{a}_{m_p}\right|^2} \approx \frac{\hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_p,m_r\}} \hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_q,m_p\}}}{\hat{\bar{\mathbf{P}}}_{\mathbf{y}\{m_q,m_r\}}} = \frac{\widehat{\phi_s \bar{a}_{m_p} \bar{a}^*_{m_r}} \widehat{\phi_s \bar{a}_{m_q} \bar{a}^*_{m_p}}}{\widehat{\phi_s \bar{a}_{m_q} \bar{a}^*_{m_r}}}. \tag{4.19}$$

Since $\hat{\bar{\mathbf{P}}}_{\mathbf{y}}$ is Hermitian, Eq. (4.19) is the conjugate of Eq. (4.18). By taking the average of Eq. (4.19) and Eq. (4.18), one can insure a real valued estimate of $\hat{\bar{\mathbf{P}}}_{x\{m_p,m_p\}}$.

The choice of m_q and m_r should satisfy that $m_q \neq m_r \neq m_p$ and $1 \leq m_q, m_r \leq M$. Therefore, there are $(M-1)(M-2)$ different estimates of $\phi_s \left|\bar{a}_{m_p}\right|^2$, say the set $\mathbb{L}$. We find all the estimates and take their mean value as the final estimate of $\hat{\bar{\mathbf{P}}}_{x\{m_p,m_p\}}$, that is,

$$\hat{\bar{\mathbf{P}}}_{x\{m_p,m_p\}} = \frac{1}{(M-1)(M-2)} \sum_{\forall \widehat{\phi_s \left|\bar{a}_{m_p}\right|^2} \in \mathbb{L}} \widehat{\phi_s \left|\bar{a}_{m_p}\right|^2} \tag{4.20}$$

4.4.4. RTF ESTIMATION

Since $\bar{\mathbf{P}}_x = \phi_s \bar{\mathbf{a}} \bar{\mathbf{a}}^H$, we can estimate a scaled version of $\bar{\mathbf{a}}$ by the principal eigenvector of $\hat{\bar{\mathbf{P}}}_x$ denoted as $\mathbf{u}$. From $\bar{\mathbf{a}} = \mathbf{U}^H \mathbf{\Psi}^{-\frac{1}{2}} \mathbf{a}$ and $a_1 = 1$, we can estimate the RTF by

$$\hat{\mathbf{a}} = \frac{\mathbf{\Psi}^{\frac{1}{2}} \mathbf{U} \mathbf{u}}{\mathbf{e}^T \mathbf{\Psi}^{\frac{1}{2}} \mathbf{U} \mathbf{u}}. \tag{4.21}$$

4.5. EXPERIMENTS

To verify the performance of our proposed method, we simulate a room with dimension $7 \times 5 \times 4$ m and place a speech source as well as 10 microphones in the room forming a line array, as depicted in Fig. 4.2. Note that for some experiments, only the first a few microphones are used from left to right. The signal received at each microphone is a convolution between the speech source and the corresponding room impulse response. The room impulse responses are simulated by the image source method [17]. Moreover, we calculate the spatial coherence matrix of the late reverberation by assuming a spherically diffuse sound field, i.e., $\mathbf{\Gamma}_{i,j}(k) = \operatorname{sinc}\left(\frac{2\pi f_s k}{K} \frac{d_{i,j}}{c}\right)$, with $\operatorname{sinc}(x) = \sin x / x$, $d_{i,j}$ the inter-distance between microphones i and j, f_s the sampling frequency, c the speed of sound and K the number of frequency bins. The spatial coherence matrix of the ambient noise is set to the identity matrix, i.e. $\mathbf{\Psi} = \mathbf{I}$ simulating microphone self-noise by a

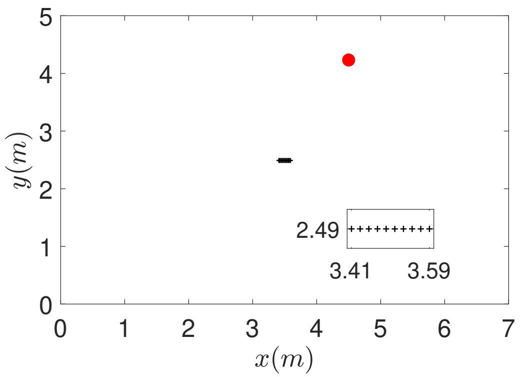
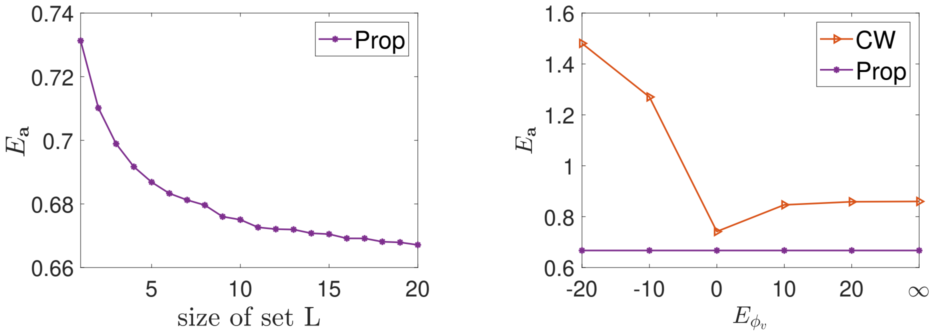


Figure 4.2: Top view of the acoustic scene with a zoom-in of microphones. The source is denoted by the red circle.

zero-mean uncorrelated Gaussian process with the same variance for each microphone. The noisy microphone signals are sampled at a frequency of $f_s = 16$ kHz and processed by the STFT procedure including windowing and FFT. We use a square-root Hann window with a duration of 12.5 ms and an overlap of 75% between two adjacent time frames. The FFT length is 256. The true RTF is calculated by 256-length FFT of the first 200 samples of the room impulse responses. The RTF estimation error is evaluated by the Hermitian angle measure (in rad) [6]

$$E_{\mathbf{a}} = \frac{\sum\limits_{\ell=1}^{L}\sum\limits_{k=1}^{K/2+1} \mathrm{acos}\left(\frac{\left|\mathbf{a}^H(\ell,k)\hat{\mathbf{a}}(\ell,k)\right|}{\left\|\mathbf{a}^H(\ell,k)\right\|_2\|\hat{\mathbf{a}}(\ell,k)\|_2}\right)}{L(K/2+1)}\,(\mathrm{rad})\,. \tag{4.22}$$

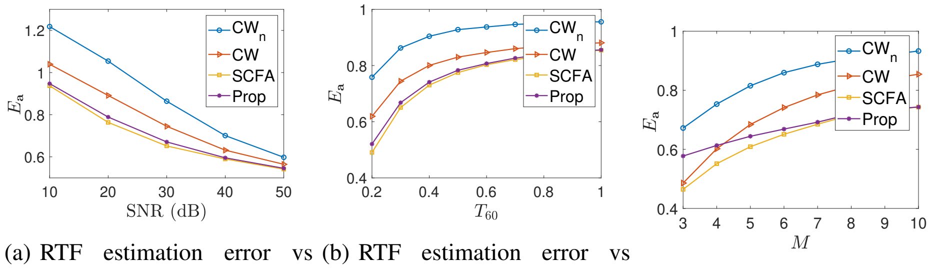


(a) Performance of the proposed method as a function of the size of $\mathbb{L}$.

(b) Performance in terms of $E_{\mathbf{a}}$ as a function of the noise PSD estimation errors.

Figure 4.3: Evaluation of the proposed and CW method.

For the results shown in Fig. 4.3, we use 6 microphones with reverberation time $T_{60} = 0.3$ s and signal-to-noise ratio (SNR) of 30 dB. Hence, we will have $(M-1)(M-2) = 20$ different estimates of each of the diagonal elements of $\bar{\mathbf{P}}_x$. As shown in Fig. 4.3a, the more estimates we average, the smaller the RTF estimation error becomes. Therefore, in the following experiments, we will average all different estimates in our proposed method. In Fig. 4.3b, the estimation performance of the CW method and our proposed method (referred to as 'Prop') are compared as a function of the noise PSD estimation error in dB, i.e., $E_{\phi_v} = 10\log_{10}\left(\phi_v/\hat{\phi}_v\right)$. Note that ϕ_v is the mean of the trace of the noise covariance matrix. E_{ϕ_v} ranges from an overestimation error of -20 dB to an underestimation error of ∞ dB (i.e. not subtracting anything before whitening) in Fig. 4.3b. Since the proposed method is independent of the noise PSD, the proposed method is not affected by E_{ϕ_v} and is presented as a horizontal line in Fig. 4.3b. Note that even at 0 dB, the proposed method outperforms 'CW', because the true noise spatial coherence matrix is not identical to, although close to, the identity matrix in the experiments.



(a) RTF estimation error vs SNR.

(b) RTF estimation error vs T_{60}.

(c) RTF estimation error vs M.

Figure 4.4: Performance comparison of the proposed method, the CW method and the SCFA method.

In Fig. 4.4, the simultaneous confirmatory factor analysis method (SCFA) [5] is also included for comparison, which minimizes the maximum likelihood cost function using the 'fmincon' MATLAB procedure after calculating the gradient and Hessian matrix at each updating step. Note that 'CW$_\mathrm{n}$' refers to CW without subtracting the noise covariance matrix, i.e., $E_{\phi_v} = \infty$ dB, while 'CW' refers to $E_{\phi_v} = 0$ dB. In Fig. 4.4a, we use 6 microphones and fix the reverberation time to 0.3 s, and only change the SNR from 10 dB to 50 dB. In Fig. 4.4b, we use 6 microphones, fix the SNR to 30 dB, and only change T_{60} from 0.2 s to 1 s. From these results, it follows that our proposed method and the SCFA method have a similar performance and both outperform the CW method in most scenarios. As the SNR increases or the T_{60} decreases, all methods improve. However, the proposed method has better performance compared to 'CW' for low SNR or small T_{60}, as the reverberation-to-noise ratio is small in both cases resulting in relatively large impact from the noise component.

In Fig. 4.4c, we fix the reverberation time to 0.3 s, the SNR to 30 dB, and only change the number of microphones from 3 to 10. The estimation performance of the proposed method is shown to be less good for a small number of microphones, but improves very fast when using more microphones and reaches almost the same performance as the

SCFA method for large M. The reason is that we use only the off-diagonal elements of the simplified covariance matrix $\hat{\hat{\mathbf{P}}}_y$ in the proposed method. The percentage of the number of elements in $\hat{\hat{\mathbf{P}}}_y$ we omit is $M/M^2 = 1/M$, which decreases as the number of microphones increases. In Table 4.1, we average and normalize the computation time over all scenarios per method. The runtime for Prop is close to CW, but much lower than for SCFA.

Table 4.1: Computation time comparison.

methods	SCFA	Prop	CW
run time	286.97	1	0.67

4.6. CONCLUSIONS

We considered the problem of estimating the RTF for a single source in a reverberant and noisy environment. We proposed a method that uses only off-diagonal elements of the simplified covariance matrix which are not affected by the late reverberation and the noise PSDs. Experiments show that the RTF estimation performance of the proposed method is insensitive to the noise PSD errors and reaches the performance of the SCFA method while using much less computation time. Both the proposed method and the SCFA method outperform the CW method, in most scenarios, especially for low SNR, low reverberation time and a large number of microphones.

REFERENCES

[1] O. L. Frost, “An algorithm for linearly constrained adaptive array processing”, *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[2] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech”, *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

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

[4] M. Tammen, S. Doclo, and I. Kodrasi, “Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[5] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust joint estimation of multimicrophone signal model parameters”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[6] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”, in *Proc. IEEE Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 11–15.

[7] I. Cohen, “Relative transfer function identification using speech signals”, *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[8] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[9] S. Markovich-Golan, S. Gannot, and W. Kellermann, “Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function”, 2018, pp. 2499–2503.

[10] S. Markovich, S. Gannot, and I. Cohen, “Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[11] O. Schwartz, S. Gannot, and E. A. P. Habets, “Multi-microphone speech dereverberation and noise reduction using relative early transfer functions”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 240–251, 2015.

[12] S. Braun and E. A. Habets, “Dereverberation in noisy environments using reference signals and a maximum likelihood estimator”, in *Proc. EURASIP Eur. Signal Process. Conf.*, 2013, pp. 1–5.

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

[14] N. Ito, H. Shimizu, N. Ono, and S. Sagayama, “Diffuse noise suppression using crystal-shaped microphone arrays”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2101–2110, 2011.

[15] A.-J. Boonstra and A.-J. van der Veen, “Gain calibration methods for radio telescope arrays”, *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 25–38, 2003.

[16] S. A. Mulaik, *Foundations of factor analysis*. CRC press, 2009.

[17] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

5

ALTERNATING LEAST-SQUARES-BASED MICROPHONE ARRAY PARAMETER ESTIMATION FOR A SINGLE-SOURCE REVERBERANT AND NOISY ACOUSTIC SCENARIO

This chapter is based on the article published as: C. Li and R. C. Hendriks. "Alternating Least-Squares-Based Microphone Array Parameter Estimation for a Single-Source Reverberant and Noisy Acoustic Scenario". In: IEEE/ACM Trans. Audio, Speech, Language Process. 31 (2023), pp. 3922–3934.

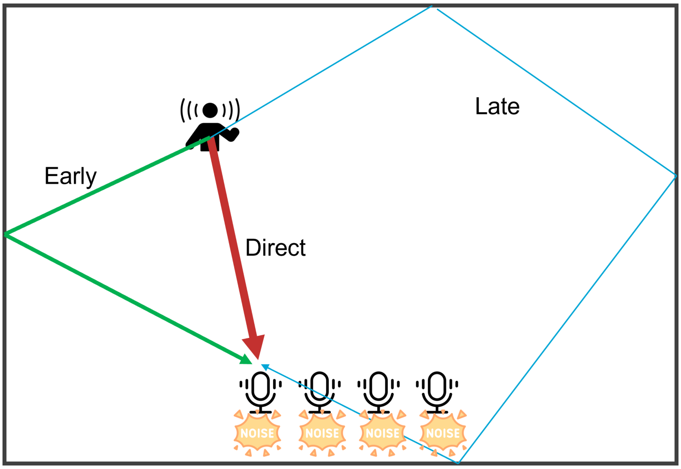


Figure 5.1: Illustration of a single source, reverberant and noisy scenario.

In this chapter, we consider a single source reverberant and noisy scenario as illustrated in Fig. 5.1, which is the same as the scenario we considered in the previous chapter. However, instead of considering the RTF estimation only, we will also estimate the PSDs of the source, the late reverberation and the noise. We will propose a joint estimator to answer research question 1.3 shown in Fig. 1.4 using the signal model presented in Fig. 2.2 (b).

Acoustic-scene-related parameters such as RTFs and PSDs of the target source, late reverberation and ambient noise are essential for microphone array signal processing but are challenging to estimate. Existing methods typically only estimate a subset of the parameters by assuming the other parameters are known. This can lead to unmatched scenarios and reduced estimation performance. Moreover, many methods process time frames independently, despite they share the same RTF. In this chapter, we will propose a joint estimator using multiple time frames. We first modify an existing alternating least squares (ALS) method using a single time frame. Then, we extend it to use multiple time frames. Furthermore, we propose more robust constraints on the PSDs to avoid large estimation errors. We compare our proposed method (JALS) to the state-of-the-art SCFA method, the JMLE method from Chapter 3 and the ALS method. The experiments on estimation accuracy, predicted speech quality, and predicted speech intelligibility demonstrate that JALS has a performance similar to SCFA, both which outperform ALS in all scenarios and outperform JMLE particularly in low SNR scenarios. Moreover, JALS is the least computational complex, confirmed by the measured processing time, which is significantly lower than SCFA.

5.1. INTRODUCTION

Hands-free speech communication applications like mobile phones and hearing aids are commonly used nowadays. Equipped with microphone arrays, these devices can record and analyze the speech signal for various applications. Unavoidably, the microphone signals are corrupted by reverberation and ambient noise, which can degrade the speech quality and intelligibility [1], [2]. Hence, techniques like spatial filtering are used to extract the target signal from the noisy microphone signals. Typically, these spatial filters depend on acoustic-scene-related parameters such as relative transfer functions (RTFs) and power spectral densities (PSDs) of the source, the late reverberation and the ambient noise. In practice, these parameters are typically unknown. Therefore, an essential problem with hands-free speech communication applications is to estimate the aforementioned parameters. Note that there are non-parametric techniques such as blind beamforming or blind source separation [3], [4] that can extract the target signal without estimating the parameters. However, in this work we only focus on parametric beamformers where the estimated parameters can be used as a prior information on the acoustic scene.

Due to the non-stationarity of the speech signal, the PSDs of the target source and the late reverberation are time-varying. The PSDs of the ambient noise can be time-varying as well, depending on the working environment of the microphone arrays. The RTFs can change over time as well depending on whether the source is moving relative to the array. The facts that these parameters can be time-varying and corruptions caused by reverberation and ambient noise are present, make the estimation of these parameters rather challenging.

In recent years, many methods have been proposed to estimate these parameters, see e.g., [5]–[14]. Many of these methods only estimate a subset of the parameters by making some strict assumptions about the acoustic scenarios and the knowledge of the remaining parameters. For example, in [5], [9], [12], the RTFs of the target source are assumed to be known such that the speech PSD, late reverberation PSD and noise PSD can be estimated. In [6], the PSD of the late reverberation is assumed to be known and the RTF of the target source is estimated. In [7], the RTFs and the PSDs of all sources and the noise covariance matrix are estimated. However, it is assumed that the late reverberation component is stationary and only a single source is active per time frequency tile. In [8], the noise covariance matrix is assumed known and the late reverberation PSD is estimated. In Chapter 3 and [13], the noiseless scenario is assumed, neglecting the estimation of the ambient noise PSD.

From the above overview, we see that existing methods for parameter estimation from the acoustic scene all assume a subset of parameters to be known. However, erroneously assuming a subset of the parameters to be known can lead to unmatched scenarios, and thus to reduced noise reduction performance. This emphasizes the importance of accurate joint parameter estimation. A second important point is the fact that, apart from a few exceptions, e.g., [11], [14], many of these methods process the time frames independently, despite the fact that they may share some common information. For instance, the RTFs corresponding to some adjacent time frames are the same if the sound source is static during these time frames. In such cases, we could use these time frames jointly to obtain better estimates of the RTFs [11], [14].

The joint estimation of parameters using multiple time frames is realized in [11] in a reverberant and noisy environment, using the simultaneous confirmatory factor analysis (SCFA) method. As expected, SCFA has much better estimation performance compared to methods using each time frame independently, especially for the RTF estimation [11]. Nevertheless, SCFA has a rather high computational cost. Therefore, we recently proposed some alternative methods that can achieve a nearly similar performance as SCFA, but at a much lower complexity [14].

In [14], we considered a single reverberant source scenario and proposed a joint maximum likelihood estimator (JMLE) for the parameters of interest. In the current work, we extend the signal model from [14] to the noisy case. Specifically, we model the noise component as a spatially homogeneous sound field characterized by a time-invariant spatial coherence matrix with a time-varying PSD. We can assume the spatial coherence matrix is known, as assumed in [9]. Further, we consider the use of multiple time frames to jointly form a segment. The RTF is considered constant across the segment, while the PSDs of the target's early reflections, the PSDs of the late reverberation and the ambient noise PSD are allowed to change from frame-to-frame. The focus herein is to jointly estimate the source's RTF, and the PSDs of the early reflections, the late reverberation and the ambient noise at low complexity. We will use the least squares (LS) error as a cost function, i.e., minimizing the Frobenious norm of model error matrices. Note that the LS cost function has been considered in [10] as well to estimate these parameters and the LS minimization was solved by an alternating least squares (ALS) method. However, we will show in this work that the ALS based method from [10] can suffer from a parameter identifiability issue and thus needs to be modified to obtain more accurate estimates. Note also that the ALS method from [10] uses each time frame separately. Hence, we will extend the modified ALS method such that it uses multiple time frames jointly to improve the estimation performance. In addition, we propose constraints on the estimated PSDs that are more robust than the ones used in [10] to avoid large estimation errors. Note that minimizing the least squares cost function for multiple time frames jointly can be seen as a special case of the joint diagonalization problems modeled in [15]–[17], except that the problem proposed in our work has additional constraints on some of the parameters and the single target source is disturbed by both the late reverberation and the ambient noise.

The remaining parts of the chapter are structured as follows. In Section 5.2, we introduce the notation used in this chapter, present the signal model and formulate the problem discussed in this chapter. In Section 5.3, we will present the existing ALS method, propose a modified ALS method and extend it to a method using multiple time frames. After that, we will compare our proposed methods to some state-of-the-art reference methods in various simulated acoustic experiments in Section 5.4. Finally, we will draw the conclusions in Section 5.5.

The matlab code of the proposed methods can be downloaded from: http://sps.ewi.tudelft.nl/Repository/

5.2. PRELIMINARIES

5.2.1. NOTATION

In this chapter, we use lower-case letters to denote scalars, bold-face lower-case letters for vectors and bold-face upper-case letters for matrices. Matrix notation with subscripts using two lower-case letters (e.g. $\mathbf{P}_{\mathbf{y}_{i,j}}$) denotes the element of the matrix. Matrix notation with superscripts $T, *, H$ denotes taking the transpose, the conjugate and the conjugate transpose of the matrix, respectively. $\Re(x)$ and $\Im(x)$ represent the real part and the imaginary part of a complex-valued variable x, respectively. Further, $\mathrm{E}[\cdot]$ refers to the expectation operator, $\mathrm{tr}(\cdot)$ refers to taking the trace of a matrix, and if not further specified, $|\cdot|$ denotes taking the determinant of a matrix. Finally, $\mathrm{diag}[a_1, \cdots, a_M]$ denotes a diagonal matrix with diagonal elements $a_1, \cdots, a_M$ and $\|\cdot\|_F$ denotes taking the Frobenius norm of a matrix.

5.2.2. SIGNAL MODEL

We consider a reverberant and noisy environment, in which a single acoustic point source is recorded by an array of M microphones with an arbitrary geometric structure. The microphone signal received at the m_{th} microphone in the short-time Fourier transform (STFT) domain is given by

$$y_m(l,k) = x_m(l,k) + r_m(l,k) + v_m(l,k), \tag{5.1}$$

where l is the time-frame index and k is the frequency bin index, $x_m(l,k)$ is the sum of the direct sound and the early reflections, $r_m(l,k)$ is the sum of all the late reflections in time frame l and frequency bin k, and $v_m(l,k)$ contains the ambient noise and microphone self-noise. Since the direct components and early reflections are beneficial for speech intelligibility [18], the combination of these components forms our target signal,

$$x_m(l,k) = a_m(l,k)\, s(l,k), \tag{5.2}$$

where $s(l,k)$ contains the direct and early speech component recorded by the reference microphone and $a_m(l,k)$ is the relative transfer function (RTF) between the reference microphone and the m_{th} microphone. By selecting the first microphone as the reference microphone, we have the prior information that $a_1 = 1$. Note that we use the multiplicative transfer function (MTF) approximation in Eq. (5.2) for ease of analyzing, instead of the convolutive transfer function (CTF) approximation [19], [20]. Stacking the M microphone STFT coefficients into a column vector, we have

$$\mathbf{y}(l,k) = \mathbf{a}(l,k)\, s(l,k) + \mathbf{r}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1}, \tag{5.3}$$

where $\mathbf{y}(l,k) = [y_1(l,k), \cdots, y_M(l,k)]^T$ and the other vectors are defined in the same way.

5.2.3. CROSS POWER SPECTRAL DENSITY MATRICES

By processing in short time frames, we can assume the three components in Eq. (5.3) to be stationary and mutually uncorrelated within a time frame. The PSD matrix of the noisy microphone recordings can therefore be expressed as

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(l,k) &= \mathrm{E}\left[\mathbf{y}(l,k)\,\mathbf{y}^{H}(l,k)\right] \\ &= \mathbf{P}_{\mathbf{x}}(l,k)+\mathbf{P}_{\mathbf{r}}(l,k)+\mathbf{P}_{\mathbf{v}}(l,k) \in \mathbb{C}^{M\times M},\end{aligned} \tag{5.4}$$

where $\mathbf{P}_{\mathbf{x}}$ is given by

$$\mathbf{P}_{\mathbf{x}}(l,k) = \phi_s(l,k)\mathbf{a}(l,k)\mathbf{a}^{H}(l,k), \tag{5.5}$$

and $\phi_s(l,k)=\mathrm{E}\left[|s(l,k)|^2\right]$ is the PSD of the target source at the reference microphone with $|\cdot|$ taking the absolute value. However, notice that across frames, s and r might be correlated.

The CPSD matrix of the late reverberation component is commonly modelled as [5], [21]

$$\mathbf{P}_{\mathbf{r}}(l,k) = \phi_\gamma(l,k)\boldsymbol{\Gamma}(k), \tag{5.6}$$

which is a spatially homogeneous and isotropic sound field with a time varying PSD $\phi_\gamma(l,k)$. The spatial coherence matrix $\boldsymbol{\Gamma}(k)$ is time-invariant. Hence, $\boldsymbol{\Gamma}(k)$ can be estimated in advance using the information on the microphone array geometry [22]–[24]. We assume a spherically isotropic noise field [25] and model the $\{i,j\}$-th element of $\boldsymbol{\Gamma}(k)$ as

$$\boldsymbol{\Gamma}_{i,j}(k) = \mathrm{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right), \tag{5.7}$$

where $\mathrm{sinc}(x)=\frac{\sin x}{x}$, $d_{i,j}$ is the inter-distance between microphones i and j, f_s is the sampling frequency, c denotes the speed of sound and K is the number of frequency bins.

Lastly, we assume that the residual noise component has a similar CPSD matrix formulation as the late reverberation, i.e.,

$$\mathbf{P}_{\mathbf{v}}(l,k) = \phi_v(l,k)\boldsymbol{\Psi}(k), \tag{5.8}$$

where $\boldsymbol{\Psi}(k)$ is the known spatial coherence matrix and $\phi_v(l,k)$ is unknown PSD. We assume that $\boldsymbol{\Psi}(k)$ is non-singular and linearly independent with $\boldsymbol{\Gamma}(k)$ (i.e. $\boldsymbol{\Psi}(k)$ is not a scaled version of $\boldsymbol{\Gamma}(k)$). Note that when considering the microphone self noise only, we have $\boldsymbol{\Psi}(k)=\mathbf{I}$.

5.2.4. PROBLEM FORMULATION

Based on the assumptions made in the previous subsection and Eqs. (5.5), (5.6) and (5.8), we can rewrite the noisy CPSD matrix for each time frame l as

$$\mathbf{P}_{\mathbf{y}}(l) = \phi_s(l)\mathbf{a}(l)\mathbf{a}^{H}(l)+\phi_\gamma(l)\boldsymbol{\Gamma}+\phi_v(l)\boldsymbol{\Psi}. \tag{5.9}$$

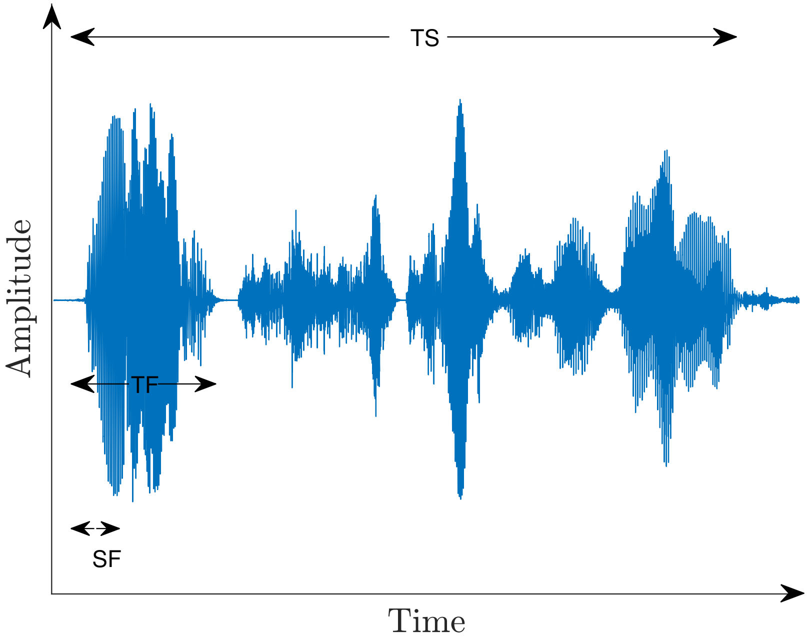


Figure 5.2: Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

Note that we omit the frequency bin index k in Eq. (5.9) and hereafter for legibility since the signals will be processed for each k independently. By making the RTF vector $\mathbf{a}$ dependent on the time-frame index l, we implicitly assume that the relative source position or room acoustics can change from time frame to time frame. However, we consider in this work a semi-static source scenario by assuming the RTF $\mathbf{a}$ does not change for N (a finite number) time frames (N ranges from 1 to 8 in our experiments, corresponding to a duration of approximately 0.5 s to 5 s). We denote the set of N time frames sharing a single RTF by a time segment with index β. The noisy CPSD matrix then becomes

$$\mathbf{P_y}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\mathbf{\Gamma} + \phi_v(l)\mathbf{\Psi}, \tag{5.10}$$

with $\beta = \lfloor \frac{l-1}{N} \rfloor + 1$.

Further, we define sub frames indexed by t_s, where T_{sf} overlapping sub frames form a time frame. See Fig. 5.2 for a visual interpretation of time segment, time frame and sub frame. Since the noisy signal is assumed to be stationary within a time frame, we can estimate the CPSD matrix per time frame i based on a sampled covariance matrix using the sub-time frames, that is,

$$\hat{\mathbf{P}}_{\mathbf{y}}(l) = \frac{1}{T_{sf}} \sum_{t_s=1+(l-1)T_{sf}}^{lT_{sf}} \mathbf{y}(t_s)\mathbf{y}(t_s)^H, \tag{5.11}$$

where $\mathbf{y}(t_s)$ denotes the STFT coefficients vector.

Accurate estimation of the parameters from the signal model in Eq. (5.10) is very important for speech enhancement and intelligibility improvement algorithms. However, this is also very challenging when the source is only stationary for a short time and microphone and source positions are time varying. The main goal of this chapter therefore is to estimate the RTF vector, the PSD of the source, the PSD of the late reverberation and the PSD of self-noise simultaneously using N sequentially estimated CPSD matrices $\hat{\mathbf{P}}_{\mathbf{y}}(l)$ for one time segment β, i.e., for N time frames, while the source is only stationary within a time frame and the RTF changes from segment-to-segment.

5.3. ALS-BASED JOINT ESTIMATION

To jointly estimate the parameters of interest, we consider the use of alternating least squares (ALS) based methods. Note that a two-step ALS method has been proposed before in this context [10]. In Section 5.3.1, we will first introduce the method proposed in [10]. Then in Section 5.3.2 we will propose a modified version of the ALS method based on two improvements over the original method to overcome parameter identifiability issues and potential numerical issues due to matrix singularities. Note that in [10] each time frame is utilized separately. However, if we assume the CPSD matrices for multiple time frames in a single time segment share the same RTF vector, we can use these time frames jointly to estimate RTF $\mathbf{a}$ with improved accuracy. Therefore, we will extend the modified ALS method to the case using the PSD matrices for multiple time frames in Section 5.3.3.

5.3.1. ALS FOR A SINGLE TIME FRAME

In [10], for each single time frame, the estimates of the RTF vector $\mathbf{a}$ and the PSD vector $\phi = \left[\phi_s, \phi_\gamma, \phi_v\right]^T$ are obtained by minimizing the Frobenius norm of a model mismatch error matrix, i.e.,

$$\underset{\mathbf{a},\phi}{\arg\min} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \phi_s \mathbf{a}\mathbf{a}^H - \phi_\gamma \hat{\mathbf{\Gamma}} - \phi_v \hat{\mathbf{\Psi}} \right\|_F^2, \tag{5.12}$$

where $\hat{\mathbf{A}}$ means the estimated $\mathbf{A}$. Note that the cost function in Eq. (5.12) is non-convex. To solve Eq. (5.12), a two-step ALS method is used by assuming that for either $\mathbf{a}$ or ϕ, an estimate is given and then estimating the other parameter vector.

More specifically, by assuming the RTF vector $\mathbf{a}$ is known or already estimated, the estimate of ϕ can be obtained by solving

$$\underset{\phi}{\arg\min} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \phi_s \hat{\mathbf{a}}\hat{\mathbf{a}}^H - \phi_\gamma \hat{\boldsymbol{\Gamma}} - \phi_v \hat{\boldsymbol{\Psi}} \right\|_F^2, \tag{5.13}$$

which has the following closed form solution [10]

$$\hat{\phi} = \boldsymbol{\Phi}_{\mathbf{a}}^{-1}\mathbf{b}, \tag{5.14}$$

where

$$\boldsymbol{\Phi}_{\mathbf{a}} = \begin{bmatrix} \left(\hat{\mathbf{a}}^H\hat{\mathbf{a}}\right)^2 & \hat{\mathbf{a}}^H\hat{\boldsymbol{\Gamma}}\hat{\mathbf{a}} & \hat{\mathbf{a}}^H\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{a}} \\ \hat{\mathbf{a}}^H\hat{\boldsymbol{\Gamma}}\hat{\mathbf{a}} & \operatorname{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Gamma}}\right\} & \operatorname{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Psi}}\right\} \\ \hat{\mathbf{a}}^H\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{a}} & \operatorname{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\boldsymbol{\Psi}}\right\} & \operatorname{tr}\left\{\hat{\boldsymbol{\Psi}}^H\hat{\boldsymbol{\Psi}}\right\} \end{bmatrix}, \tag{5.15}$$

and

$$\mathbf{b} = \begin{bmatrix} \hat{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{y}}\hat{\mathbf{a}} \\ \operatorname{tr}\left\{\hat{\boldsymbol{\Gamma}}^H\hat{\mathbf{P}}_{\mathbf{y}}\right\} \\ \operatorname{tr}\left\{\hat{\boldsymbol{\Psi}}^H\hat{\mathbf{P}}_{\mathbf{y}}\right\} \end{bmatrix}. \tag{5.16}$$

When assuming the PSD vector $\hat{\phi}$ is already estimated, the RTF vector $\mathbf{a}$ can be estimated by minimizing the cost function with respect to $\mathbf{a}$, that is

$$\underset{\mathbf{a}}{\arg\min} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_s \mathbf{a}\mathbf{a}^H - \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} - \hat{\phi}_v \hat{\boldsymbol{\Psi}} \right\|_F^2, \tag{5.17}$$

which also has a closed form solution [26] given by the scaled principal eigenvector of the matrix $\hat{\mathbf{P}}_{\mathbf{x}} = \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} - \hat{\phi}_v \hat{\boldsymbol{\Psi}}$, which is

$$\hat{\mathbf{a}} = \sqrt{\frac{\lambda}{\hat{\phi}_s}}\nu, \tag{5.18}$$

where λ and ν are the principal eigenvalue and eigenvector of $\hat{\mathbf{P}}_{\mathbf{x}}$. The two steps are performed iteratively.

For the first step, the method in [10] finds an initial estimate of the RTF vector $\mathbf{a}$ by taking a random value or using a coarse estimate of the direction of arrival of the target source. For the second step, the PSD vector ϕ is estimated via Eq. (5.14) with $\boldsymbol{\Phi}_{\mathbf{a}}$ and $\mathbf{b}$ calculated using the initial estimate $\hat{\mathbf{a}}$. Using the estimate of ϕ, matrix $\hat{\mathbf{P}}_{\mathbf{x}}$ is calculated in the second step and the RTF vector $\mathbf{a}$ can be estimated again via Eq. (5.18). For the next iterations, the two steps are repeated and the estimates of $\mathbf{a}$ and ϕ are updated in an alternating fashion until a given convergence criterion is achieved or a certain number of iterations I are executed. Note that since each step reduces the cost function value, this method can converge to a local minimum even though the global minimum is not guaranteed. The ALS method is summarized in Algorithm 1. Note that the convergence rate of alternating least square based methods is slow [27].

Since PSDs should be positive by definition, all the estimated PSDs need to be lower bounded. In [10], the estimates of the PSDs are updated in the following way[1]:

$$\left\{\phi_s, \phi_\gamma, \phi_v\right\} = \max\left(\left\{\phi_s, \phi_\gamma, \phi_v\right\}, \varepsilon\right), \tag{5.19}$$

and

$$\left\{\phi_s, \phi_\gamma, \phi_v\right\} = \min\left(\left\{\phi_s, \phi_\gamma, \phi_v\right\}, \frac{\operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{y}\right)}{M}\right), \tag{5.20}$$

where ε is the machine precision.

Algorithm 1: ALS method

```
Input: $\hat{\mathbf{P}}_\mathbf{y}$, $\hat{\mathbf{\Gamma}}$,$\hat{\mathbf{\Psi}}$,init.$\hat{\bar{\mathbf{a}}}$, $I$
Output: $\mathbf{a}$, $\phi$
for all $k,l$ do
    for iter=1:I do
        Compute $\mathbf{\Phi}_a$ using Eq. (5.15) and $\mathbf{b}$ using Eq. (5.16).
        Estimate $\phi$ using Eq. (5.14).
        Constrain the estimates of PSDs using Eq. (5.19) and Eq. (5.20).
        Calculate $\hat{\mathbf{P}}_\mathbf{x} = \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_\gamma\hat{\mathbf{\Gamma}} - \hat{\phi}_v\hat{\mathbf{\Psi}}$.
        Take EVD of $\hat{\mathbf{P}}_\mathbf{x}$ to find its principal eigenvalue and eigenvector.
        Estimate $\mathbf{a}$ using Eq. (5.18).
    for next time frame: use $\mathbf{a} = \mathbf{a}/a_1$ as the initial estimate.
```

5.3.2. MODIfiED-ALS FOR A SINGLE TIME FRAME

An important condition for parameter estimation is the fact that the estimation problem itself needs to be identifiable [28]. Specifically, in the problem of jointly estimating the RTF vector $\mathbf{a}$ and the PSDs, the following condition should be satisfied for any two sets of parameters $\left\{\mathbf{a}, \phi_s, \phi_\gamma, \phi_v\right\}$ and $\left\{\bar{\mathbf{a}}, \bar{\phi}_s, \bar{\phi}_\gamma, \bar{\phi}_v\right\}$:

$$\begin{gathered} \phi_s\mathbf{a}\mathbf{a}^H + \phi_\gamma\mathbf{\Gamma} + \phi_v\mathbf{\Psi} = \bar{\phi}_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H + \bar{\phi}_\gamma\mathbf{\Gamma} + \bar{\phi}_v\mathbf{\Psi} \\ \Leftrightarrow \\ \phi_s = \bar{\phi}_s, \mathbf{a} = \bar{\mathbf{a}}, \phi_\gamma = \bar{\phi}_\gamma, \phi_v = \bar{\phi}_v \end{gathered} \tag{5.21}$$

In the ALS method [10], however, Eq. (5.21) does not hold. To see this, let $\bar{\phi}_s = 4\phi_s$ and $\bar{\mathbf{a}} = \frac{\mathbf{a}}{2}$, we have $\phi_s\mathbf{a}\mathbf{a}^H = \bar{\phi}_s\bar{\mathbf{a}}\bar{\mathbf{a}}^H$ but $\bar{\phi}_s \neq \phi_s$ and $\bar{\mathbf{a}} \neq \mathbf{a}$. Therefore, any proper scaling of $\mathbf{a}$ and ϕ_s can be a solution as well. To solve this issue, we use the prior information that

[1] Note that this step can be replaced by solving bounded-variable least squares (BVLS) problem, but using BVLS does not improve the estimation performance while increasing the computation cost based on our experimental tests.

$a_1 = 1$. In the final iteration, after estimating $\mathbf{a}$ using Eq. (5.18), we add a normalization step for both $\mathbf{a}$ and ϕ_s using the constant $c = \hat{a}_1$:

$$\hat{\mathbf{a}} \leftarrow \frac{\hat{\mathbf{a}}}{c} \tag{5.22}$$

and

$$\hat{\phi}_s \leftarrow \hat{\phi}_s \left|c\right|^2. \tag{5.23}$$

Notice also that in each iteration of the ALS method, if the estimated ϕ_s has an unusually small value (e.g. eps), the elements of the estimate of $\mathbf{a}$ in Eq. (5.18) will have rather large values. This will lead to large values of the first column and the first row of the matrix $\mathbf{\Phi_a}$ in Eq. (5.15), which means $\mathbf{\Phi_a}$ is close to being singular or badly scaled. To solve this issue, we can constrain the norm of the estimate of the scaled RTF vector to 1 by simply using the principal eigenvector instead of the scaled one in Eq. (5.18). Note that estimating the scaled $\mathbf{a}$ and ϕ_s is allowable because we will normalize them using Eqs. (5.22) and (5.23) eventually in the last step.

The modified alternating least squares (MALS) method aims at minimizing the following cost function

$$\underset{\tilde{\mathbf{a}},\tilde{\phi}_s,\phi_\gamma,\phi_v}{\arg\min} \left\| \hat{\mathbf{P}}_{\mathbf{y}} - \tilde{\phi}_s \tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \phi_\gamma \hat{\mathbf{\Gamma}} - \phi_v \hat{\mathbf{\Psi}} \right\|_F^2, \tag{5.24}$$

where $\tilde{\mathbf{a}} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{a}}}$ and $\tilde{\phi}_s = \phi_s \mathbf{a}^H\mathbf{a}$. Since $\tilde{\phi}_s \tilde{\mathbf{a}}\tilde{\mathbf{a}}^H = \phi_s \mathbf{a}\mathbf{a}^H$, the solution to Eq. (5.24) will also be the solution to Eq. (5.12). Once the estimates $\tilde{\mathbf{a}}$ and $\tilde{\phi}_s$ are obtained, the estimates of the RTF vector and the PSD of the source are given by

$$\mathbf{a} \leftarrow \frac{\tilde{\mathbf{a}}}{\tilde{a}_1}, \tag{5.25}$$

and

$$\phi_s \leftarrow \tilde{\phi}_s |\tilde{a}_1|^2. \tag{5.26}$$

Similarly as in [10] and as described in Section 5.3.1, The optimization problem in Eq. (5.24) can be solved in an alternating fashion. Assuming $\tilde{\mathbf{a}}$ is already available (from a previous iteration or initialization), $\tilde{\phi} = \left[\tilde{\phi}_s, \phi_\gamma, \phi_v\right]$ is estimated by the least squares estimate

$$\hat{\tilde{\phi}} = \mathbf{\Phi}_{\tilde{\mathbf{a}}}^{-1}\tilde{\mathbf{b}}, \tag{5.27}$$

where

$$\mathbf{\Phi}_{\tilde{\mathbf{a}}} = \begin{bmatrix} 1 & \hat{\tilde{\mathbf{a}}}^H \mathbf{\Gamma} \hat{\tilde{\mathbf{a}}} & \hat{\tilde{\mathbf{a}}}^H \mathbf{\Psi} \hat{\tilde{a}} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Gamma}} \hat{\tilde{\mathbf{a}}} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Gamma}}\right\} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Psi}} \hat{\tilde{a}} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} & \operatorname{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{\Psi}}\right\} \end{bmatrix}, \tag{5.28}$$

and

$$\tilde{\mathbf{b}} = \begin{bmatrix} \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{y}} \hat{\tilde{\mathbf{a}}} \\ \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{P}}_{\mathbf{y}}\right\} \\ \operatorname{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{P}}_{\mathbf{y}}\right\} \end{bmatrix}. \tag{5.29}$$

When an estimate of $\hat{\tilde{\phi}}$ is known from the previous iteration, we calculate the matrix $\hat{\mathbf{P}}_{\mathbf{x}} = \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_{\gamma}\hat{\mathbf{\Gamma}} - \hat{\phi}_{v}\hat{\mathbf{\Psi}}$ and obtain the estimate of $\tilde{\mathbf{a}}$ by

$$\hat{\tilde{\mathbf{a}}} = v, \tag{5.30}$$

where v is the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{x}}$. We terminate the algorithm after a sufficient number of iterations. $\mathbf{a}$ and ϕ are then obtained using Eq. (5.25) and Eq. (5.26).

The MALS method is summarized in Algorithm 2.

Algorithm 2: MALS method

```
Input: $\hat{\mathbf{P}}_{\mathbf{y}}$, $\hat{\mathbf{\Gamma}}$,$\hat{\mathbf{\Psi}}$,init.$\hat{\tilde{\mathbf{a}}}$, $I$
Output: $\mathbf{a}$, $\phi$
for all $k,l$ do
    for iter=1:I do
        Compute $\mathbf{\Phi}_{\tilde{\mathbf{a}}}$ using Eq. (5.28) and $\tilde{\mathbf{b}}$ using Eq. (5.29).
        Estimate $\tilde{\phi}$ using Eq. (5.27).
        Calculate $\hat{\mathbf{P}}_{\mathbf{x}} = \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\gamma}\mathbf{\Gamma} - \hat{\phi}_{v}\hat{\mathbf{\Psi}}$.
        Take EVD of $\hat{\mathbf{P}}_{\mathbf{x}}$ to find its principal eigenvector.
        Estimate $\tilde{\mathbf{a}}$ using Eq. (5.30).
    Estimate $\mathbf{a}$ and $\phi_s$ using Eq. (5.25) and Eq. (5.26).
```

5.3.3. ALS FOR MULTIPLE TIME FRAMES

In the previous subsections, the joint estimation of the RTF vector $\mathbf{a}$ and the PSD vector ϕ is performed for a single time frame based on the ALS approach. However, in many cases, $\mathbf{a}$ can be assumed to be constant across multiple frames in a time segment. With this prior information, we consider in this subsection the joint estimation of $\mathbf{a}$, and the PSD vector $\phi = \left[\phi\left(1+(\beta-1)N\right)^T, \cdots, \phi\left(\beta N\right)^T\right]^T$ using all time-frames in a segment, where $\phi\left(l\right) = \left[\phi_s\left(l\right), \phi_\gamma\left(l\right), \phi_v\left(l\right)\right]^T$ for $l = 1+(\beta-1)N, \cdots, \beta N$.

The alternating least squares method using multiple time frames jointly (JALS) aims at minimizing the sum of the Frobenius norms of the model mismatch error matrices for all time frames l that fall in the same segment β, i.e.,

$$\underset{\mathbf{a},\phi}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\|\hat{\mathbf{P}}_{\mathbf{y}}\left(l\right) - \phi_s\left(l\right)\mathbf{a}\mathbf{a}^H - \phi_\gamma\left(l\right)\hat{\mathbf{\Gamma}} - \phi_v\left(l\right)\hat{\mathbf{\Psi}}\right\|_F^2. \tag{5.31}$$

Like the MALS method, we reparameterize $\mathbf{a}$ and $\phi_s(l)$ for $l = 1+(\beta-1)N, \cdots, \beta N$ by $\tilde{\mathbf{a}} = \frac{\mathbf{a}}{\sqrt{\mathbf{a}^H\mathbf{a}}}$ and $\tilde{\phi}_s(l) = \phi_s(l)\mathbf{a}^H\mathbf{a}$, which gives us the following cost function

$$\underset{\tilde{\mathbf{a}},\tilde{\phi}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\|\hat{\mathbf{P}}_{\mathbf{y}}\left(l\right) - \tilde{\phi}_s\left(l\right)\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \phi_\gamma\left(l\right)\hat{\mathbf{\Gamma}} - \phi_v\left(l\right)\hat{\mathbf{\Psi}}\right\|_F^2. \tag{5.32}$$

To solve Eq. (5.32), we also use a two-step ALS method by either assuming $\tilde{\mathbf{a}}$ is given and estimating $\tilde{\phi}$ or assuming $\tilde{\phi}$ is estimated and estimating $\tilde{\mathbf{a}}$.

When an estimate of $\tilde{\mathbf{a}}$ is already given, the minimization with respect to $\tilde{\phi}$ is

$$\underset{\tilde{\phi}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \tilde{\phi}_s(l)\, \hat{\tilde{\mathbf{a}}}\hat{\tilde{\mathbf{a}}}^H - \phi_\gamma(l)\, \hat{\mathbf{\Gamma}} - \phi_v(l)\, \hat{\mathbf{\Psi}} \right\|_F^2, \tag{5.33}$$

which is equivalent to minimizing the cost function for each time frame l separately, i.e.

$$\underset{\substack{\tilde{\phi}(l),\\ \forall l \in 1+(\beta-1)N,\cdots,\beta N}}{\arg\min} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \tilde{\phi}_s(l)\, \hat{\tilde{\mathbf{a}}}\hat{\tilde{\mathbf{a}}}^H - \phi_\gamma(l)\, \hat{\mathbf{\Gamma}} - \phi_v(l)\, \hat{\mathbf{\Psi}} \right\|_F^2, \tag{5.34}$$

as $\tilde{\phi}(l)$ is defined per time frame. For each time frame l, Eq. (5.34) has a closed form solution

$$\tilde{\phi}(l) = \tilde{\mathbf{\Phi}}_{\tilde{\mathbf{a}}}^{-1} \tilde{\mathbf{b}}(l), \tag{5.35}$$

where

$$\tilde{\mathbf{\Phi}}_{\tilde{\mathbf{a}}} = \begin{bmatrix} 1 & \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Gamma}} \hat{\tilde{\mathbf{a}}} & \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Psi}} \hat{\tilde{\boldsymbol{a}}} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Gamma}} \hat{\tilde{\mathbf{a}}} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Gamma}}\right\} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} \\ \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{\Psi}} \hat{\tilde{\boldsymbol{a}}} & \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{\Psi}}\right\} & \operatorname{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{\Psi}}\right\} \end{bmatrix}, \tag{5.36}$$

and

$$\tilde{\mathbf{b}}(l) = \begin{bmatrix} \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{y}}(l) \hat{\tilde{\mathbf{a}}} \\ \operatorname{tr}\left\{\hat{\mathbf{\Gamma}}^H \hat{\mathbf{P}}_{\mathbf{y}}(l)\right\} \\ \operatorname{tr}\left\{\hat{\mathbf{\Psi}}^H \hat{\mathbf{P}}_{\mathbf{y}}(l)\right\} \end{bmatrix}. \tag{5.37}$$

When an estimate of $\tilde{\phi}$ is given, $\tilde{\mathbf{a}}$ can be obtained for a segment β by minimizing

$$\underset{\tilde{\mathbf{a}}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\tilde{\phi}}_s(l)\, \tilde{\mathbf{a}}\tilde{\mathbf{a}}^H - \hat{\gamma}(l)\, \hat{\mathbf{\Gamma}} - \hat{\phi}_v(l)\, \hat{\mathbf{\Psi}} \right\|_F^2. \tag{5.38}$$

We define $\hat{\mathbf{P}}_{\mathbf{x}}(l) = \hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\phi}_\gamma(l)\, \hat{\mathbf{\Gamma}} - \hat{\phi}_v(l)\, \hat{\mathbf{\Psi}}$ and reformulate Eq. (5.38) as

$$\begin{aligned} &\underset{\tilde{\mathbf{a}}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left\| \hat{\mathbf{P}}_{\mathbf{x}}(l) - \hat{\tilde{\phi}}_s(l)\, \tilde{\mathbf{a}}\tilde{\mathbf{a}}^H \right\|_F^2 \\ &= \underset{\tilde{\mathbf{a}}}{\arg\min} \sum_{l=1+(\beta-1)N}^{\beta N} \left[\left(\hat{\phi}_s(l)\, \tilde{\mathbf{a}}^H \tilde{\mathbf{a}} \right)^2 - 2\hat{\tilde{\phi}}_s(l)\, \tilde{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{x}}(l)\, \tilde{\mathbf{a}} \right] \\ &= \underset{\tilde{\mathbf{a}}}{\arg\min} -2\tilde{\mathbf{a}}^H \left(\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l)\, \hat{\mathbf{P}}_{\mathbf{x}}(l) \right) \tilde{\mathbf{a}} \\ &= \underset{\tilde{\mathbf{a}}}{\arg\max}\, \tilde{\mathbf{a}}^H \left(\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l)\, \hat{\mathbf{P}}_{\mathbf{x}}(l) \right) \tilde{\mathbf{a}} \end{aligned}, \tag{5.39}$$

where we have used the fact that $\tilde{\mathbf{a}}^H \tilde{\mathbf{a}} = 1$. The solution for $\tilde{\mathbf{a}}$ is the principal eigenvector of $\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l) \left[\hat{\mathbf{P}}_{\mathbf{y}}(l) - \hat{\phi}_\gamma(l)\, \hat{\mathbf{\Gamma}} - \hat{\phi}_v(l)\, \hat{\mathbf{\Psi}} \right]$.

The alternating least squares method using multiple time frames jointly (JALS) is summarized in Algorithm 3.

Algorithm 3: JALS method

```
Input: $\hat{\mathbf{P}}_\mathbf{y}$, $\hat{\boldsymbol{\Gamma}}$,$\hat{\boldsymbol{\Psi}}$,init.$\hat{\tilde{\mathbf{a}}}$, $I$
Output: $\mathbf{a}, \phi$
for all $k, \beta$ do
    for iter=1:I do
        Calculate $\boldsymbol{\Phi}_{\tilde{\mathbf{a}}}$ using Eq. (5.36) and $\tilde{\mathbf{b}}(l)$ using Eq. (5.37).
        Estimate $\tilde{\phi}(l)$ using Eq. (5.35) for each $l$.
        Calculate $\hat{\mathbf{P}}_\mathbf{x}(l) = \hat{\mathbf{P}}_\mathbf{y}(l) - \hat{\phi}_\gamma(l)\hat{\boldsymbol{\Gamma}} - \hat{\phi}_v(l)\hat{\boldsymbol{\Psi}}$.
        Estimate $\tilde{\mathbf{a}}$ using the principal eigenvector of $\sum_{l=1+(\beta-1)N}^{\beta N} \hat{\tilde{\phi}}_s(l)\hat{\mathbf{P}}_\mathbf{x}(l)$.
    Estimate $\mathbf{a}$ and $\phi_s(l)$ using Eq. (5.25) and Eq. (5.26).
```

5.3.4. ROBUST PSD CONSTRAINTS

In [11], it has been shown that linear inequality constraints on the parameters of interest can be used to improve the robustness of the estimation. In [10], the PSD of the source, the PSD of the late reverberation and the PSD of the ambient noise are constrained by Eqs. (5.19) and (5.20). In this section, we introduce more robust constraints on the PSDs to avoid large underestimation and overestimation errors.

UPPER BOUNDS

To avoid large overestimation errors, we can use upper bounds for the PSDs. For the diagonal elements of $\mathbf{P}_y$, it holds that

$$\mathbf{P}_{y_{m,m}}(l) = \tilde{\phi}_s(l)\,|\tilde{a}_m|^2 + \phi_\gamma(l)\,\boldsymbol{\Gamma}_{m,m} + \phi_v(l)\,\boldsymbol{\Psi}_{m,m}. \tag{5.40}$$

Since the three additive terms in Eq. (5.40) are positive, we have

$$\left\{\tilde{\phi}_s(l)\,|\tilde{a}_m|^2, \phi_\gamma(l)\,\boldsymbol{\Gamma}_{m,m}, \phi_v(l)\,\boldsymbol{\Psi}_{m,m}\right\} \le \mathbf{P}_{y_{m,m}}(l), \tag{5.41}$$

for all m. Hence, the upper bound for the PSDs of the target source is

$$\tilde{\phi}_s(l) \le \min_m \left\{ \frac{\mathbf{P}_{\mathbf{y}_{m,m}}(l)}{|\tilde{a}_m|^2} \right\}. \tag{5.42}$$

Similarly, the upper bounds for the PSDs of the late reverberation and the ambient noise are

$$\phi_\gamma(l) \le \min_m \left\{ \frac{\mathbf{P}_{\mathbf{y}_{m,m}}(l)}{\boldsymbol{\Gamma}_{m,m}} \right\}, \tag{5.43}$$

$$\phi_v(l) \leq \min_m \left\{ \frac{\mathbf{P}_{\mathbf{y}_{m,m}}(l)}{\mathbf{\Psi}_{m,m}} \right\}. \tag{5.44}$$

Note that $\mathbf{\Gamma}_{m,m} = 1$ in Eq. (5.7) and that $\mathbf{\Psi}_{m,m} = 1$ when considering only self-noise and each microphone has the same self-noise PSD. In that case we thus have

$$\left\{\phi_\gamma(l), \phi_v(l)\right\} \leq \min_m \left\{\mathbf{P}_{\mathbf{y}_{m,m}}(l)\right\} \leq \frac{\operatorname{tr}(\mathbf{P}_\mathbf{y})}{M}, \tag{5.45}$$

which is tighter than the bound in Eq. (5.20) as used in [10]. Hence, by using Eqs. (5.43) and (5.44), the overestimation errors for the PSDs of the late reverberation and the ambient noise are smaller than the errors using Eq. (5.20), resulting in better speech intelligibility performance [29], [30].

LOWER BOUNDS

To avoid large underestimation errors, we need lower bounds for the PSDs as well. In both [10] and [11], the prior information was used that the PSDs should be positive, setting the lower bounds for all PSDs to ε. That is, when obtaining negative incorrect estimates of the PSDs, these are replaced by the minimum value ε. However, this will lead to very large under estimation errors. Therefore, we propose the use of tighter lower bounds derived from other prior information on the PSDs.

For the normalized PSD of the source $\tilde{\phi}_s$ and the PSD of the late reverberation ϕ_γ, we can see that they have a similar distribution on the time-frequency domain as illustrated in Fig. 5.3.

Based on this, we make the assumption that the ratio between the normalized PSD of the source and the PSD of the late reverberation is bounded on both sides, i.e.

$$C_1 \leq \frac{\tilde{\phi}_s(l)}{\phi_\gamma(l)} \leq \frac{1}{C_2}, \tag{5.46}$$

or

$$C_1 \phi_\gamma(l) \leq \tilde{\phi}_s(l), \tag{5.47}$$

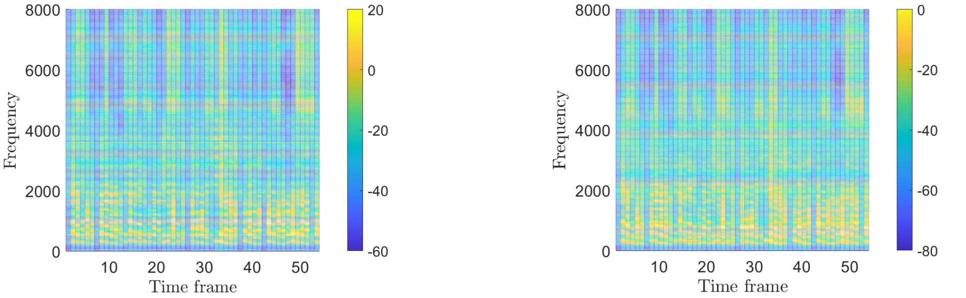
and

$$C_2 \tilde{\phi}_s(l) \leq \phi_\gamma(l), \tag{5.48}$$

for all (l,k) pairs. Using Eqs. (5.47) and (5.48), we can constrain the estimated PSDs of the source and the PSDs of the late reverberation in the following way. We first initialize C_1 and C_2 by an initial value like $C_1 = C_2 = 1$ for the first time frame $l = 1$. For the l-th time frame, we update C_1 and C_2 while making $\hat{\tilde{\phi}}_s(l)$ and $\hat{\phi}_\gamma(l)$ positive in the way shown in Fig. 5.4.

We first update $C_1(l)$ and $C_2(l)$ by

$$C_1(l) = \begin{cases} \min\left\{C_1(l-1), \frac{\hat{\tilde{\phi}}_s(l)}{\hat{\phi}_\gamma(l)}\right\} & \text{if} \hat{\tilde{\phi}}_s(l) > 0, \hat{\phi}_\gamma(l) > 0 \\ C_1(l-1) & \text{else.} \end{cases}, \tag{5.49}$$



(a) Normalized PSD of the target source $\tilde{\phi}_s$ (in dB).

(b) PSD of the late reverberation ϕ_γ (in dB).

Figure 5.3: Time frame and frequency distribution of the target source PSD and the late reverberation PSD. Each time frame has a length of 0.64 s.

and

$$C_2\left(l\right)=\begin{cases} \min\left\{C_2\left(l-1\right), \frac{\hat{\phi}_\gamma(l,k)}{\hat{\tilde{\phi}}_s(l)}\right\} & \text{if} \hat{\tilde{\phi}}_s\left(l\right)>0, \hat{\phi}_\gamma\left(l\right)>0 \\ C_2\left(l-1\right) & \text{else.} \end{cases} \tag{5.50}$$

With $C_1\left(l\right)$ and $C_2\left(l\right)$, we update $\hat{\tilde{\phi}}_s\left(l\right)$ by

$$\hat{\tilde{\phi}}_s\left(l\right)=\begin{cases} \hat{\tilde{\phi}}_s\left(l\right) & \text{if} \hat{\tilde{\phi}}_s\left(l\right)>0 \\ C_1\left(l\right)\hat{\phi}_\gamma\left(l\right) & \text{if} \hat{\phi}_\gamma\left(l\right)>0, \hat{\tilde{\phi}}_s\left(l\right)\leq 0 \\ \hat{\tilde{\phi}}_s^{\min}\left(l\right) & \text{if} \hat{\phi}_\gamma\left(l\right)\leq 0, \hat{\tilde{\phi}}_s\left(l\right)\leq 0 \end{cases} \tag{5.51}$$

where $\hat{\tilde{\phi}}_s^{\min}\left(l\right)$ is calculated by

$$\hat{\tilde{\phi}}_s^{\min} = \min\left\{\left|\frac{\hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\mathbf{P}}_{\mathbf{y}_{m+1,m+1}}}{\left|\hat{\tilde{a}}_m\right|^2 - \left|\hat{\tilde{a}}_{m+1}\right|^2}\right|\right\}_{m=1}^{M-1}, \tag{5.52}$$

where we used the fact that $\mathbf{P}_{\mathbf{y}_{m,m}} = \tilde{\phi}_s|\tilde{a}_m|^2 + \phi_\gamma + \phi_v$ for $m = 1, \cdots, M$ and

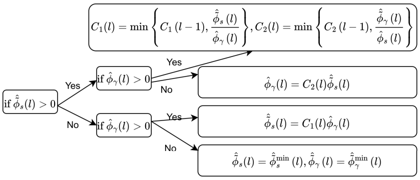

$C_1(l) = \min\left\{C_1(l-1), \frac{\hat{\tilde{\phi}}_s(l)}{\hat{\phi}_\gamma(l)}\right\}, C_2(l) = \min\left\{C_2(l-1), \frac{\hat{\tilde{\phi}}_\gamma(l)}{\hat{\phi}_s(l)}\right\}$
Yes
Yes
if $\hat{\phi}_\gamma(l) > 0$
No
$\hat{\phi}_\gamma(l) = C_2(l)\hat{\tilde{\phi}}_s(l)$
if $\hat{\tilde{\phi}}_s(l) > 0$
No
if $\hat{\phi}_\gamma(l) > 0$
Yes
$\hat{\tilde{\phi}}_s(l) = C_1(l)\hat{\phi}_\gamma(l)$
No
$\hat{\tilde{\phi}}_s(l) = \hat{\tilde{\phi}}_s^{\min}(l), \hat{\tilde{\phi}}_\gamma(l) = \hat{\tilde{\phi}}_\gamma^{\min}(l)$

Figure 5.4: Decision flow for updating $C_1, C_2, \hat{\tilde{\phi}}_s$ and $\hat{\phi}_\gamma$.

$\mathbf{P}_{\mathbf{y}_{m,m}} - \mathbf{P}_{\mathbf{y}_{m+1,m+1}} = \tilde{\phi}_s \left(|\tilde{a}_m|^2 - |\tilde{a}_{m+1}|^2 \right)$. Then, we update $\hat{\phi}_\gamma(l)$ by

$$\hat{\phi}_\gamma(l) = \begin{cases} \hat{\phi}_\gamma(l) & \text{if} \hat{\hat{\phi}}_\gamma(l) > 0 \\ C_2(l)\, \hat{\hat{\phi}}_s(l) & \text{if} \hat{\phi}_s(l) > 0, \hat{\hat{\phi}}_\gamma(l) \le 0 \\ \hat{\phi}_\gamma^{\min}(l) & \text{if} \hat{\phi}_s(l) \le 0, \hat{\phi}_\gamma(l) \le 0 \end{cases} \tag{5.53}$$

where $\hat{\phi}_\gamma^{\min}(l)$ is calculated by

$$\hat{\phi}_\gamma^{\min} = \min \left\{ \left| \hat{\mathbf{P}}_{\mathbf{y}_{m,m}} - \hat{\hat{\phi}}_s^{\min} |\hat{\hat{a}}_m|^2 - \hat{\phi}_v \right| \right\}_{m=1}^{M}. \tag{5.54}$$

For the PSD of the ambient noise, the lower bounds depend on the stochastic property of the noise component. We use the following way to constrain $\phi_v(l)$. First, we give the lower bound C_3 an initial small value ε. Then, we update C_3 as

$$C_3(l) = \begin{cases} \frac{C_3(l-1) + \hat{\phi}_v(l)}{2} & \text{if} \hat{\phi}_v(l) > 0 \\ C_3(l-1) & \text{else} \end{cases}. \tag{5.55}$$

With C_3, we estimate $\phi_v(l)$ by

$$\hat{\phi}_v(l) = \begin{cases} \hat{\phi}_v(l) & \text{if} \hat{\phi}_v(l) > 0 \\ C_3(l-1) & \text{else} \end{cases}, \tag{5.56}$$

Note that the above procedure dealing with non-positive estimates of the PSDs might give us values larger than the upper bounds we derived before in Eqs. (5.42) to (5.44). Therefore, we first execute the above procedure and then upper bound all the estimates.

5.4. EXPERIMENTS

In this section, we will evaluate our proposed ALS-based methods in various scenarios. In addition to the ALS method proposed in [10], we introduce in Section 5.4.1 two more reference methods, namely JMLE [14] and SCFA [11]. In Section 5.4.2, we present the evaluation metrics for all methods. We compare the performance of all methods in various scenarios in Sections 5.4.3 and 5.4.4.

5.4.1. REFERENCE METHODS

The two reference methods introduced here are both based on the maximum likelihood (ML) cost function:

$$\min \sum_{l=1}^{N} \log |\mathbf{P}_\mathbf{y}(l)| + \operatorname{tr}\left(\hat{\mathbf{P}}_\mathbf{y}(l)\, \mathbf{P}_\mathbf{y}^{-1}(l) \right). \tag{5.57}$$

JMLE

In our recent work [14], we assumed a noiseless scenario and proposed a joint maximum likelihood estimator (JMLE) to estimate the RTF of the target source, the PSDs of the target source and the PSDs of the late reverberation jointly. The JMLE method performances well and has low computational complexity. However, the performance of JMLE is not robust for low SNR scenarios due to the noiseless signal model assumed in [14], which is

$$\mathbf{P}_\mathbf{y}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\boldsymbol{\Gamma}. \tag{5.58}$$

SCFA

The last reference method we use for comparison is the simultaneous confirmatory factor analysis (SCFA) method [11]. SCFA performs well in reverberant and noisy environments. However, SCFA comes with a high computational cost due to solving the following non-convex optimization problem

$$\begin{aligned} &\operatorname*{arg\,min}_{\substack{\phi_s(l),\mathbf{a}(\beta)\\ \phi_\gamma(l),\phi_v}} \sum_{l=1}^{N} \log|\mathbf{P}_\mathbf{y}(l)| + \mathrm{tr}\left(\hat{\mathbf{P}}_\mathbf{y}(l)\mathbf{P}_\mathbf{y}^{-1}(l)\right) \\ &\text{s.t. } \mathbf{P}_\mathbf{y}(l) = \phi_s(l)\mathbf{a}(\beta)\mathbf{a}^H(\beta) + \phi_\gamma(l)\boldsymbol{\Gamma} + \phi_v\mathbf{I}, \\ &\quad a_1(\beta) = 1, \phi_s(l) \geq 0, \phi_\gamma(l) \geq 0, \phi_v \geq 0, \end{aligned}, \tag{5.59}$$

where $\phi_v\mathbf{I}$ corresponds to the microphone self noise, which is assumed to be white Gaussian noise. In [11], the above optimization problem is computed iteratively. At each iteration, the parameters are updated and the cost function value is reduced by solving a non-linear constrained optimization problem. The updating procedure is terminated when meeting a local minimum. Note that due to the non-convexity of the optimization problem, the number of iterations needed is large. Hence the computational cost of this method is relatively high.

5.4.2. EVALUATION METRICS

ESTIMATION ERRORS

The first evaluation metric is the estimation error of the parameters of interest. For the RTF vector, we calculate the Hermitian angles between the estimated RTFs and the true RTFs and average them over different frequency bins and time segments, that is,

$$E_\mathbf{a} = \frac{\sum_{\beta=1}^{B}\sum_{k=1}^{K/2+1} \arccos\left(\frac{|\mathbf{a}(\beta,k)^H\hat{\mathbf{a}}(\beta,k)|}{\|\mathbf{a}(\beta,k)\|_2\|\hat{\mathbf{a}}(\beta,k)\|_2}\right)}{\mathrm{B}\left(K/2+1\right)}. \tag{5.60}$$

Note that this metric evaluates the alignment of the estimated RTF with the ground-truth RTF, but cannot reflect scaling errors. For all types of PSDs, we use the symmetric log-error distortion measure [31]

$$E_i = \frac{10 \sum\limits_{\beta=1}^{B} \sum\limits_{l=1+(\beta-1)N}^{\beta N} \sum\limits_{k=1}^{K/2+1} \left|\log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right|}{\mathrm{B}N\left(K/2+1\right)}, \tag{5.61}$$

with $i \in \{s, \gamma, v\}$. In the following experiments, we will also show the detailed PSD estimation performance by using the overestimating errors (denoted as $E_{\phi_i}^{\mathrm{ov}}$) and the underestimation errors (denoted as $E_{\phi_i}^{\mathrm{un}}$) as used in [29]

$$E_i^{\mathrm{ov}} = \frac{10 \sum\limits_{\beta=1}^{B} \sum\limits_{l=1+(\beta-1)N}^{\beta N} \sum\limits_{k=1}^{K/2+1} \left|\min\left\{0, \log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right\}\right|}{\mathrm{B}N\left(K/2+1\right)}, \tag{5.62}$$

and

$$E_i^{\mathrm{un}} = \frac{10 \sum\limits_{\beta=1}^{B} \sum\limits_{l=1+(\beta-1)N}^{\beta N} \sum\limits_{k=1}^{K/2+1} \max\left\{0, \log\left(\frac{\phi_i(l,k)}{\hat{\phi}_i(l,k)}\right)\right\}}{\mathrm{B}N\left(K/2+1\right)}. \tag{5.63}$$

Note that, typically, large underestimation errors in the source PSDs and large overestimation errors in the noise PSDs can cause large target source distortions when applying the estimates in a noise reduction framework. Also, large underestimation errors in the noise PSD are likely to cause musical noise [29]. We therefore also quantify the performance in terms of predicted quality and intelligibility when used in combination with a noise reduction algorithm, as explained below.

PREDICTED QUALITY AND INTELLIGIBILITY

We can construct the following multi-channel Wiener filter (MWF) [32] based on the estimated parameters to extract the target signal,

$$\hat{\mathbf{w}}\left(l\right) = \frac{\hat{\phi}_s\left(l\right)\hat{\mathbf{w}}_{\mathrm{MVDR}}\left(l\right)}{\hat{\phi}_s\left(l\right) + \hat{\mathbf{w}}_{\mathrm{MVDR}}\left(l\right)^H \hat{\mathbf{R}}_{nn}\left(l\right)\hat{\mathbf{w}}_{\mathrm{MVDR}}\left(l\right)}, \tag{5.64}$$

where $\mathbf{w}_{\mathrm{MVDR}}(l)$ is the minimum variance distortionless response (MVDR) beamformer [33]

$$\hat{\mathbf{w}}_{\mathrm{MVDR}}(l) = \frac{\hat{\mathbf{R}}_{nn}^{-1}(l)\hat{\mathbf{a}}(l)}{\hat{\mathbf{a}}(l)^H \hat{\mathbf{R}}_{nn}^{-1}(l)\hat{\mathbf{a}}(l)}, \tag{5.65}$$

and

$$\hat{\mathbf{R}}_{nn}(l) = \hat{\phi}_\gamma(l)\hat{\mathbf{\Gamma}} + \hat{\phi}_v(l)\hat{\mathbf{\Psi}}, \tag{5.66}$$

and where $\hat{\mathbf{w}}(l)$ is used as $\hat{s}(l) = \hat{\mathbf{w}}(l)^H \mathbf{y}(l)$. After estimating $\hat{s}(l)$, the time domain signal is reconstructed by calculating the IFFT followed by an overlap-add procedure. Note that for the JMLE method, $\hat{\mathbf{R}}_{nn}(l) = \hat{\phi}_\gamma(l)\hat{\mathbf{\Gamma}}$ due to its noiseless signal model.

After applying the MWF filter to the noisy signal, we obtain the estimated target signal and evaluate the noise reduction performance using the segmental signal-to-noise-ratio

(SSNR) [34], the speech intelligibility performance using the speech intelligibility in bits (SIIB) measure [35], [36] and the speech quality performance using the perceptual evaluation of speech quality (PESQ) measure [37].

COMPUTATION TIME

The last evaluation metric is the computational time comparison between our proposed methods and the reference methods.

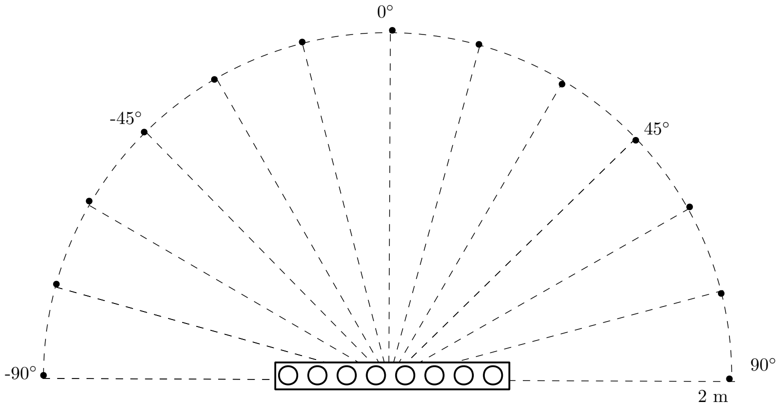


Figure 5.5: Geometric setup for the real RIRs.

5.4.3. EXPERIMENTS WITH SIMULATED NOISE

SETUP

We use speech signals originating from the TIMIT database [38] and recorded RIRs to simulate realistic acoustic scenarios. The RIRs are downloaded from the database in [39], which were recorded in a room with size 6 × 6 × 2.4 m. The geometric setup for the recording is shown in Fig. 5.5. The sound source was placed 2 m away from the center of the uniform linear microphone array at 0°. This array has 8 microphones and 8 cm inter-distances. At each microphone, we synthesize the reverberant signal by convolving the speech source (with a duration of 35 s) with the corresponding RIR. Subsequently, we add noise components to the reverberant signals simulating the microphone noise at specified signal-to-noise ratios (SNRs) to synthesize the microphone signals. In the following, we will consider white Gaussian noise to simulate microphone selfnoise with variance σ_v^2 calculated from given SNR values for each microphone. Since the signal is non-stationary, we calculate the SNR by averaging the target signal-to-noise ratio over the whole time duration.

In this experiment, we used the following parameters setting: The sampling rate is $f_s = 16$ kHz. The sampled noisy microphone signals are processed by the STFT for each sub-time frame. As analysis and synthesis window we use the square-root Hann window with a length of 32 ms with 50% overlap between adjacent sub-time frames. Note that each time frame consists of $T_{sf} = 40$ overlapping sub-time frames and has a duration of 0.64 s. The FFT length is 512. The speed of sound is set to 344 m/s. Note that the first 512 samples of the RIRs are used to calculate the true RTFs as these parts of the RIRs fall within each current sub-time frame and the remaining parts are considered as the late reverberation. Note that for ALS-based methods, we use the same random vector as an initial estimate of the RTF for the first time frame in a time segment (ALS and MALS) or for a time segment (JALS).

RESULTS

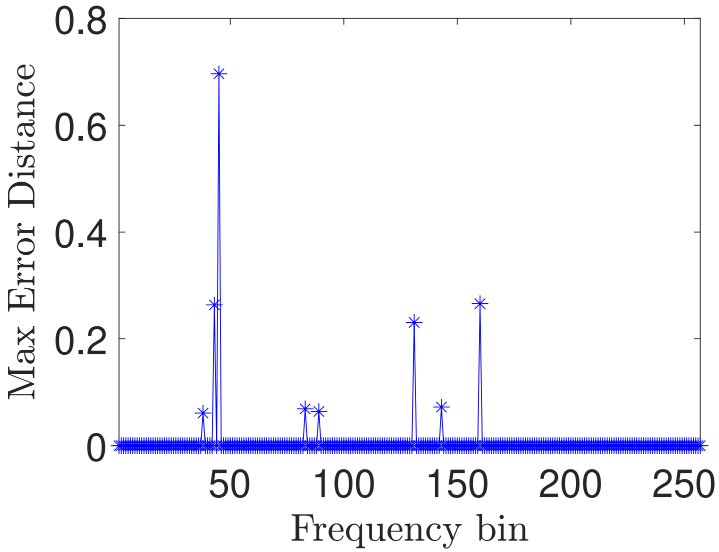


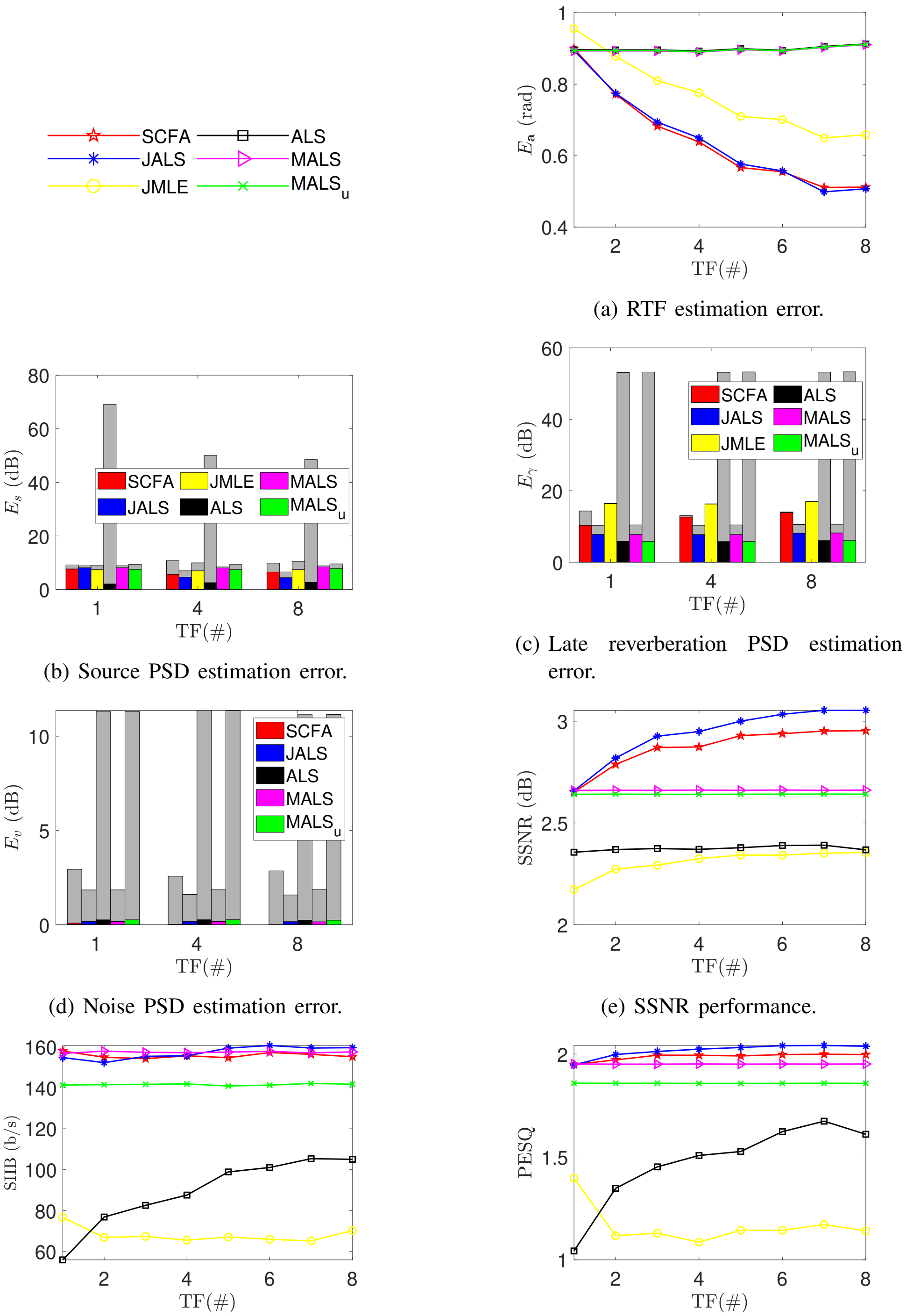
Figure 5.6: Maximum error distance for different frequency bins.

We first show the sensitivity of the ALS method to the initialization in Fig. 5.6. For each frequency bin, we generate the initial RTF estimates randomly and independently 100 times. Each time we generate two Gaussian distributed vectors as the real part and the imaginary part of a complex random vector and normalize it with its first element. With the initial RTF estimate, we apply the ALS method and obtain an estimated RTF. We calculate the Hermitian angle error of the estimated RTF for each initial RTF estimate, resulting in 100 different RTF errors for each frequency bin. We subtract the maximum with the minimum of these errors as the maximum error distance in Fig. 5.6. As shown in this figure, the initialization does not have a significant impact on the RTF estimation errors for the ALS method.

In Fig. 5.7, we compare our proposed methods with all the other reference methods as a function of the number of time frames in a time segment varying from 1 to

8. The reverberation time is 0.61 s and the SNR is fixed at 0 dB. To evaluate how the robust constraints of the PSDs proposed in Section 5.3.4 help the estimation of the parameters of interest, we also included the modified ALS method without using the robust constraints in Section 5.3.4 but using Eqs. (5.19) and (5.20), referred to as MALS$_u$. When using only a single time frame in each time segment, the RTF estimation errors for the ALS-based methods and the SCFA method have similar values which are much lower than the JMLE method as shown in Fig. 5.7a. The reason is that JMLE was derived from a noiseless signal model [40]. The signal model mismatch error is thus large for the JMLE in a 0 dB environment. When increasing the number of time frames in each time segment, the RTF estimation errors for ALS, MALS$_u$ and MALS (the three ALS-based methods using a single time frame) do not vary much; while the RTF estimation errors for JALS, JMLE and SCFA (methods using multiple time frames) become much lower. The RTF estimation errors $E_{\mathbf{a}}$ for methods using a single time frame fluctuate slightly because the first time frame of a time segment use random initial estimate of the RTF. The other time frames use the estimate in the previous time frame as the initial estimate. $E_{\mathbf{a}}$ for ALS and MALS$_u$ are close to MALS due to the normalization process in the Hermitian angle metric in Eq. (5.60). The drawback of the Hermitian angle metric is that any scaled estimate will have the same value. The bad scaling of ALS can be reflected in the target source PSD estimation errors, where ALS has much larger errors compared to the other methods as shown in Fig. 5.7b. In Figs. 5.7c and 5.7d, we can see that MALS$_u$ has similar performance with ALS, which both use the PSDs constraints in Eqs. (5.19) and (5.20). While, MALS using the robust constraints of the PSDs proposed in Section 5.3.4 has much lower errors compared to ALS and MALS$_u$. As expected, the PSDs estimation errors do not change much as a function of the number of frames in a segment since the PSDs are time frame variant parameters. In Figs. 5.7b to 5.7d, we show the underestimation error and overestimation error for the PSDs. Our proposed methods (MALS and JALS) have improved performance compared to ALS and similar performance compared to SCFA. As shown, ALS has the worst underestimation errors for all the PSDs. This is due to the lack of a normalization step and using the value ε to replace negative values in the ALS method. JMLE has the largest overestimation errors for PSDs of the late reverberation. This is due to the noiseless signal model that is assumed with JMLE. In a low SNR environment, the JMLE method considers the ambient noise as late reverberation and gives larger values when estimating the PSDs of the late reverberation. For noise reduction performance evaluated by SSNR in Fig. 5.7e, our proposed JALS has the best performance, which is slightly better than SCFA but much better than the other methods. For the speech intelligibility performance evaluated by SIIB and the speech quality performance evaluated by PESQ in Figs. 5.7f and 5.7g, the proposed JALS, MALS and the reference method SCFA outperform the other methods.

In Fig. 5.8, we compare all the methods while changing the variance of the ambient noise component such that the SNR increases from 0 dB to 40 dB. The reverberation time is 0.36 s and each time segment contains 8 time frames. As shown in Fig. 5.8a, the RTF estimation errors become lower for all methods when the SNR becomes larger. SCFA and our proposed method JALS have the best overall performance, which is much better than methods using a single time frame (ALS and MALS). For low SNR, JMLE



(a) RTF estimation error.

(b) Source PSD estimation error.

(c) Late reverberation PSD estimation error.

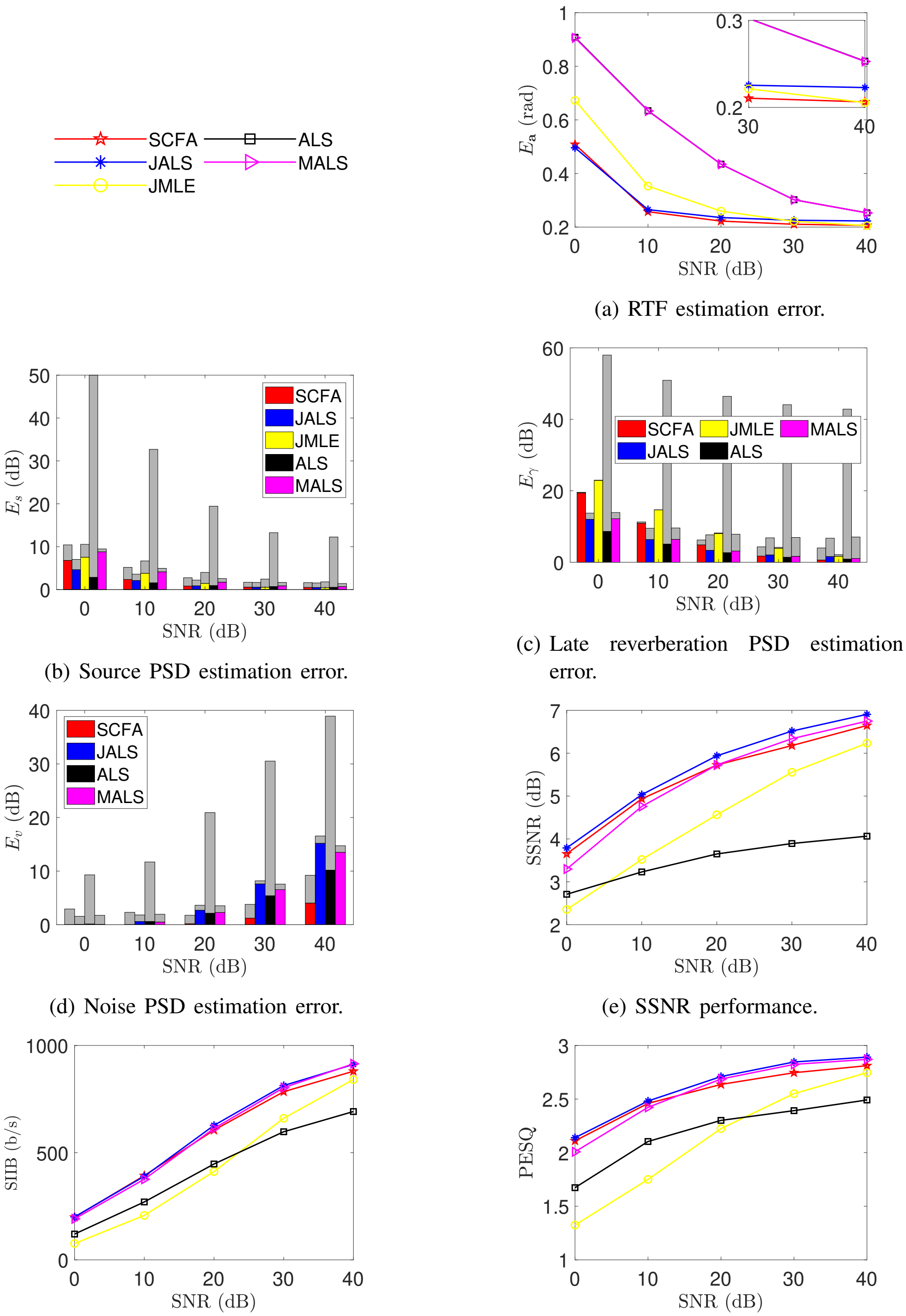
(d) Noise PSD estimation error.

(e) SSNR performance.

(f) SIIB performance.

(g) PESQ performance.

Figure 5.7: Performance vs the number of time frames. In Figs. b, c and d, the gray bars indicate the underestimation errors, the colored bars indicate overestimation errors and the methods from left to right are SCFA, JALS, JMLE(in Figs. b and c), ALS and MALS.



(a) RTF estimation error.

(b) Source PSD estimation error.

(c) Late reverberation PSD estimation error.

(d) Noise PSD estimation error.

(e) SSNR performance.

(f) SIIB performance.

(g) PESQ performance.

Figure 5.8: Performance vs SNR. In Figs. b, c and d, the gray bars indicate the underestimation errors and the colored bars indicate overestimation errors.

is worse than JALS, but when increasing the SNR, JMLE improves the fastest as its model mismatch error is smaller and has a smaller RTF estimation error than JALS for 40 dB SNR. We can see that when the signal model mismatch error is neglectable, the MLE-based methods (SCFA and JMLE) perform better than the ALS-based methods (JALS). For the PSDs estimation errors in Figs. 5.8b to 5.8d, SCFA has the best performance with only JMLE reaching a similar performance for high SNR scenarios. Our proposed ALS-based methods (MALS and JALS) perform much better than ALS. For noise reduction and speech intelligibility performance in Figs. 5.8e to 5.8g, MALS and JALS have similar performance with SCFA and much better performance than ALS. When increasing the SNR, JMLE has the most significant improvement and gets close to the performance of MALS, JALS and SCFA for 40 dB SNR.

Table 5.1: Computation time comparison.

method	SCFA	ALS	MALS	JMLE	JALS
Normalized run time	154.65	6.27	5.7	1.66	1

We also evaluate the computation time for all methods and average these over all cases shown in Fig. 5.8. Then, we averaged and normalized the run time for all methods with respect to the run time for JALS as shown in Table 5.1. We sort the run time for all the methods in descending order from left to right. As expected, SCFA is the most time-consuming method. JALS and JMLE are the two fastest methods. The computational cost mainly comes from the inversion of a 3×3 matrix (complexity of order 3^3) and the eigenvalue decomposition of an $M\times M$ matrix (complexity of order M^3) for the ALS-based methods (ALS, MALS and JALS). For the case that each time segment has N time frames, ALS and MALS process each time frame separately and execute I iterations N times. Hence, they have a complexity of order $I\times N\times\left(3^3+M^3\right)$. For JALS, we only need to calculate the eigenvalue decomposition I times. Hence, its complexity order is $I\times M^3+I\times N\times 3^3$. The complexity order of JMLE is $(N+I)\times M^3$ [40]. In this experiment, we have $M=8$, $N=8$ and $I=10$. Therefore, the time cost ratio among ALS/MALS, JMLE and JALS is $I\times N\times\left(3^3+M^3\right):(N+I)\times M^3:I\times M^3+I\times N\times 3^3\approx 5.92:1.27:1$, which is approximately similar to the real averaged run time ratio in Table 5.1.

5.4.4. EXPERIMENTS WITH RECORDED NOISE

SETUP

In this experiment, we generate the RIRs using the image source method [41]. The dimension of the room is $7\times5\times4$ m. In this simulated room, we have a single speaker, four microphones and a recorded wash machine noise from the ESC-50 database [42] as shown in Fig. 5.9. Note that we also added a white Gaussian noise to each microphone signal to simulate the microphone selfnoise at a SNR of 50 dB. The other settings are the same as those of Experiment 1. For ALS-based methods, we assume an ideal voice activity detector is used and the spatial coherence matrix of the ambient noise is

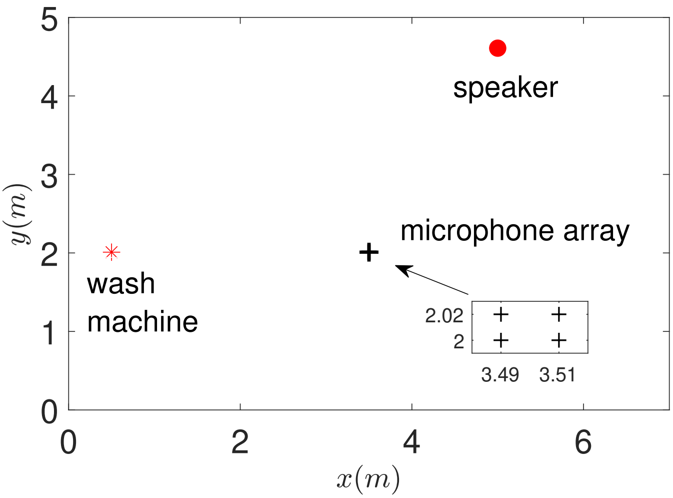


Figure 5.9: Top view of the acoustic scene with a zoom-in of microphones.

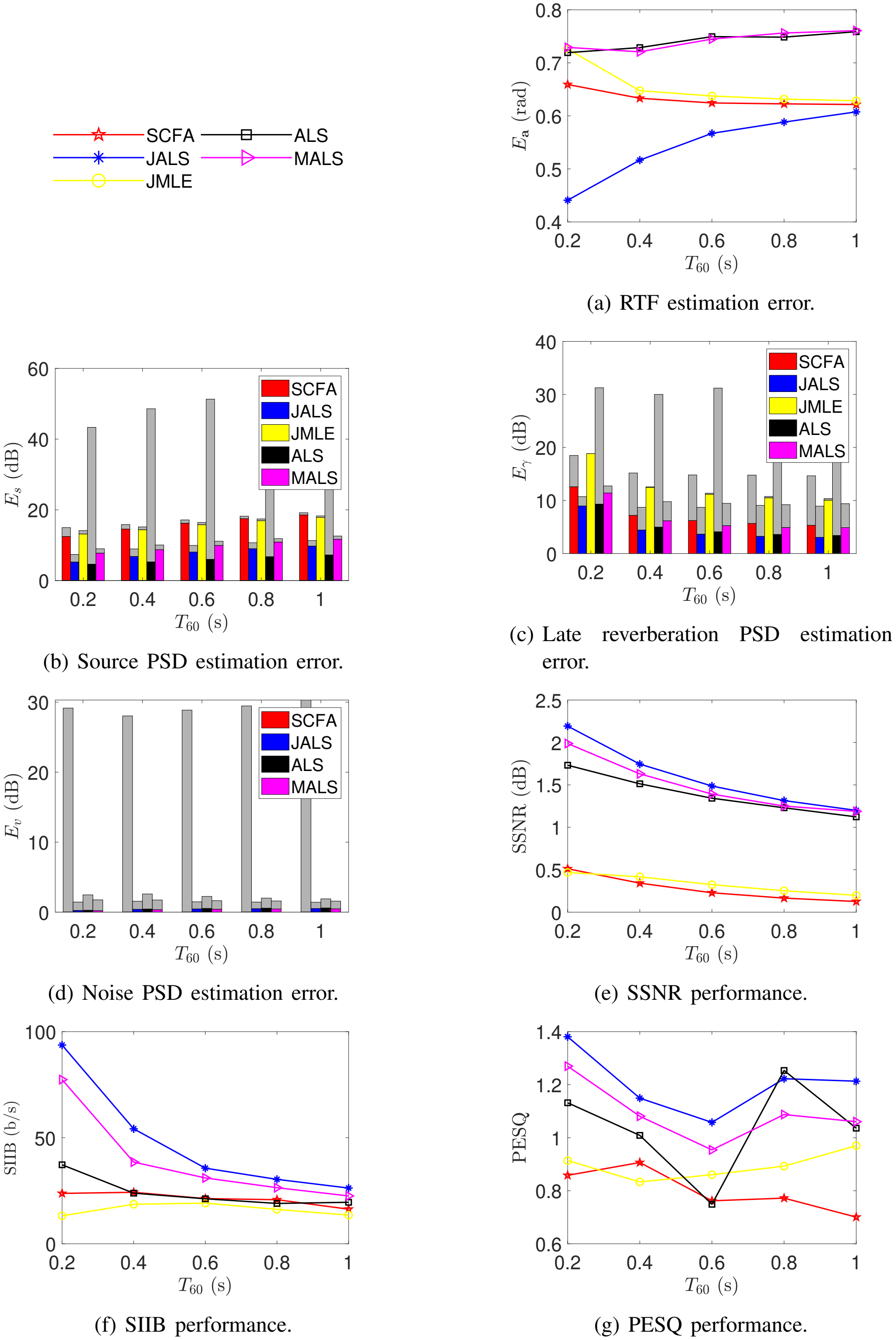
calculated using the noise only time frame with the following equation

$$\boldsymbol{\Psi}_{i,j}(k)=\frac{\sum_{t_n=1}^{T_n} y_i(t_n,k)\,y_j(t_n,k)^*}{\sqrt{\left(\sum_{t_n=1}^{T_n}|y_i(t_n,k)|^2\right)\left(\sum_{t_n=1}^{T_n}|y_j(t_n,k)|^2\right)}},\tag{5.67}$$

with $|x|$ the absolute value of x and $\{i,j\}$ the microphone indices. For SCFA, the spatial coherence matrix of the ambient noise is modeled as the identity matrix in [11]. For JMLE, the ambient noise is not considered. Hence, these two methods will have sever model mismatch errors in this experiment. Note that SCFA can be extended to handle spatial coherence matrices different from the identity matrix. However, it takes some effort to calculate the gradient and the Hessian matrix of the cost function and will not be addressed in this work.

RESULTS

In Fig. 5.10, we compare all the methods while changing the reverberation time T_{60} of the RIRs from 0.2 s to 1 s. Each time segment contains 8 time frames. We can see that our proposed JALS method has the best performance in all the metrics evaluated. For the RTF estimation error in Fig. 5.10a, the ALS-based methods ALS, MALS and JALS have degraded performance as T_{60} increases. However, SCFA and JMLE have improved performance. This is due to the model mismatch caused by the ambient noise component. When increasing T_{60}, the ratio between the correctly modeled late reverberation component and the incorrectly modeled ambient noise component becomes



(a) RTF estimation error.

(b) Source PSD estimation error.

(c) Late reverberation PSD estimation error.

(d) Noise PSD estimation error.

(e) SSNR performance.

(f) SIIB performance.

(g) PESQ performance.

Figure 5.10: Performance vs T_{60}. In Figs. b, c and d, the gray bars indicate the underestimation errors and the colored bars indicate overestimation errors.

larger. For the PSDs estimation errors of the target source and the late reverberation in Figs. 5.10b and 5.10c, SCFA and JMLE have large over estimation errors due to considering the ambient noise as the target source and the late reverberation. The ALS method still has the worst performance in Figs. 5.10b and 5.10c. While, for the noise PSD estimation error in Fig. 5.10d, SCFA has the worst performance due to erroneous spatial coherence matrix used. In Figs. 5.10e to 5.10g, our proposed multiple time frames method JALS has improved performance over our single time frame method MALS, which both outperform all the other reference methods.

5.5. CONCLUDING REMARKS

We proposed alternating least square (ALS) based methods to estimate the RTFs, the PSDs of the source, the PSDs of the late reverberation, and the PSDs of the ambient noise jointly for a single reverberant and noisy scenario. We first modified an existing ALS method to obtain more accurate estimates using a single time frame. Then, we extend the method to use multiple time frames that share the same RTF jointly. Furthermore, we imposed more robust constraints on the estimated PSDs. Experimental results demonstrated that the proposed methods achieve similar performance compared to the SCFA method in terms of estimation accuracy, noise reduction performance, speech quality, and speech intelligibility. The proposed methods outperform the existing ALS-based method and the JMLE method assuming a noiseless signal model, especially in low SNR scenarios.

REFERENCES

[1] K. L. Payton, R. M. Uchanski, and L. D. Braida, "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing", *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994.

[2] J. Xia, B. Xu, S. Pentony, J. Xu, and J. Swaminathan, "Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners", *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1523–1533, Mar. 2018.

[3] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010.

[4] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From Blind to Guided Audio Source Separation: How models and side information can improve the separation of sound", *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 107–115, 2014.

[5] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[6] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[7] B. Schwartz, S. Gannot, and E. A. Habets, "Two model-based EM algorithms for blind source separation in noisy environments", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[8] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[9] I. Kodrasi and S. Doclo, "Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[10] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[11] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[12] Y. Laufer and S. Gannot, “Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[13] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[14] C. Li, J. Martinez, and R. C. Hendriks, “Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 695–705, 2023.

[15] A.-J. Van Der Veen, “Joint diagonalization via subspace fitting techniques”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, IEEE, vol. 5, 2001, pp. 2773–2776.

[16] K. Rahbar and J. Reilly, “A frequency domain method for blind source separation of convolutive audio mixtures”, *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 832–844, 2005.

[17] S. Degerine and E. Kane, “A Comparative Study of Approximate Joint Diagonalization Algorithms for Blind Source Separation in Presence of Additive Noise”, *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 3022–3031, 2007.

[18] J. S. Bradley, H. Sato, and M. Picard, “On the importance of early reflections for speech in rooms”, *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[19] R. Talmon, I. Cohen, and S. Gannot, “Relative Transfer Function Identification Using Convolutive Transfer Function Approximation”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[20] Y. Avargel and I. Cohen, “System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[21] S. Braun and E. A. Habets, “Dereverberation in noisy environments using reference signals and a maximum likelihood estimator”, in *Proc. EURASIP Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[22] E. A. P. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields”, *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

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

[24] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[25] B. F. Cron and C. H. Sherman, “Spatial-correlation functions for various noise models”, *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[26] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank”, *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936.

[27] A. M. Sardarabadi and A.-J. van der Veen, "Complex factor analysis and extensions", *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 954–967, 2018.

[28] T. Söderström and P. Stoica, *System Identification*. Prentice-Hall International, 1989.

[29] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay", *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[30] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[31] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement", in *Proc. Interspeech*, 2007, pp. 830–833.

[32] H. L. V. Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., Mar. 2002.

[33] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

[34] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[35] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory", *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, 2017.

[36] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[37] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", *Rec. ITU-T P. 862*, 2001.

[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[39] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments", in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sep. 2014.

[40] C. Li, J. Martinez, and R. C. Hendriks, "Low Complex Accurate Multi-Source RTF Estimation", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4953–4957.

[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[42] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification", in *Proc. ACM Int. Conf. Multimed.*, Brisbane, Australia: ACM Press, 2015, pp. 1015–1018, ISBN: 978-1-4503-3459-4.

6

LOW COMPLEX ACCURATE MULTI-SOURCE RTF ESTIMATION

This chapter is based on the article published as: C. Li, J. Martinez, and R. C. Hendriks. “Low Complex Accurate Multi-Source RTF Estimation”. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. 2022, pp. 4953–4957.

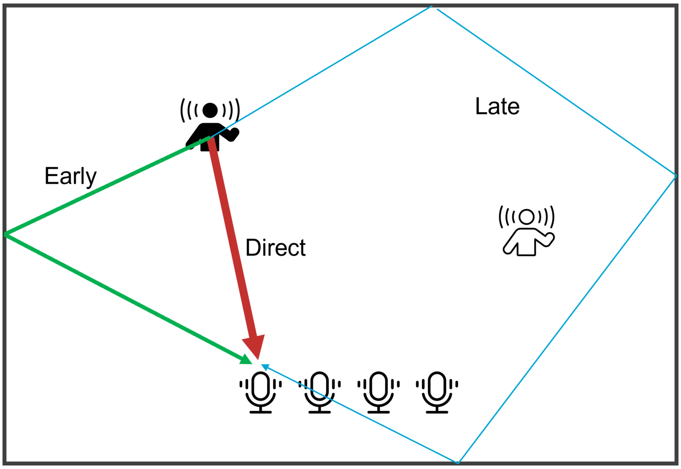


Figure 6.1: Illustration of multi-source reverberant scenario.

In previous chapters, we only considered a single source scenario. In this chapter and the next chapter, we will consider a multi-source scenario. In this chapter, we assume the late reverberation has little energy and the environment is noiseless as illustrated in Fig. 6.1. We will propose an RTF estimator to answer research question 2.1 shown in Fig. 1.4 using the signal model presented in Fig. 2.2 (c).

Many multi-microphone algorithms depend on knowing the relative acoustic transfer functions (RTFs) of the individual sound sources in the acoustic scene. However, accurate joint RTF estimation for multiple sources is a challenging problem. Existing methods to jointly estimate the RTF for multiple sources have either no satisfying performance, or, suffer from a very large computational complexity. In this chapter, we propose a method for robust estimation of the individual RTFs in a multi-source acoustic scenario. The presented algorithm is based on linear algebraic concepts and therefore of lower computational complexity compared to a recently presented state-of-the-art algorithm, while having a similar performance. Experimental results are presented to demonstrate the RTF estimation performance as well as the noise reduction performance when combining the estimated RTFs with a beamformer.

6.1. INTRODUCTION

Microphone arrays are ubiquitous these days and can be used for applications like source separation [1]–[3], dereverberation [4]–[6], noise reduction [7]–[10] and sources localization [11]. These applications have in common that they heavily rely on acoustic-scene dependent parameters like relative acoustic transfer functions (RTFs), power spectral densities (PSDs) of the sources, PSDs of the late reverberation and PSDs of the microphone self-noise. In particular the RTF plays a very important role in beamforming applications. knowing and having an accurate estimate of the RTF per source is very important, for example, to steer a beamformer in the right direction [12], or preserve the spatial cues in binaural noise reduction algorithms [8]. However, accurate RTF estimation is also rather challenging. In this chapter we therefore specifically focus on estimating the RTFs and present an algorithm to jointly estimate the individual RTFs of the sources in the acoustic scene.

RTF estimation for a single point source in noise is a problem that has been addressed before in several papers, e.g. [13]–[15]. In this work, we consider the more general and more challenging case of simultaneously RTF estimation for multiple sources. A few methods have been proposed for multiple source RTF estimation in recent years [16]–[18]. In [16], the RTFs are estimated by updating the initial estimate of the RTFs in an iterative fashion. However in reality, the a priori information of the RTFs might be unknown. In [17], the expectation maximization (EM) method is used to estimate the RTFs by assuming that, in each time-frequency bin, only a single source is active, which thus puts limitations on the acoustic scenarios. In [18], a simultaneous confirmatory factor analysis (SCFA) method was proposed to estimate the RTFs and also the PSDs of sources, late reverberation and the microphone self-noise jointly. However, due to the non-convexity of the problem formulation, the SCFA method in [18] has a rather high computational cost and is therefore currently less applicable for real-time applications.

To accurately estimate the RTFs jointly for multiple sources, our starting point is the algorithm proposed in [1]. This algorithm was developed for blind source separation and is based on linear algebraic concepts. We start with presenting the method from [1], but from a different perspective, such that our proposed algorithm can be better understood. Next, we propose a more robust method, which is also based on linear algebraic concepts and has relatively low computational complexity. The simulations demonstrate that our method is more accurate compared to the reference algorithm [1] and of much lower complexity compared to the state-of-the-art SCFA method from [18], while having a comparable performance.

6.2. PRELIMINARIES

6.2.1. SIGNAL MODEL

We consider R acoustic point sources observed by a microphone array consisting of M microphones with an arbitrary geometric structure under the assumption that the signal-to-noise ratio (SNR), i.e., the SNR due to the diffuse noise, is relatively high, the late reverberation is neglectable and the number of microphones is larger than the number of the sources (i.e., $M > R$). In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone can be modelled as

$$y_m(i,k) = \sum_{r=1}^{R} a_{mr}(\beta,k)\, s_r(i,k), \tag{6.1}$$

where i is the time-frame index, k is the frequency bin index and $a_{mr}(\beta,k)$ is the m-th element of the RTF vector $\mathbf{a}_r(\beta,k)$ corresponding to source s_r in time segment β at microphone m. In this work, we differentiate between time segments (indexed by β) and time frames (indexed by i). Each time segment consists of multiple time frames. We assume that the RTF vector is constant during a time segment (thus during multiple time frames that fall within one segment) and $a_{1r} = 1$ for $r = 1,...,R$, which means that the first microphone is selected as the reference microphone. Stacking the M microphone STFT coefficients into a vector, we have

$$\mathbf{y}(i,k) = \sum_{r=1}^{R} \mathbf{a}_r(\beta,k)\, s_r(i,k) \in \mathbb{C}^{M\times 1}. \tag{6.2}$$

We assume that all the sources are mutually uncorrelated for each frame of a time segment, which leads to the following second-order statistical signal model

$$\mathbf{P}_\mathbf{y}(i,k) = \sum_{r=1}^{R} p_r(i,k)\mathbf{a}_r(\beta,k)\mathbf{a}_r^H(\beta,k) \in \mathbb{C}^{M\times M}, \tag{6.3}$$

where $p_r(i,k) = E\left[|s_r(i,k)|^2\right]$ is the power spectral density (PSD) of the r-th source at the reference microphone. The covariance matrix can be rewritten in the following matrix form

$$\mathbf{P}_\mathbf{y}(i,k) = \mathbf{A}(\beta,k)\,\mathbf{P}(i,k)\,\mathbf{A}^H(\beta,k), \tag{6.4}$$

where the RTF matrix is given by

$$\mathbf{A}(\beta,k) \;=\; [\mathbf{a}_1(\beta,k),\cdots,\mathbf{a}_R(\beta,k)] \tag{6.5}$$

and the PSD matrix is given by

$$\mathbf{P}(i,k) = \mathrm{diag}\left[p_1(i,k),\cdots,p_R(i,k)\right]. \tag{6.6}$$

The main goal of this chapter is to estimate the RTF matrix $\mathbf{A}(\beta,k)$ using estimated covariance matrices $\{\mathbf{P}_\mathbf{y}(i,k)\}$ with $i = 1+(\beta-1)N,\cdots,\beta N$, where N is the number of time frames in a time segment.

6.2.2. COVARIANCE MATRIX ESTIMATION

In addition to time frames and time segments, we now also define sub time-frames. Each time frame consists of N_s overlapping sub-frames indexed by n_s with equal length T_s, where the sub-frame length is much smaller than the time frame length such that N_s is a large integer. Assuming the signal is stationary across a time frame, we can estimate the covariance matrix per time frame i based on the sample covariance matrix using the sub-frames' samples, i.e.,

$$\hat{\mathbf{P}}_{\mathbf{y}}(i,k) = \frac{1+(i-1)N_s}{iN_s}\sum_{n_s=1}^{N_s}\mathbf{y}(n_s,k)\,\mathbf{y}(n_s,k)^H, \tag{6.7}$$

where $\mathbf{y}(n_s,k)$ is the STFT coefficient vector. Notice that within the time frames of one time segment, the RTF matrix is a constant matrix and the PSDs of the sources are assumed to be non-stationary, which means that the signal powers can change over the frames.

6.3. RTF ESTIMATION

In Section 6.3.2, we propose an improved algorithm to estimate the RTF matrix. The starting point is the method presented in [1], which is originally meant for blind source separation. Since the RTF is defined per frequency and per time segment, from now on, frequency indices and time segment indices are neglected for ease of notation.

We first write the covariance matrices $\mathbf{P}_{\mathbf{y}}(i)$ into the form

$$\mathbf{P}_{\mathbf{y}}(i) = \tilde{\mathbf{A}}(i)\,\tilde{\mathbf{A}}^H(i), \text{for } i = 1,\cdots,N, \tag{6.8}$$

where $\tilde{\mathbf{A}}(i) = \mathbf{A}\sqrt{\mathbf{P}(i)}$ and the diagonal matrix $\sqrt{\mathbf{P}(i)}$ is the unique non-negative square root of $\mathbf{P}(i)$. Note that $\mathbf{A}$ equals the normalized version of matrix $\tilde{\mathbf{A}}(i)$ where the columns of $\mathbf{A}(i)$ are normalized with respect to their first element, which is the square root of the PSD of each corresponding source. Hence, estimation of $\mathbf{A}$ and $\mathbf{P}(i)$ can be converted into the estimation of $\tilde{\mathbf{A}}(i)$ for any time frame i. With this conversion, the covariance matrices for all the other time frames in the same segment can be represented by $\tilde{\mathbf{A}}(i)$. That is

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(j) &= \mathbf{A}\mathbf{P}(j)\,\mathbf{A}^H\\ &= \mathbf{A}\sqrt{\mathbf{P}(i)}\sqrt{\mathbf{P}^{-1}(i)\mathbf{P}(j)}\sqrt{\mathbf{P}^{-1}(i)}\sqrt{\mathbf{P}(i)}\mathbf{A}^H\\ &= \tilde{\mathbf{A}}(i)\,\tilde{\mathbf{P}}(j)\,\tilde{\mathbf{A}}^H(i), \text{for } j = 1,\cdots,N,\end{aligned} \tag{6.9}$$

where $\tilde{\mathbf{P}}(j) = \sqrt{\mathbf{P}^{-1}(i)\mathbf{P}(j)}\sqrt{\mathbf{P}^{-1}(i)}$ is a diagonal matrix.

6.3.1. JOINT DIAGONALIZATION METHOD

We first summarize in this section the joint diagonalization method from [1] to put our work in perspective. This method was originally proposed for blind source separation and used in, e.g., [1], [19], to estimate the mixing matrix instead of the RTF matrix. Therefore, although the estimation steps are the same as in [1], we summarize this method when used in a different context to better understand our proposed method that we present in Section 6.3.2.

The method in [1] focuses on estimating $\tilde{\mathbf{A}}(1)$. Then, matrices $\mathbf{P}_{\mathbf{y}}(i)$ in the segment can be represented by $\tilde{\mathbf{A}}(1)$ using

$$\mathbf{P}_{\mathbf{y}}(i) = \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{P}}(i)\,\tilde{\mathbf{A}}^{H}(1), \text{for } i = 2, \cdots, N, \tag{6.10}$$

where

$$\tilde{\mathbf{P}}(i) = \sqrt{\mathbf{P}^{-1}(1)}\,\mathbf{P}(i)\sqrt{\mathbf{P}^{-1}(1)} \tag{6.11}$$

is diagonal. Notice that $\mathbf{P}_{\mathbf{y}}(1) = \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{A}}^{H}(1)$.

Consider the singular value decomposition (SVD) of $\tilde{\mathbf{A}}(1)$, i.e.,

$$\tilde{\mathbf{A}}(1) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{H}, \tag{6.12}$$

where $\mathbf{U}$ is an $M \times R$ complex sub-unitary matrix (i.e., $\mathbf{U}^{H}\mathbf{U} = \mathbf{I}$), $\boldsymbol{\Sigma}$ is a $R \times R$ diagonal matrix and $\mathbf{V}$ is a complex valued $R \times R$ unitary matrix. The estimation of $\tilde{\mathbf{A}}$ is decomposed into the estimation of the three matrices $\mathbf{U}$, $\boldsymbol{\Sigma}$ and $\mathbf{V}$.

The estimates of $\mathbf{U}$ and $\boldsymbol{\Sigma}$ can be obtained from $\mathbf{P}_{\mathbf{y}}(1)$. Using the SVD of $\tilde{\mathbf{A}}(1)$ in (6.8), $\mathbf{P}_{\mathbf{y}}(1)$ can be expressed as:

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(1) &= \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{A}}^{H}(1)\\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{H}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^{H}\\ &= \mathbf{U}\boldsymbol{\Sigma}^{2}\mathbf{U}^{H}.\end{aligned} \tag{6.13}$$

Since $\mathbf{U}$ is a sub-unitary matrix and $\boldsymbol{\Sigma}^{2}$ is a diagonal matrix, (6.13) is an eigenvalue decomposition of the matrix $\mathbf{P}_{\mathbf{y}}(1)$. Hence we can calculate $\mathbf{U}$ and $\boldsymbol{\Sigma}$ by taking the EVD of $\mathbf{P}_{\mathbf{y}}(1)$.

The estimation of $\mathbf{V}$ can be solved by using estimated $\mathbf{U}$, $\boldsymbol{\Sigma}$, and the covariance matrices for all other time frames in the same segment. Taking the SVD of $\tilde{\mathbf{A}}(1)$ in (Eq. (6.10)), $\mathbf{P}_{\mathbf{y}}(i)$ for $i = 2, \cdots, N$ can be expressed as

$$\begin{aligned}\mathbf{P}_{\mathbf{y}}(i) &= \tilde{\mathbf{A}}(1)\,\tilde{\mathbf{P}}(i)\,\tilde{\mathbf{A}}^{H}(1)\\ &= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{H}\tilde{\mathbf{P}}(i)\,\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^{H}.\end{aligned} \tag{6.14}$$

Now we construct a new set of matrices $\mathbf{P}_{\mathbf{w}}(i)$ using $\mathbf{U}$ and $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathbf{P}_{\mathbf{w}}(i) &= \boldsymbol{\Sigma}^{-1}\mathbf{U}^{H}\mathbf{P}_{\mathbf{y}}(i)\,\mathbf{U}\boldsymbol{\Sigma}^{-1}\\ &= \mathbf{V}^{H}\tilde{\mathbf{P}}(i)\,\mathbf{V}.\end{aligned} \tag{6.15}$$

As $\mathbf{V}$ is an orthogonal matrix, it can be obtained by computing the eigenvectors of the matrices $\{\mathbf{P}_{\mathbf{w}}(i)\}$, with $i = 2, \cdots, N$.

If $N = 2$, we can estimate $\mathbf{V}$ by taking the EVD of $\mathbf{P_w}(2)$. In case of equal eigenvalues, the corresponding eigenvectors are not unique. Hence, in order to obtain the correct estimate of $\mathbf{V}$, we need to assume that the diagonal matrix $\tilde{\mathbf{P}}(i)$ has distinct diagonal elements, which means that the following inequalities should be satisfied for every two sources r_1 and r_2,

$$\frac{p_{r_1}(2)}{p_{r_1}(1)} \neq \frac{p_{r_2}(2)}{p_{r_2}(1)}, \tag{6.16}$$

where $p_r(i)$ denotes the PSD of the r_{th} source in the i_{th} time frame.

If $N > 2$, the estimation of $\mathbf{V}$ becomes a joint diagonalization problem: find a unitary matrix $\mathbf{V}$ such that $\left\{\mathbf{V}\mathbf{P_w}(i)\mathbf{V}^H\right\}$ with $i = 2, \cdots, N$ is a set of diagonal matrices or has minimal off-diagonal elements. The Jacobi-like algorithm proposed in [20] can be used to solve this joint diagonalization problem, which reduces the original joint diagonalization problem into finite sub-problems having closed-form solutions (see [20] for more details). To make sure the joint diagonalization problem has a satisfying solution, we also need to make an assumption on the PSDs of the sources: for every source r_1, there exists one time frame i_0 such that the following inequality holds for any other source r_2:

$$\frac{p_{r_2}(i_0)}{p_{r_2}(1)} \neq \frac{p_{r_1}(i_0)}{p_{r_1}(1)} \quad \text{for any } r_2 \neq r_1. \tag{6.17}$$

Finally, we estimate $\tilde{\mathbf{A}}(1)$ by multiplying the estimated $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ as in (Eq. (6.12)). Normalizing $\tilde{\mathbf{A}}(1)$, we eventually obtain the estimate of the RTF matrix $\mathbf{A}$. Note that having estimated $\tilde{\mathbf{A}}(1)$, we can also estimate the individual source PSDs. To do so, we use the diagonal elements of $\mathbf{V}\mathbf{P_w}(i)\mathbf{V}^H$ to estimate $\tilde{\mathbf{P}}(i)$. Using the definition $\tilde{\mathbf{A}}(1) = \mathbf{A}\sqrt{\mathbf{P}(1)}$ in combination with (Eq. (6.11)) we obtain the PSDs of all sources for all time frames in the segment.

This algorithm is summarized in Algorithm description 1.

Algorithm 4: Joint Diagonalization Method (JOINT)

```
Input: Estimated $\hat{\mathbf{P}}_\mathbf{y}(i)$, for $i = 1, \cdots, N$,
Output: $\mathbf{A}$ and $\mathbf{P}(i)$ for $i = 1, \cdots, N$,
Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\hat{\mathbf{P}}_\mathbf{y}(1)$.
Construct new matrices $\mathbf{P_w}(i)$ for $i = 2, \cdots, N$.
Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$ using the Jacobi-like algorithm [20].
Estimate $\tilde{\mathbf{A}}(1)$ by multiplying $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
Estimate $\mathbf{A}$ by normalizing $\tilde{\mathbf{A}}(1)$ with its first row.
Estimate $\mathbf{P}(i)$ for $i = 1, \cdots, N$ using the first row of $\tilde{\mathbf{A}}(1)$ and $\tilde{\mathbf{P}}(i)$ for
  $i = 2, \cdots, N$.
```

6.3.2. ROBUST JOINT DIAGONALIZATION

The algorithm introduced in Section 6.3.1 focuses on estimating the RTF matrix $\mathbf{A}$ using the estimated covariance matrices $\hat{\mathbf{P}}(i)$ for $i = 1, \cdots, N$. However, instead of using the individual matrices $\hat{\mathbf{P}}(i)$ as done in the first step in Algorithm 1, we can also choose to estimate the RTF matrix from any linear combination of estimated covariance matrices in segment β . By using an average of estimated covariance matrices instead of a single estimated $\mathbf{P}_{\mathbf{y}}(1)$ in step 1 from Algorithm 1, we are able to significantly reduce the estimation error on estimating $\mathbf{A}$ if we are able to also select the best estimated covariance matrices to form this average. To see this, let us first look at the error on the estimated covariance matrix $\Delta\mathbf{P}_{\mathbf{y}}(i)$. This error can be decomposed into:

$$\Delta\mathbf{P}_{\mathbf{y}}(i) = \mathbf{A}\left(\mathbf{P}(i) - \hat{\mathbf{P}}(i)\right)\mathbf{A}^H - \mathbf{E}(i), \tag{6.18}$$

where the first part $\Delta\mathbf{P}(i) = \left(\mathbf{P}(i) - \hat{\mathbf{P}}(i)\right)$ is indeed the estimation error between the sampled covariance matrix and the true covariance matrix of sources, and the second part $\mathbf{E}(i)$ is due to the late reverberation component and the microphone self noise component, which can be assumed to be positive definite.

It is well known that the estimation error between a sampled covariance matrix and the true covariance matrix can be reduced by increasing the number of samples. Hence, to decrease $\Delta\mathbf{P}(i)$, we can average covariance matrices for as many time frames as possible in a time segment. However, the second error matrix $\mathbf{E}(i)$ might increase when using more time frames. The question now is, which estimated $\mathbf{P}_{\mathbf{y}}(i)$ for $i = 1, \cdots, N$, should we average to replace $\hat{\mathbf{P}}_{\mathbf{y}}(1)$ in step 1 from Algorithm 1 to reduce the estimation error. Notice that the rank of the true covariance matrix $\mathbf{P}_{\mathbf{y}}(i)$ is R, the rank of the estimated covariance matrix $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ is M and we have assumed that $M > R$. Therefore the $R+1$ largest eigenvalue $\lambda_{R+1}(i)$ of $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ can be used to evaluate how large the error matrix $\mathbf{E}(i)$ is.

Based on the analysis of the estimation error of covariance matrices, the next steps of the robust joint diagonalization algorithm are as follows: Take the EVD for the N estimated covariance matrices $\hat{\mathbf{P}}_{\mathbf{y}}(i)$ from a segment and reorder the time frame index such that $\lambda_{R+1}(i)$ is in an ascending order. Use the first estimated covariance matrix (i.e., the one with the smallest error $\mathbf{E}$) to do Algorithm 1 and obtain the first estimates of the RTF matrix $\hat{\mathbf{A}}_1$ and PSDs of the R sources $\left\{\hat{\mathbf{P}}_{\mathbf{1}}(i)\right\}$. Use these estimates to calculate the following weighted cost function:

$$C(1) = \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_{R+1}^2(i)} ||\hat{\mathbf{P}}_{\mathbf{y}}(i) - \hat{\mathbf{A}}_1\hat{\mathbf{P}}_1(i)\hat{\mathbf{A}}_1^H||_2, \tag{6.19}$$

where $||\cdot||_2$ denotes the matrix 2-norm. Next, average the first two estimated covariance matrices from the ordered sequence and use this in combination with Algorithm 1 to obtain the second estimates of the RTF matrix and PSDs of the R sources, and calculate the cost function:

$$C(2) = \sum_{i=1}^{N} \frac{1}{\hat{\lambda}_{R+1}^2(i)} ||\hat{\mathbf{P}}_{\mathbf{y}}(i) - \hat{\mathbf{A}}_2\hat{\mathbf{P}}_2(i)\hat{\mathbf{A}}_2^H||_2, \tag{6.20}$$

In each next iteration, we include an additional covariance matrix from the ordered sequence in the average and use the averaged covariance matrix in combination with Algorithm 1 to estimate the RTF matrix and PSDs of R sources until all the N covariance matrices are averaged and N cost function values are calculated. We then select the minimum cost function value with respect to iteration q, and use the estimate of the RTF matrix in the q_{th} iteration as the final estimate of the RTF matrix.

The algorithm steps are given in algorithm two. Note that the computational cost of the proposed algorithm is about N times higher than for Algorithm 1.

```
Algorithm 5: Robust Joint Diagonalization (PROP)
Input: Estimated $\mathbf{P_y}(i)$, for $i = 1, \cdots, N$,
Output: $\mathbf{A}$
Estimate $\lambda_{R+1}$ from EVD of $\mathbf{P_y}(i)$, for $i = 1, \cdots, N$.
Reorder time frame index such that $\lambda_{R+1}$ is ascending.
for $q = 1 : N$ do
    Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\sum_{i=1}^{q} \frac{1}{q}\hat{\mathbf{P}}_{\mathbf{y}}(i) = \tilde{\mathbf{A}}_{\mathbf{q}}\tilde{\mathbf{A}}_{\mathbf{q}}^{H}$.
    Construct new matrices $\mathbf{P_w}(i)$ for $i = 2, \cdots, N$.
    Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(i)$ for $i = 2, \cdots, N$ using the Jacobi-like algorithm [20].
    Estimate $\tilde{\mathbf{A}}_{\mathbf{q}}$ by multiplying $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
    Estimate $\mathbf{A}_q$ by normalizing $\tilde{\mathbf{A}}_{\mathbf{q}}$ with its first row.
    Estimate $\mathbf{P}(i)$ using the first row of $\tilde{\mathbf{A}}_{\mathbf{q}}$ and $\tilde{\mathbf{P}}(i)$ for $i = 1, \cdots, N$.
    Use the estimate to calculate the cost function Eq. (6.19)
Find the minimum cost function value with respect to the $q_{th}$ estimate of $\mathbf{A}$ and
  use it as the final estimate of the RTF matrix.
```

6.4. EXPERIMENTS

The performance of the proposed methods is evaluated in the context of noise reduction with four microphones and three sources each with a duration of 25 s. The acoustic setup is depicted in Fig. 6.2. Each speech signal is convolved with a room impulse response in the time domain. The room impulse responses are generated using the image method [21]. To simulate a nearly non-reverberant noisy signal, we set the reflection coefficients of the six walls as $[0.5, -0.25, 0.1, -0.5, 0.25, -0.1]$ in the first scenario (the reverberation time is about 0.04 s). Besides, we also evaluate the performance of our proposed methods in a second scenario where the reverberation time of the room impulse response is 0.2 s. The sampling frequency is $f_s = 16$ kHz. The microphone self-noise is a zero-mean uncorrelated Gaussian process with variance σ_v^2, such that the SNR due to the self-noise is equal to the values as specified in Fig. 6.3 per microphone. The noisy speech signal is converted into the STFT domain using a square-root Hann window with a length of 800 samples (i.e. 50 ms) and an overlap of 50%. The FFT length is 1024. Note that the true RTF matrix is calculated using the 1024-length FFT coefficients of the first 800 samples of the room impulse responses. Each time segment consists of $N = 8$ time frames and each time frame consists of $N_s = 40$ sub frames. For comparison, we used the SCFA method from [18] and the original joint diagonalization method from [1] as a reference as SCFA and JOINT, respectively. The proposed method will be referred to as PROP.

The RTF estimation error is evaluated by the Hermitian angle [22].

$$\frac{\sum\limits_{r=1}^{R}\sum\limits_{\beta}^{\mathrm{B}}\sum\limits_{k=1}^{K/2+1} \operatorname{acos}\left(\frac{\left|\mathbf{a}_r^H(\beta,k)\hat{\mathbf{a}}_r(\beta,k)\right|}{\left\|\mathbf{a}_r^H(\beta,k)\right\|_2\left\|\hat{\mathbf{a}}_r(\beta,k)\right\|_2}\right)}{R\mathrm{B}\left(K/2+1\right)}\,(\mathrm{rad})\,, \tag{6.21}$$

where K and B are the number of frequency bins and time segments, respectively. In Fig. 6.3(a), we show the estimation performance in the nearly no reverberation case (with subscript 'nr'), and $T_{60} = 0.2s$ (with subscript 'r'). For both scenarios, PROP and SCFA have a similar and much better performance compared to JOINT. For the nearly no reverberation case, SCFA has a somewhat better performance than PROP, because SCFA can model microphone self-noise and can better reduce the model mismatch error caused mainly by the diffuse noise. However, for the $T_{60} = 0.2s$ and high SNR case, PROP has a slightly better estimation performance than SCFA, because the model mismatch error now is mainly caused by the late reverberation component, which is not considered in the referenced version of SCFA. For the $T_{60} = 0.2s$ case, we also evaluated the noise reduction performance in combination with three minimum variance distortionless response (MVDR) beamformers [23], where we use each time one of the three estimated RTFs as the target and the remaining two sources as interferers. We then calculate the segmental-signal-to-noise-ratio (SSNR) and average this over the three sources. Note that for the SSNR calculation, we omit the sub frames in which the signal energy is zero. In addition to the methods PROP, JOINT and SCFA we also show the performance when using the true RTF. As shown in Fig. 6.3(b), the SSNR for each method increases as the SNR increases. PROP has an almost similar performance compared to SCFA, while both PROP and SCFA improve over JOINT with slightly less than 1 dB in terms of SSNR. In Table 6.1 we show the normalized computation time for all methods after averaging the run time over all scenarios. As expected, the runtime for PROP is about $N = 8$ times larger than for JOINT, but PROP is significantly less complex than SCFA.

Table 6.1: Computation time comparison.

method	SCFA	PROP	JOINT
Normalized run time	1	0.0163	0.0024

6.5. CONCLUSIONS

We considered the problem of estimating the RTF for multiple sources jointly. We proposed a robust method which averages covariance matrices for as many time frames

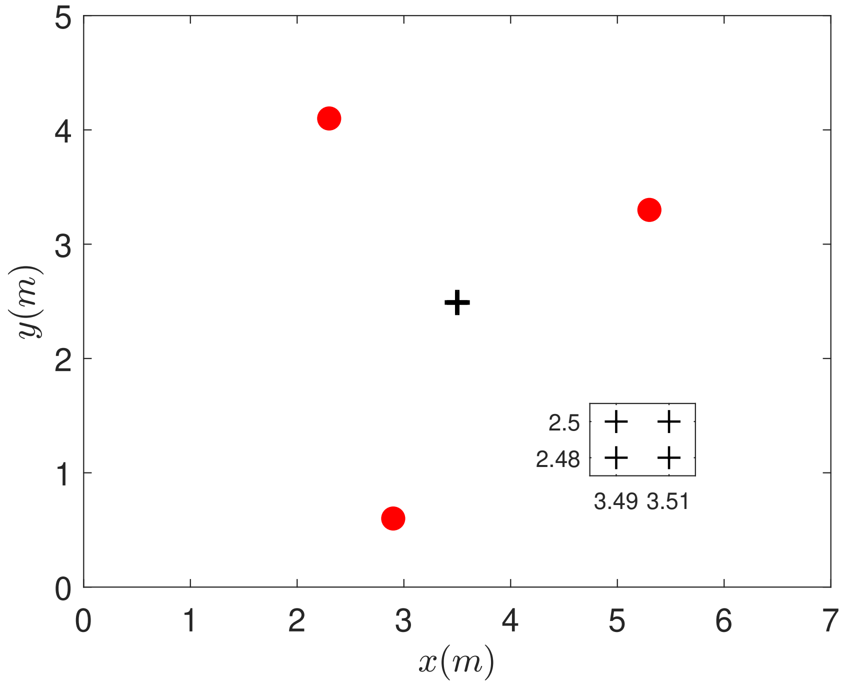


Figure 6.2: Acoustic scene. The three red circles denote the sources. The cross in the center denotes the set of microphones. A zoom-in of that set of four microphones is provided in the little square.

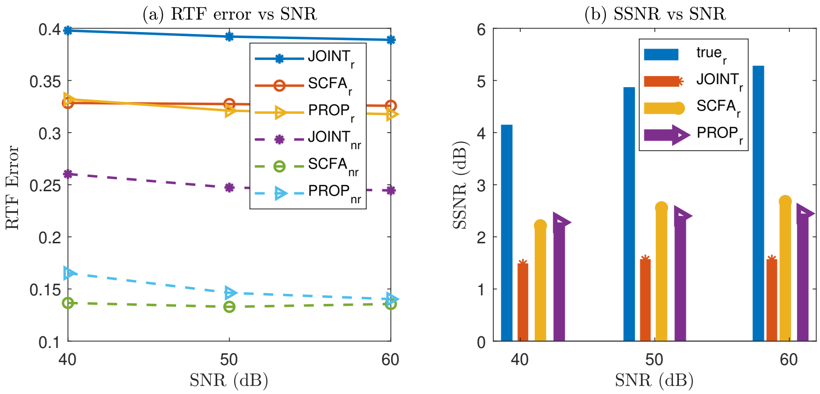


Figure 6.3: RTF estimation error and SSNR vs SNR.

as possible without suffering too much from model mismatch errors caused by late reverberation and microphone self noise. Experiments show that the RTF estimation performance of the proposed method is similar to the SCFA method, but at a significantly lower complexity, and much better than the joint diagonalization method from [1].

REFERENCES

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics”, *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.

[2] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources”, *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[3] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Frequency-domain blind source separation of many speech signals using near-field and far-field models”, *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[4] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[5] I. Kodrasi and S. Doclo, “Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[6] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, “Evaluation and comparison of late reverberation power spectral density estimators”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[7] S. Markovich, S. Gannot, and I. Cohen, “Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[8] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Relaxed binaural LCMV beamforming”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 137–152, 2017.

[9] J. Zhang, S. P. Chepuri, R. C. Hendriks, and R. Heusdens, “Microphone subset selection for MVDR beamformer based noise reduction”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 3, pp. 550–563, 2018.

[10] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, “A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks With Arbitrary Topology”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1434–1448, 2018.

[11] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, “Informed sound source localization using relative transfer functions for hearing aid applications”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[12] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech”, *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[13] I. Cohen, “Relative transfer function identification using speech signals”, *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

[14] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[15] S. Markovich-Golan, S. Gannot, and W. Kellermann, “Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function”, 2018, pp. 2499–2503.

[16] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, “Square Root-Based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 755–769, 2020.

[17] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[18] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust joint estimation of multimicrophone signal model parameters”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[19] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals”, *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[20] J.-F. Cardoso and A. Souloumiac, “Jacobi Angles For Simultaneous Diagonalization.”, *SIAM J. Matrix Anal. Appl.*, vol. 17, Jan. 1996.

[21] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[22] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”, in *Proc. IEEE Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 11–15.

[23] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

7

MULTI-MICROPHONE SIGNAL PARAMETER ESTIMATION IN A MULTI-SOURCE NOISY REVERBERANT SCENARIO

This chapter is based on the article submitted to: C. Li and R. C. Hendriks. "Multimicrophone Signal Parameter Estimation in A Multi-Source Noisy Reverberant Scenario". In: IEEE/ACM Trans. Audio, Speech, Language Process. 33 (2025), pp. 678-692.

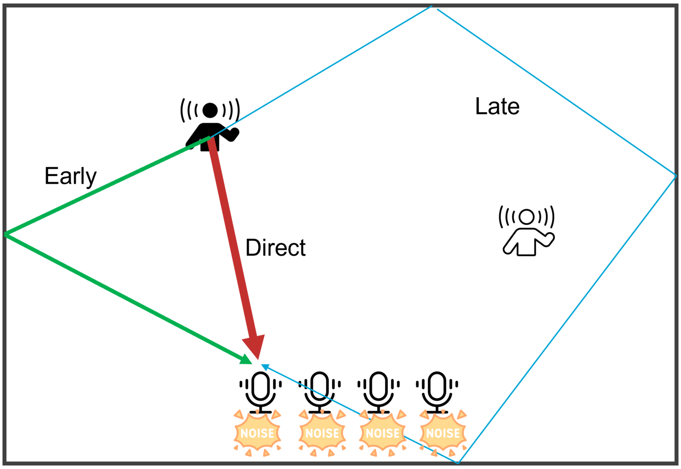


Figure 7.1: Illustration of multi-source, reverberant and noisy scenario.

In this chapter, we consider the most complex scenario in this dissertation, which is the multi-source reverberant and noisy scenario. As illustrated in Fig. 7.1, it can be seen as an extension from the scenario in Chapter 6 to the reverberant and noisy case or from the scenario in Chapters 4 and 5 to the multi-source case. We will propose a method to answer research question 2.2 shown in Fig. 1.4 using the signal model presented in Fig. 2.2 (d).

Estimation of acoustic parameters is of great interest but very challenging in the multichannel microphone signal processing area. Existing methods either assume simple, but less realistic scenarios, or suffer from very high computational costs. In this work, we consider the more general scenario where multiple sources, late reverberation and noise exist concurrently. The parameters of interest include the relative transfer functions (RTFs) of the point sources (both target and interferers) and individual power spectral densities (PSDs) of the sources and the late reverberation. We first propose a robust late reverberation PSD estimator using an iterative compensation scheme. Then, based on an analysis of the variance of the sample covariance matrices, we propose a robust and joint estimator for the sources RTFs and PSDs using multiple time frames that share the same RTFs. We compare the proposed method with the state-of-the-art simultaneously confirmatory factor analysis (SCFA) method and the second order blind identification (SOBI) method. Experiments show that our proposed method reaches the estimation performance of SCFA, which significantly outperforms SOBI, but has much less computational costs compared to SCFA.

7.1. INTRODUCTION

Microphone arrays are widely used in various devices, such as mobile phones, ear/headphones, hearing aids and all sorts of speech recognition applications. Typically, signals recorded by the microphones include not only the direct sounds from one or more point sources, but also reflections and ambient noise. In particular, the late reflections, known as late reverberation, are, next to the direct sound of interfering point sources, harmful to the speech quality and intelligibility [1], [2], even if these late reflections originate from the target source. Therefore, to achieve satisfying speech communication performance, microphone signals are processed by multi-microphone or single-microphone noise reduction and dereverberation algorithms [3], [4]. Multi-microphone noise reduction algorithms typically perform significantly better than their single-microphone counterparts [5] and typically depend on the relative transfer functions (RTFs) of the sources, the power spectral densities (PSDs) of the sources, the late reverberation and the ambient noise. However, in practice these parameters are unknown and their estimation is thus an essential problem for microphone array signal processing.

Many methods have been proposed in recent years to estimate these acoustic parameters [6]–[19]. However, when considering multiple sources and the coexistence of late reverberation and ambient noise, the estimation of the aforementioned parameters can be very challenging. Therefore, many of these works consider simplified signal models [7]–[12], [14], [15], [17]–[19], where either simplifying assumptions are used, or a subset of the parameters is assumed known. For instance, in [9], it is assumed that there is only a single active source in each time-frequency bin. In [8]–[10], [17], [18], either the late reverberation or the noise component is not considered in the model. Note that these methods based on simplified signal models have been widely used in practice, due to their simplicity and the properties of speech signals such as sparsity.

Some works considered a more general signal model, but have some other strict assumptions. For example, in [15], only the direct sound is considered as the target signal. The RTFs are assumed to only depend on the direction of arrival (DOA) of the source position and the microphone array geometry. By further assuming the DOA is known, the RTFs are considered known. However, the early reflections, which are beneficial to speech intelligibility [20], are sometimes included in the target sound. The number of unknown real parameters in each RTF vector is $2(M-1)$ with M the number of microphones. Some methods use prior knowledge (like hearing aids that assume a target in front). When considering both the direct sound and the early reflections as the target signal without prior knowledge of the scene, it is very challenging to estimate the RTFs. In [21], [22], the sound sources are assumed to be active successively and in [23], the interferers are assumed to be active earlier than the target sound source, which means that these methods cannot be used if two or more sources become active simultaneously.

The joint estimation of all the parameters considering multiple sources, late reverberation and ambient noise is achieved in [13] using the simultaneous confirmatory factor analysis (SCFA) method. Although this method is very effective, it comes with a very high computational cost. The goal of this chapter is therefore to develop a method that can estimate the signal parameters (RTFs and PSDs of multiple sources, as well as the late reverberation PSD) at high accuracy and low complexity.

An important aspect of this problem formulation is the estimation of the late reverberation PSD. In [24], a comparison between many state-of-the-art late reverberation PSD estimators was published. All methods in this comparison considered only a single source and the RTF was assumed to be known for the spatial coherence-based methods. In this work, as part of the joint estimation of all unknown parameters, we propose a late reverberation PSD estimator that does not require knowledge on the RTFs. This can be seen as an extension of the method in [10] from a single-source to the multi-source scenarios.

In [25], a low complexity blind source separation method was proposed based on a joint diagonalization of a set of covariance matrices. In the previous chapter, we modified this method to estimate the RTFs of multiple sources in a nearly non-reverberant and noiseless environment. In the current work, we extend the methods from [25] and Chapter 6 to jointly estimate not only the RTFs in a noise-free and non-reverberant environment as in Chapter 6, but to estimate both the RTFs and the PSDs of the sources in a reverberant and noisy environment. Note that eventually, the noise component in this work refers to microphone self-noise. Although not strictly necessary for the proposed method, this is often modelled as spatially white Gaussian noise. Given a set of covariance matrices corresponding to a sequence of time-windows, [25] exploits the covariance matrix of the first time-window and Chapter 6 exploits an average of a subset of these covariance matrices to jointly diagonalize the complete set and then estimate the RTFs. We show in this chapter that any proper linear combination (e.g., a random combination or their average) of these matrices can be used and propose the optimal linear combination that minimizes the variances of the error matrix of the sample covariance matrix.

This chapter is structured in the following way. Section 7.2 presents the signal model, statistical assumptions and problem formulation of this work. In Section 7.3, we will first propose our late reverberation PSD estimator. Then in Section 7.3.2, we modify the second order blind identification (SOBI) method from [25] to our estimation problem. After that, we will analyze the variance of the sample covariance matrices and propose our minimum variance joint diagonalization (MVJD) method to estimate the RTFs and the PSDs of the sources. In Section 7.4, experiments in different scenarios will be presented to compare our proposed method to some state-of-the-art reference methods. Finally, Section 7.5 concludes the chapter.

7.2. SIGNAL MODEL

We consider the presence of R acoustic point sources recorded by a microphone array of M microphones in a reverberant and noisy environment. The number of sources R is assumed known in this work. (In practice, it can be estimated using some existing methods such as [26], [27].) The microphones can be placed compactly with various geometric structures (e.g., linear, circular or spherical). Each microphone records the signals generated from sound sources via both a direct propagation path and (infinite) reflections of surrounding objects (e.g. walls). These signals can be modeled as the convolution between the sound sources and the room impulse response (RIR). In the short-time Fourier transform (STFT) domain, the signal received at the m-th microphone can then be modeled as

$$y_m(l,k) = \underbrace{\sum_{r=1}^{R} x_{mr}(l,k)}_{x_m(l,k)} + \underbrace{\sum_{r=1}^{R} d_{mr}(l,k)}_{d_m(l,k)} + v_m(l,k) \ , \tag{7.1}$$

where l is the time index of the STFT window, which we will refer to as a sub-time frame, and k is the frequency-bin index. In addition to sub-time frames indexed by l, we will later also define time frames and time segments. The source reflections are typically labeled as direct component, early reflections (typically the first 50 ms), and late reflections. When considering the target source, these early reflections are actually beneficial for the speech intelligibility [20]. For a source r, we will therefore consider the direct component and early reflections combined, denoted by $x_{mr}(l,k)$, and differentiate these from the late reflections, denoted by $d_{mr}(l,k)$. The additive noise component is denoted by $v_m(l,k)$. In addition to potential interfering sources, both the late reverberation and additive noise are detrimental to speech intelligibility and quality.

As multiplication in the STFT domain can approximate the convolution in the time domain [28], we can model the r-th source at the m-th microphone as

$$x_{mr}(l,k) = a_{mr}(l,k)\, s_r(l,k) \ , \tag{7.2}$$

where $s_r(l,k)$ contains the direct sound and early reflections at the reference microphone and $a_{mr}(l,k)$ is the relative transfer function (RTF) [28] of the r-th source between the m-th microphone and the reference microphone. Without any limitation, we use the first microphone as our reference (i.e., $a_{1r} = 1$). For the duration that the sources are static relative to the microphone array, we can assume that the RTFs are constant. We refer to this duration as a time segment (TS) indexed by β. In vector form, the multi-microphone signal model is then given by

$$\mathbf{y}(l,k) = \underbrace{\sum_{r=1}^{R} \mathbf{a}_r(\beta,k)\, s_r(l,k)}_{\mathbf{x}(l,k)} + \mathbf{d}(l,k) + \mathbf{v}(l,k) \in \mathbb{C}^{M\times 1} \ , \tag{7.3}$$

where each column vector is stacked with M elements such as $\mathbf{y}(l,k) = [y_1(l,k), \cdots, y_M(l,k)]^T$.

Although speech-related signals $s_r(l,k)$ and $\mathbf{d}(l,k)$ are realizations of non-stationary processes, they can be assumed stationary for a short duration of a time frame (TF). The duration of a TF is much longer than that of the STFT window, which we already denoted as a sub-time frame (SF). Hence, we assume the t-th TF contains T consecutive SFs indexed by l from $l = 1+(t-1)T$ to $l = tT$. In addition, we assume in this work that all sources are static for N consecutive TFs (e.g. $N = 8$ for approximately 2.5

s in our experiments), which means that the β-th TS contains N TFs indexed by t from $t = 1 + (\beta - 1)N$ to $t = \beta N$. The relation between TS, TF and SF is visualized in Fig. 7.2. In the situation that sources are not static for the duration of a TS, we can use an adaptive time-segmentation e.g. as proposed in [29].

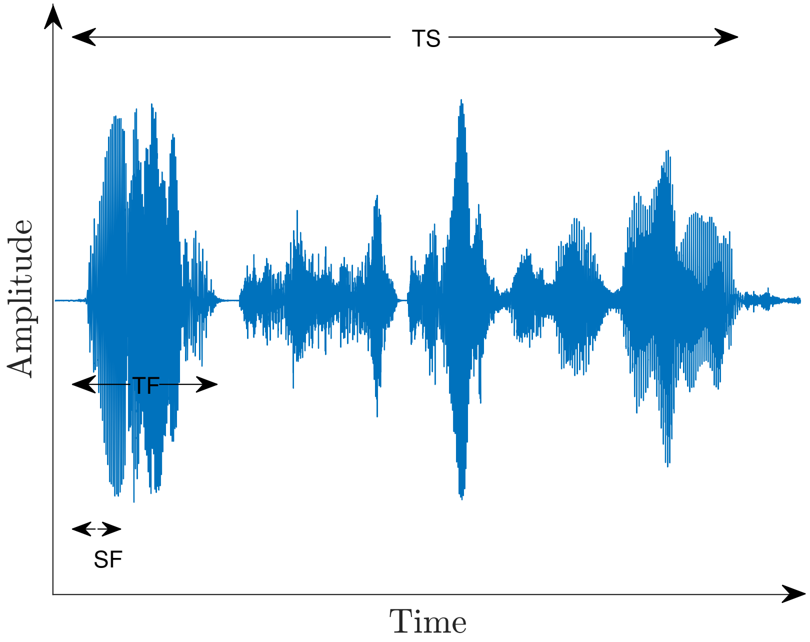


Figure 7.2: Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

Within the t-th TF, the STFT coefficients vector $\mathbf{y}(l,k)$, with sub-frame index $l = 1 + (t-1)T, \cdots, tT$, is assumed to follow a circularly-symmetric complex Gaussian distribution with zero mean and cross power spectral density (CPSD) matrix $\mathbf{P_y}(t,k) \in \mathbb{C}^{M\times M}$. Since $\mathbf{x}(l,k)$, $\mathbf{d}(l,k)$ and $\mathbf{v}(l,k)$ are commonly assumed to be mutually uncorrelated (even though strictly speaking $\mathbf{x}$ and $\mathbf{d}$ are weakly correlated), we can decompose $\mathbf{P_y}(t,k)$ into

$$\begin{aligned}\mathbf{P_y}(t,k) &= \mathrm{E}\left[\mathbf{y}(l,k)\,\mathbf{y}^H(l,k)\right] \\ &= \mathbf{P_x}(t,k) + \mathbf{P_l}(t,k) + \mathbf{P_v}(t,k) \in \mathbb{C}^{M\times M}\ .\end{aligned} \tag{7.4}$$

For the source component $\mathbf{x}(l,k)$, containing the direct and early reflections for all sources, the CPSD matrix $\mathbf{P_x}(t,k)$ is given by

$$\begin{aligned}\mathbf{P}_{\mathbf{x}}(t,k) &= \sum_{r=1}^{R} \phi_r(t,k)\,\mathbf{a}_r(\beta,k)\,\mathbf{a}_r^H(\beta,k) \\ &= \mathbf{A}(\beta,k)\,\mathbf{P}(t,k)\,\mathbf{A}^H(\beta,k)\end{aligned} \tag{7.5}$$

with $\mathbf{A}(\beta,k) = [\mathbf{a}_1(\beta,k),\cdots,\mathbf{a}_R(\beta,k)]$, $\mathbf{P}(t,k) = \mathrm{diag}[\phi_1(t,k),\cdots,\phi_R(t,k)]$ and $\phi_r(t,k) = \mathrm{E}\left[|s_r(l,k)|^2\right]$ the power spectral density (PSD) of the r-th source at the reference microphone with $|\cdot|$ denoting the absolute value. Note that in Eq. (7.5), we used the assumption that all sources are mutually uncorrelated and made explicit that the RTFs $\mathbf{a}_r(\beta,k)$ are constant over a time segment β.

For the late reverberation component, $\mathbf{P}_{\mathbf{l}}(t,k)$ is commonly assumed to be the product of a time-invariant full rank spatial coherence matrix $\mathbf{\Gamma}(k)$ and a time-varying PSD $\phi_\gamma(t,k)$ [7], [30], that is,

$$\mathbf{P}_{\mathbf{l}}(t,k) = \phi_\gamma(t,k)\,\mathbf{\Gamma}(k)\ . \tag{7.6}$$

Here, $\mathbf{\Gamma}(k)$ is assumed to be measured or calculated *a priori* since it is time-invariant and independent of the microphone array position [31]–[33]. For instance, if a spherically isotropic noise field is assumed [34] and inter-microphone distances are assumed known, $\mathbf{\Gamma}(k)$ can be calculated to be

$$\mathbf{\Gamma}_{i,j}(k) = \mathrm{sinc}\left(\frac{2\pi f_s k}{K}\frac{d_{i,j}}{c}\right)\ , \tag{7.7}$$

with $\mathrm{sinc}(x) = \frac{\sin x}{x}$, $d_{i,j}$ the inter-distance between microphones i and j, f_s the sampling frequency, c the speed of sound and K the total frequency bin number. Also note that when a room has ceilings and floors that are more absorbing than the walls, the cylindrical isotropic noise field is a more realistic model. Note that with Eqs. (7.6) and (7.7), $\mathbf{P}_{\mathbf{l}}$ could also model other isotropic noise sources, i.e., noise sources that are not due to the late reverberation.

The noise component $\mathbf{v}$ is usually a summation of the microphone self-noise and other non-point noise sources that are approximately spatially uncorrelated. For this kind of noise, we assume that it has a time-invariant covariance matrix $\mathbf{P}_{\mathbf{v}}(k)$ for each frequency. Therefore, we can also measure $\mathbf{P}_{\mathbf{v}}(k)$ *a priori* by assuming a noise-only segment is available. In this work, we consider only the microphone self-noise to be present with each microphone having the same spatially white Gaussian noise distribution, which means $\mathbf{P}_{\mathbf{v}}(k) = \phi_v\mathbf{I}$. However, notice that we can always introduce a whitening step to guarantee $\mathbf{P}_{\mathbf{v}}(k)$ is spatially white.

With these assumptions, we can now write the covariance matrix of $\mathbf{y}(l)$ as

$$\mathbf{P}_{\mathbf{y}}(t) = \mathbf{A}(\beta)\,\mathbf{P}(t)\,\mathbf{A}^H(\beta) + \phi_\gamma(t)\,\mathbf{\Gamma} + \phi_v\mathbf{I}\ , \tag{7.8}$$

Note that we omitted the frequency indices for legibility in Eq. (7.8) and will do so for all the following equations since the estimators proposed in this work are independent across frequency. Based on the previously discussed stationarity of the signal, we can estimate $\mathbf{P}_{\mathbf{y}}(t)$ using the sample covariance matrix

$$\hat{\mathbf{P}}_{\mathbf{y}}(t) = \frac{1}{T} \sum_{l=1+(t-1)T}^{tT} \mathbf{y}(l)\mathbf{y}(l)^H . \tag{7.9}$$

Note that, to compute an STFT with a meaningful frequency resolution at 16kHz, the subframe duration or STFT window length cannot be too small. Meanwhile, to estimate the second-order statistics in practice, each time frame is composed of many subframes. Therefore, the time frame here is longer than the commonly assumed duration for stationarity. This will lead to an average of the PSDs but will maintain the RTF matrix or the spatial coherence matrix of the signal components [32]. Within the β-th TS, the *a priori* known or estimated parameters from the signal model given in Eq. (7.8) now include N sample covariance matrices $\left\{\hat{\mathbf{P}}_{\mathbf{y}}(t)\right\}_{t=1+(\beta-1)N}^{\beta N}$ (i.e., for the N time frames in segment β), the estimated spatial coherence matrix of the late reverberation $\hat{\mathbf{\Gamma}}$ and the estimated noise PSD $\hat{\phi}_v$. Note that as analyzing the errors of estimated $\hat{\mathbf{\Gamma}}$ is outside of the scope of this work, we assume that $\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$ in the next section. The main goal of this chapter is to develop an algorithm that can estimate the RTF matrix $\mathbf{A}(\beta)$, the diagonal PSD matrices of the sources $\{\mathbf{P}(t)\}_{t=1+(\beta-1)N}^{\beta N}$ and the PSDs of the late reverberation $\left\{\phi_\gamma(t)\right\}_{t=1+(\beta-1)N}^{\beta N}$ for each segment β.

7.3. PARAMETER ESTIMATION

7

In this section, we propose our joint estimator based on a joint diagonalization scheme. We first introduce the estimator of the late reverberation PSDs in Section 7.3.1. Then, we use the estimated late reverberation PSDs and the other *a priori* given parameters to estimate the RTF matrix and the source PSDs in Section 7.3.2.

7.3.1. ESTIMATOR OF THE LATE REVERBERATION PSDS

We assume here that the late reverberation PSDs $\phi_\gamma(t)$ across time frames are unrelated and will estimate these per time frame t using $\hat{\mathbf{P}}_{\mathbf{y}}(t)$ for the t-th time frame only. Hence, for legibility, we will omit the time frame index in this subsection. Subtracting the true noise covariance matrix $\mathbf{P}_{\mathbf{v}}$ from $\mathbf{P}_{\mathbf{y}}$, we get

$$\mathbf{P}_\gamma = \mathbf{P}_{\mathbf{y}} - \mathbf{P}_{\mathbf{v}} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \phi_\gamma \mathbf{\Gamma} . \tag{7.10}$$

Taking the square-root decomposition such as the Cholesky decomposition of the full rank matrix $\mathbf{\Gamma}$, we have $\mathbf{\Gamma} = \mathbf{L}\mathbf{L}^H$. Using $\mathbf{L}$, we can whiten matrix $\mathbf{P}_\gamma$ by calculating

$$\bar{\mathbf{P}}_\gamma = \mathbf{L}^{-1}\mathbf{P}_\gamma\mathbf{L}^{-H} = \left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H + \phi_\gamma \mathbf{I} . \tag{7.11}$$

Since the rank of $\left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H$ is R, we can see that, after whitening, the $M-R$ smallest eigenvalues of $\bar{\mathbf{P}}_\gamma$ should be equal to ϕ_γ. To see this, we can take the eigenvalue decomposition (EVD) of $\left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ with $\mathbf{U}$ unitary and $\mathbf{\Lambda}$ diagonal. The $M-R$ smallest diagonal elements of $\mathbf{\Lambda}$ are all zero. Taking the EVD of $\bar{\mathbf{P}}_\gamma$ using $\mathbf{U}$ we get

$$\mathbf{U}^H\bar{\mathbf{P}}_\gamma\mathbf{U} = \mathbf{\Lambda} + \phi_\gamma\mathbf{I}\ , \tag{7.12}$$

which shows that the $M-R$ smallest eigenvalues of $\bar{\mathbf{P}}_\gamma$ equal ϕ_γ. Because $\mathbf{P_y}$ is estimated from limited data, the $M-R$ smallest eigenvalues will have some distribution around ϕ_γ. Therefore, we take their mean value as our estimate of ϕ_γ. That is,

$$\hat{\phi}_\gamma = \sum_{i=R+1}^{M} \frac{\lambda_{\gamma i}}{M-R}\ . \tag{7.13}$$

Note that, we assume all the eigenvalues in this work are ordered in descending order, i.e., λ_1 is the largest eigenvalue. The error of the estimator in Eq. (7.13) is analyzed for the special case with $R=1$ in [10]. Eq. (7.13) is indeed a biased estimate of ϕ_γ (underestimation) due to using the subset of the ordered eigenvalues. Although Eq. (7.13) is not the optimal estimate of the late reverberation PSD, we choose this estimator since it does not need any RTF information.

Note that this method can be seen as an extension of the method proposed in [10] where a single source scenario (i.e., $R=1$) was assumed. However, we work with estimates of $\mathbf{P_y}, \mathbf{\Gamma}$ and $\mathbf{P_v}$. Therefore, similar to other spatial coherence-based methods as evaluated in [24], this method can have overestimation errors or underestimation errors when the late reverberation PSD is relatively small compared to the noise PSD (e.g. under low reverberant signal-to-noise ratios (RSNRs) in [24]). Even when the true covariance matrix $\mathbf{P_y}$ is used, we can only obtain an estimated noise PSD, implying a residual noise PSD error will remain. Hence, we have

$$\hat{\mathbf{P}}_\gamma = \mathbf{P_y} - \hat{\phi}_v\mathbf{I} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \phi_\gamma\mathbf{\Gamma} + \underbrace{\left(\phi_v - \hat{\phi}_v\right)\mathbf{I}}_{\text{residual noise}}\ . \tag{7.14}$$

The whitened matrix is then given by

$$\begin{aligned}\hat{\bar{\mathbf{P}}}_\gamma &= \mathbf{L}^{-1}\left(\mathbf{P_y} - \hat{\phi}_v\mathbf{I}\right)\mathbf{L}^{-H} \\ &= \left(\mathbf{L}^{-1}\mathbf{A}\right)\mathbf{P}\left(\mathbf{L}^{-1}\mathbf{A}\right)^H + \phi_\gamma\mathbf{I} + \left(\phi_v - \hat{\phi}_v\right)\mathbf{\Gamma}^{-1}\ .\end{aligned} \tag{7.15}$$

If $\left(\phi_v - \hat{\phi}_v\right) \gg \phi_\gamma$, the $M-R$ smallest eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ can be much larger than ϕ_γ resulting in large overestimation errors of ϕ_γ. If $-\left(\phi_v - \hat{\phi}_v\right) \gg \phi_\gamma$, the eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ can be negative. A common way to deal with negative PSD estimates is to replace the negative estimates with ε as done in [12]. However, this will result in very large underestimation errors. To avoid large overestimation errors and underestimation errors, we propose the following estimation procedure for ϕ_v and ϕ_γ.

First of all, notice that $\mathbf{P_x} = \mathbf{P_y} - \phi_v\mathbf{I} - \phi_\gamma\mathbf{\Gamma} = \mathbf{A}\mathbf{P}\mathbf{A}^H$ is positive semi-definite with rank R. In practice, we have the estimated matrix

$$\hat{\mathbf{P}}_\mathbf{x} = \hat{\mathbf{P}}_\mathbf{y} - \hat{\phi}_v\mathbf{I} - \hat{\phi}_\gamma\mathbf{\Gamma}\ , \tag{7.16}$$

which can have negative eigenvalues even when we know the actual values of the PSDs ϕ_v and ϕ_γ, since we only have an estimated $\hat{\mathbf{P}}_{\mathbf{y}}$. Therefore, instead of adjusting $\hat{\phi}_v$ and $\hat{\phi}_\gamma$ to make $\hat{\mathbf{P}}_{\mathbf{x}}$ positive semi-definite with a rank R, we only constrain the estimated matrix $\hat{\mathbf{P}}_{\mathbf{x}}$ to have no less than R positive eigenvalues to overcome adjustments that will lead to overestimation of ϕ_γ. We now consider three cases in which this constraint is violated due to large overestimation errors of $\hat{\phi}_v$ and $\hat{\phi}_\gamma$.

1. If the given initial estimate $\hat{\phi}_v$ (estimated from speech absence frames) is larger than λ_{yR}, with λ_{yR} the R-th largest eigenvalue of $\hat{\mathbf{P}}_{\mathbf{y}}$, for any non-negative $\hat{\phi}_\gamma$, we have

$$\begin{aligned}\hat{\mathbf{P}}_{\mathbf{x}} &= \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_v \mathbf{I} - \hat{\phi}_\gamma \mathbf{\Gamma} \\ &\preceq \hat{\mathbf{P}}_{\mathbf{y}} - \lambda_{yR} \mathbf{I} - \hat{\phi}_\gamma \mathbf{\Gamma} \\ &\preceq \hat{\mathbf{P}}_{\mathbf{y}} - \lambda_{yR} \mathbf{I} \ ,\end{aligned} \tag{7.17}$$

   where the matrix inequality $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semi-definite. Since $\hat{\mathbf{P}}_{\mathbf{y}} - \lambda_{yR}\mathbf{I}$ has at most $R-1$ positive eigenvalues, $\hat{\mathbf{P}}_{\mathbf{x}}$ has less than R positive eigenvalues. Therefore, to make sure $\hat{\mathbf{P}}_{\mathbf{x}}$ has no less than R positive eigenvalues, we need $\hat{\phi}_v < \lambda_{yR}$. In this work, we update $\hat{\phi}_v$ by

$$\hat{\phi}_v \leftarrow \min\left\{\hat{\phi}_v, \frac{\sum_{i=R}^{M} \lambda_{yi}}{M-R+1}\right\} \ , \tag{7.18}$$

   such that $\hat{\phi}_v \leq \frac{\sum_{i=R}^{M} \lambda_{yi}}{M-R+1} \leq \lambda_{yR}$, where, for the second inequality, the equality holds only when $\lambda_{yR} = \lambda_{yR+1} = \cdots = \lambda_{yM}$.

2. Next, $\hat{\phi}_v$ can still be largely overestimated such that the eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$ in Eq. (7.15) are too small to get a positive $\hat{\phi}_\gamma$ using Eq. (7.13). Therefore, we iteratively update $\hat{\phi}_v$ by $\hat{\phi}_v \leftarrow c_v \hat{\phi}_v$ with $0 < c_v < 1$ a constant value such as $c_v = 0.9$ and estimate $\hat{\phi}_\gamma$ using Eq. (7.13) again until a positive $\hat{\phi}_\gamma$ is obtained. Note that this procedure has at most $\left\lceil \log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)\right\rceil + 1$ iterations since after these iterations, we have

$$\hat{\phi}_v c_v^{\left\lceil \log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)\right\rceil + 1} < \hat{\phi}_v c_v^{\log_{c_v}\left(\frac{\lambda_{yM}}{\hat{\phi}_v}\right)} = \lambda_{yM} \ , \tag{7.19}$$

   and $\hat{\bar{\mathbf{P}}}_\gamma$ will be positive definite. This results in positive eigenvalues of $\hat{\bar{\mathbf{P}}}_\gamma$, and hence, a positive $\hat{\phi}_\gamma$.

3. Finally, $\hat{\phi}_\gamma$ can be overestimated such that $\hat{\mathbf{P}}_{\mathbf{x}}$ has less than R positive eigenvalues. Therefore, we iteratively update $\hat{\phi}_\gamma$ by $\hat{\phi}_\gamma \leftarrow c_\gamma \hat{\phi}_\gamma$ with $0 < c_\gamma < 1$ a constant value such as $c_\gamma = 0.1$ until $\hat{\mathbf{P}}_{\mathbf{x}}$ has R positive eigenvalues. Since we have updated $\hat{\phi}_v$ by Eq. (7.18), we have $\hat{\phi}_v \leq \lambda_{yR}$. Hence, in the worst case that we need many iterations, $\hat{\phi}_\gamma$ approaches zero and $\hat{\mathbf{P}}_{\mathbf{x}} \approx \hat{\mathbf{P}}_{\mathbf{y}} - \hat{\phi}_v \mathbf{I}$ can have R positive eigenvalues.

The late reverberation PSD estimator is summarized in Algorithm 1.

Algorithm 6: ϕ_γ estimator

```
Input: Estimated $\mathbf{P}_\mathbf{y}$, $\mathbf{\Gamma}$,init.$\hat{\phi}_v$, $IterN$
Output: $\hat{\phi}_\gamma$,$\hat{\mathbf{P}}_\mathbf{x}$
for all $k,l$ do
    Calculate the EVD of $\mathbf{P}_\mathbf{y}$ and update $\hat{\phi}_v$ using Eq. (7.18).
    Use $\hat{\phi}_v$ and $\mathbf{\Gamma}$ to do subtraction and whitening using Eq. (7.14) and Eq. (7.15).
    Calculate the EVD of $\hat{\hat{\mathbf{P}}}_\gamma$.
    Calculate $\hat{\phi}_\gamma$ using Eq. (7.13).
    while $\hat{\phi}_\gamma < 0$ do
        Update $\hat{\phi}_v$ by $\hat{\phi}_v \leftarrow c_v \hat{\phi}_v$
        Calculate the EVD of $\hat{\hat{\mathbf{P}}}_\gamma$.
        Calculate $\hat{\phi}_\gamma$ using Eq. (7.13).
    Calculate $\hat{\mathbf{P}}_\mathbf{x}$ using Eq. (7.16).
    Calculate the $R$-th largest eigenvalue of $\hat{\mathbf{P}}_\mathbf{x}$, $\lambda_{xR}$. while $\lambda_{xR} < 0$ do
        Update $\hat{\phi}_\gamma \leftarrow c_\gamma \hat{\phi}_\gamma$. Calculate $\hat{\mathbf{P}}_\mathbf{x}$ using Eq. (7.16).
        Calculate the $R$-th largest eigenvalue of $\hat{\mathbf{P}}_\mathbf{x}$, $\lambda_{xR}$.
```

7.3.2. ESTIMATOR OF THE RTF MATRIX AND THE SOURCE PSDS

Without loss of generality, we consider the estimator of the RTF matrix and the source PSDs for the first time segment (i.e., $\beta = 1$) and neglect the index β for notational convenience. Since all time frames in a time segment are assumed to share the same RTFs, we can estimate the RTF matrix with improved accuracy using all time frames jointly, similar to the recently proposed methods in [13] and Chapter 6. Having estimated $\hat{\phi}_\gamma$ and $\hat{\phi}_v$, we can subtract both the late reverberation and noise components from $\hat{\mathbf{P}}_\mathbf{y}(t)$ for $t = 1, \cdots, N$, and we get

$$\hat{\mathbf{P}}_\mathbf{x}(t) = \hat{\mathbf{P}}_\mathbf{y}(t) - \hat{\phi}_v \mathbf{I} - \hat{\phi}_\gamma(t) \mathbf{\Gamma} = \widehat{\mathbf{A}\mathbf{P}(t)\mathbf{A}^H}, \tag{7.20}$$

for $t = 1, \cdots, N$.

PARAMETER IDENTIFIABILITY

Before estimating the RTF matrix and the source PSDs, we need to analyze the parameter identifiability to avoid biased estimates. In general, the parameters are said to be identifiable meaning that if two matrices have the form as in Eq. (7.20) (i.e., $\mathbf{P}_{\mathbf{x}1}(t) = \mathbf{A}_1 \mathbf{P}_1(t) \mathbf{A}_1{}^H$ and $\mathbf{P}_{\mathbf{x}2}(t) = \mathbf{A}_2 \mathbf{P}_2(t) \mathbf{A}_2{}^H$) for $t = 1, \cdots, N$, then $\mathbf{P}_{\mathbf{x}1} = \mathbf{P}_{\mathbf{x}2}$ is equivalent to $\mathbf{A}_1 = \mathbf{A}_2$ and $\mathbf{P}_1 = \mathbf{P}_2$. Note that for a given matrix $\mathbf{P}_\mathbf{x} = \mathbf{A}\mathbf{P}\mathbf{A}^H$, we can find different solutions by simply permuting the columns of $\mathbf{A}$ and corresponding diagonal elements of $\mathbf{P}$. However, since this permutation ambiguity can be further solved by methods such as post-processing [35], we consider the parameters to be equal to their permuted versions in this work. Note that the first row of $\mathbf{A}$ are all ones and $\mathbf{P}$ is diagonal.

We now show that for multiple sources (i.e., $R > 1$) and a time-segment consisting of one time-frame (i.e., $N = 1$), the parameters are not identifiable. That means, for any matrix $\mathbf{A}_1$ with its first row all ones and $\mathbf{P}_1(1)$ diagonal, we can find $\mathbf{A}_2 \neq \mathbf{A}_1$ and $\mathbf{P}_2 \neq \mathbf{P}_1$ while $\mathbf{A}_1\mathbf{P}_1\mathbf{A}_1{}^H = \mathbf{A}_2\mathbf{P}_2\mathbf{A}_2{}^H$, where the first row of $\mathbf{A}_2$ are all ones and $\mathbf{P}_2$ is diagonal.

For any unitary matrix $\mathbf{Q} \in \mathbb{C}^{R\times R}$, we can construct

$$\mathbf{A}_2 = \mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\left(\mathrm{diag}\left(\mathbf{e}_1^H\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\right)\right)^{-1} \tag{7.21}$$

and

$$\mathbf{P}_2 = \left(\mathrm{diag}\left(\mathbf{e}_1^H\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\right)\right)\left(\mathrm{diag}\left(\mathbf{e}_1^H\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\right)\right)^H \tag{7.22}$$

with $\mathbf{e}_1 = [1,0,\cdots,0]^T \in \mathbb{C}^{M\times 1}$ (where the subcript in $\mathbf{e}_1$ indicates that the first microphone is the reference). The diagonal matrix $\mathbf{P} = \mathrm{diag}\left(\mathbf{e}_1^H\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\right)$ is used to make the first row of $\mathbf{A}_2$ all ones and $\mathbf{P}_2$ diagonal. We then have

$$\begin{aligned}\mathbf{P}_{\mathbf{x}2} &= \mathbf{A}_2\mathbf{P}_2\mathbf{A}_2{}^H \\ &= \left(\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\mathbf{P}^{-1}\right)\left(\mathbf{P}\mathbf{P}^H\right)\left(\mathbf{A}_1\mathbf{P}_1^{\frac{1}{2}}\mathbf{Q}\mathbf{P}^{-1}\right)^H \\ &= \mathbf{A}_1\mathbf{P}_1\mathbf{A}_1{}^H \\ &= \mathbf{P}_{\mathbf{x}1}\ ,\end{aligned} \tag{7.23}$$

but $\mathbf{A}_2 \neq \mathbf{A}_1$ and $\mathbf{P}_2 \neq \mathbf{P}_1$ for the non-diagonal unitary $\mathbf{Q}$ and $R > 1$ ($\mathbf{Q}$ is a scalar when $R = 1$). Hence, if no other prior information is used, the parameters for a single time frame are not identifiable, and we need multiple time frames, i.e., $N \geq 2$, for each time segment to estimate the RTF matrix and the PSDs uniquely. Note that $N \geq 2$ is only a necessary condition for the identifiability of the parameters. For a sufficient condition, We need further assumptions on the PSDs of the sources, which we will introduce in Section 7.3.2.

SOBI

Although the SOBI method was proposed in [25] to estimate the mixing matrix and separate the source signals directly, we slightly modify this method and use it to estimate the RTF matrix and the PSDs. We therefore first introduce a modified SOBI as the reference method for RTF estimation in this subsection. Subsequently, in the next subsection, we propose a significantly improved method based on SOBI, referred to as the minimum variance joint diagonalization method (MVJD).

Given is a set of covariance matrices $\{\mathbf{P}_{\mathbf{x}}(t)\}_{t=1}^N$, with $N \geq 2$. To find $\mathbf{A}$ and $\mathbf{P}(t)$, such that $\mathbf{P}_{\mathbf{x}}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H$, for $t = 1,\cdots,N$, we can make use of a joint diagonalization of the set $\{\mathbf{P_x}(t)\}_{t=1}^{N}$. That means, instead of estimating $\mathbf{A}$ and $\mathbf{P}(t)$ directly, we first estimate $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}$ and $\tilde{\mathbf{P}}(t) = \mathbf{P}(1)^{-\frac{1}{2}}\mathbf{P}(t)\mathbf{P}(1)^{-\frac{H}{2}}$ via solving a joint diagonalization problem, as we will show later. Let us for now assume we know $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$. In that case, we can estimate the RTF matrix and $\mathbf{P}(t)$ by

$$\mathbf{A} = \tilde{\mathbf{A}}\mathrm{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^{-1} , \tag{7.24}$$

$$\mathbf{P}(1) = \mathrm{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)\mathrm{diag}\left(\mathbf{e}_1^H\tilde{\mathbf{A}}\right)^H , \tag{7.25}$$

and

$$\mathbf{P}(t) = \mathbf{P}(1)^{\frac{1}{2}}\tilde{\mathbf{P}}(t)\mathbf{P}(1)^{\frac{H}{2}} , \tag{7.26}$$

where $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$.

Now, we show how to estimate $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$. Consider estimating the SVD components of $\tilde{\mathbf{A}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^H$. We can reformulate $\mathbf{P_x}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H$ by

$$\begin{aligned}\mathbf{P_x}(t) &= \underbrace{\mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}}_{\tilde{\mathbf{A}}}\underbrace{\mathbf{P}(1)^{-\frac{1}{2}}\mathbf{P}(t)\mathbf{P}(1)^{-\frac{H}{2}}}_{\tilde{\mathbf{P}}(t)}\mathbf{P}(1)^{\frac{H}{2}}\mathbf{A}^H \\ &= \tilde{\mathbf{A}}\tilde{\mathbf{P}}(t)\tilde{\mathbf{A}}^H \\ &= \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\underbrace{\mathbf{V}^H\tilde{\mathbf{P}}(t)\mathbf{V}}_{\mathbf{P_w}(t)}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{U}^H .\end{aligned} \tag{7.27}$$

For $t = 1$, we have $\mathbf{P_x}(1) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ as then $\tilde{\mathbf{P}}(1) = \mathbf{I}$ and $\mathbf{V}^H\mathbf{V} = \mathbf{I}$. Therefore, both $\mathbf{U}$ and $\mathbf{\Sigma}$ are known from the above EVD of $\mathbf{P_x}(1)$. Then we can use $\mathbf{U}$ and $\mathbf{\Sigma}$ to calculate

$$\begin{aligned}\mathbf{P_w}(t) &= \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}^H\mathbf{P_x}(t)\mathbf{U}\mathbf{\Sigma}^{-\frac{1}{2}} \\ &= \mathbf{V}^H\tilde{\mathbf{P}}(t)\mathbf{V} .\end{aligned} \tag{7.28}$$

Since $\mathbf{V}$ is unitary and $\tilde{\mathbf{P}}(t)$ is diagonal with its r-th diagonal element $\frac{\phi_r(t)}{\phi_r(1)}$, Eq. (7.28) is indeed the EVD of $\mathbf{P_w}(t)$ with $\left\{\frac{\phi_r(t)}{\phi_r(1)}\right\}_{r=1}^{R}$ the eigenvalues and $\mathbf{V}$ the joint eigenvector matrix for all t in the segment. To make sure we get a unique estimate of $\mathbf{V}$, we need to assume that for any r_1-th and r_2-th eigenvectors (i.e. the r_1-th and r_2-th columns of $\mathbf{V}$), there exist one time frame t_0 such that the r_1-th and r_2-th eigenvalues are distinct [25], i.e.,

$$\frac{\phi_{r_1}(t_0)}{\phi_{r_1}(1)} \neq \frac{\phi_{r_2}(t_0)}{\phi_{r_2}(1)} . \tag{7.29}$$

The joint eigenvector matrix $\mathbf{V}$ diagonalizes $\{\mathbf{P_w}(t)\}_{t=1}^{N}$ simultaneously, i.e.,

$$\mathbf{V}\mathbf{P_w}(t)\mathbf{V}^H = \tilde{\mathbf{P}}(t), \forall t \in \{1, \cdots, N\} . \tag{7.30}$$

However, such a joint diagonalization might not be achieved in practice since we only have the estimated $\mathbf{P_w}(t)$. Therefore, an approximate joint diagonalization was pursued in [25] by minimizing the off-diagonal elements of $\mathbf{V}\mathbf{P}_{\mathbf{w}}(t)\mathbf{V}^H$, which is

$$\begin{aligned}&\min_{\mathbf{V}} \sum\nolimits_{t=2}^{N} \mathrm{off}\left(\mathbf{V}\mathbf{P}_{\mathbf{w}}(t)\mathbf{V}^H\right)\\ &\text{s.t. } \mathbf{V}^H\mathbf{V}=\mathbf{I}\ ,\end{aligned} \tag{7.31}$$

where $\mathrm{off}(\mathbf{C}) = \sum_{1\leq i\neq j\leq M} \left|C_{i,j}\right|^2$ for a matrix $\mathbf{C} \in \mathcal{C}^{M\times M}$. Then, the algorithm proposed in [36] is used to solve Eq. (7.31), which is numerically very efficient. With $\mathbf{V}$ estimated, we use $\mathrm{diag}\left(\mathbf{V}\mathbf{P}_{\mathbf{w}}(t)\mathbf{V}^H\right)$ as the estimate of $\tilde{\mathbf{P}}(t)$.

The SOBI method is summarized in Algorithm 2.

```
Algorithm 7: SOBI
  Input: Estimated $\hat{\mathbf{P}}_{\mathbf{x}}(t)$, for $t = 1,\cdots,N$,
  Output: $\mathbf{A}$ and $\mathbf{P}(t)$ for $t = 1,\cdots,N$,
  Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\hat{\mathbf{P}}_{\mathbf{x}}(1)$.
  Construct new matrices $\mathbf{P}_{\mathbf{w}}(t)$ for $t = 2,\cdots,N$ using Eq. (7.28).
  Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(t)$ for $t = 2,\cdots,N$ using the Jacobi-like algorithm [36].
  Estimate $\tilde{\mathbf{A}}$ with $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$.
  Estimate $\mathbf{A}$ and $\mathbf{P}(t)$ using Eqs. (7.24) to (7.26).
```

Note that, with this SOBI-based algorithm, the matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ in the SVD of $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{P}(1)^{\frac{1}{2}}$ are first estimated before estimating the RTF matrix and the PSDs. The estimation accuracy of $\mathbf{U}$ and $\mathbf{\Sigma}$ depends fully on the estimation accuracy of the first covariance matrix $\mathbf{P}_{\mathbf{x}}(1) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$, which can be hugely erroneous. For instance, when the late reverberation and noise have large energy during the first time frame. Instead of using $\mathbf{P}_{\mathbf{x}}(1)$ to do the EVD at the first step, we can use any proper linear combination of all the covariance matrices $\sum_{t=1}^{N} c_t\mathbf{P}_{\mathbf{x}}(t)$ with $c_t \geq 0$, such as the average of a subset of the covariance matrices as we proposed in Chapter 6. The estimation accuracy of the RTF matrix and the PSDs can be improved by using values for c_t that minimize the error between $\sum_{t=1}^{N} c_t\mathbf{P}_{\mathbf{x}}(t)$ and its estimated counterpart $\sum_{t=1}^{N} c_t\hat{\mathbf{P}}_{\mathbf{x}}(t)$.

MVJD

In this subsection, we first show our generalization of the SOBI method. Then we propose our minimum variance joint diagonalization method (MVJD) based on the analysis of the variance of the sample covariance matrices.

Instead of using the first covariance matrix $\mathbf{P}_{\mathbf{x}}(1)$ to do the EVD at the first step of SOBI, we can use any proper linear combination of all the covariance matrices $\sum_{t=1}^{N} c_t\mathbf{P}_{\mathbf{x}}(t) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H$ with $c_t \geq 0$. Therefore, $\mathbf{U}$ and $\mathbf{\Sigma}$ can be obtained from $\sum_{t=1}^{N} c_t\mathbf{P}_{\mathbf{x}}(t)$ for

$$\tilde{\mathbf{A}} = \mathbf{A}\left(\sum_{t=1}^{N} c_t\mathbf{P}(t)\right)^{\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{V}^H\ . \tag{7.32}$$

Then, using Eqs. (7.28) and (7.31), we can get $\mathbf{V}$ and $\tilde{\mathbf{P}}(t)$ with $\tilde{\mathbf{P}}(t) = \left(\sum_{t=1}^{N} c_t \mathbf{P}(t)\right)^{-\frac{1}{2}} \mathbf{P}(t) \left(\sum_{t=1}^{N} c_t \mathbf{P}(t)\right)^{-\frac{H}{2}}$. To get a unique estimate of $\mathbf{V}$, we assume that for any r_1-th and r_2-th columns of $\mathbf{V}$, there exists one time frame t_0 such that

$$\frac{\phi_{r_1}(t_0)}{\sum_{t=1}^{N} c_t \phi_{r_1}(t)} \neq \frac{\phi_{r_2}(t_0)}{\sum_{t=1}^{N} c_t \phi_{r_2}(t)} . \tag{7.33}$$

With $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ estimated, we can estimate $\tilde{\mathbf{A}}$ using Eq. (7.32), which further gives us the RTF matrix $\mathbf{A} = \tilde{\mathbf{A}}\mathrm{diag}\left(\mathbf{e}_1^H \tilde{\mathbf{A}}\right)^{-1}$ and $\sum_{t=1}^{N} c_t \mathbf{P}(t) = \mathrm{diag}\left(\mathbf{e}_1^H \tilde{\mathbf{A}}\right) \mathrm{diag}\left(\mathbf{e}_1^H \tilde{\mathbf{A}}\right)^H$. Finally, from $\sum_{t=1}^{N} c_t \mathbf{P}(t)$ and $\tilde{\mathbf{P}}(t)$, we can calculate the PSDs matrix for all time frames by $\mathbf{P}(t) = \tilde{\mathbf{P}}(t) \sum_{t=1}^{N} c_t \mathbf{P}(t)$.

Since the estimation errors for the estimated covariance matrices $\mathbf{P}_{\mathbf{x}}(t)$ are different for different t, using different coefficients c_t in step 1 will result in a different $\mathbf{U}$ and $\mathbf{\Sigma}$, and thus in different estimates of the RTF matrix and the PSDs.

We will now explain how we can optimally select the coefficients c_t such that the summation of the variances of the error matrix in the estimated $\mathbf{P}_{\mathbf{x}}(t)$ is minimized. Suppose we have the true PSDs of the late reverberation and noise. The estimated covariance matrix for $\mathbf{P}_{\mathbf{x}}(t)$ is then given by

$$\begin{aligned}
\hat{\mathbf{P}}_{\mathbf{x}}(t) =& \hat{\mathbf{P}}_{\mathbf{y}}(t) - \phi_v \mathbf{I} - \phi_\gamma(t) \mathbf{\Gamma} \\
=& \sum_{l=1+(t-1)T}^{tT} \frac{\mathbf{y}(l)\mathbf{y}(l)^H}{T} - \phi_v \mathbf{I} - \phi_\gamma(t) \mathbf{\Gamma} \\
=& \sum_{l} \frac{(\mathbf{x}(l) + \mathbf{n}(l))(\mathbf{x}(l)^H + \mathbf{n}(l)^H)}{T} \\
& - \phi_v \mathbf{I} - \phi_\gamma(t) \mathbf{\Gamma} \\
=& \sum_{l} \frac{\mathbf{x}(l)\mathbf{x}(l)^H}{T} \\
& + \sum_{l} \frac{\mathbf{x}(l)\mathbf{n}(l)^H + \mathbf{n}(l)\mathbf{x}(l)^H}{T} \\
& + \sum_{l} \frac{\mathbf{n}(l)\mathbf{n}(l)^H}{T} - \phi_\gamma(t) \mathbf{\Gamma} - \phi_v \mathbf{I} ,
\end{aligned} \tag{7.34}$$

where $\mathbf{n}(l) = \mathbf{d}(l) + \mathbf{v}(l)$. Since we assumed that $\mathbf{x}$, $\mathbf{l}$ and $\mathbf{v}$ are uncorrelated, we omit the cross correlation terms in Eq. (7.34) and get

$$\begin{aligned}
\hat{\mathbf{P}}_{\mathbf{x}}(t) \approx& \sum_{l} \frac{\mathbf{x}(l)\mathbf{x}(l)^H}{T} \\
& + \sum_{l} \frac{\mathbf{n}(l)\mathbf{n}(l)^H}{T} - \phi_\gamma(t) \mathbf{\Gamma} - \phi_v \mathbf{I} .
\end{aligned} \tag{7.35}$$

Applying this to the weighted sum of the estimated covariance matrices used at the first step of our proposed method, we get

$$\begin{aligned}\sum_{t=1}^{N} c_t \hat{\mathbf{P}}_{\mathbf{x}}(t) \approx \sum_{t=1}^{N} c_t \frac{\sum\limits_{l=1+(t-1)T}^{tT} \mathbf{x}(l)\,\mathbf{x}(l)^H}{T} \\ \underbrace{+\sum_{t=1}^{N} c_t \frac{\sum\limits_{l=1+(t-1)T}^{tT} \mathbf{n}(l)\,\mathbf{n}(l)^H}{T} - \sum_{t=1}^{N} c_t \left(\phi_\gamma(t)\,\boldsymbol{\Gamma} + \phi_v \mathbf{I}\right)}_{\mathbf{W}} ,\end{aligned} \tag{7.36}$$

where the first term is a weighted sum of the sample covariance matrices for the target sources and the remaining terms are unwanted errors that we will denote by matrix $\mathbf{W}$. Since $\mathbf{n}(l)$, for $l = 1 + (t-1)T, \cdots, tT$, is assumed to follow a circularly-symmetric complex Gaussian distribution with zero mean and covariance matrix $\mathbf{P}_{\mathbf{n}}(t) = \phi_\gamma(t)\,\boldsymbol{\Gamma} + \phi_v \mathbf{I}$, the random matrix $\mathbf{W}_t = \sum\limits_{l=1+(t-1)T}^{tT} \mathbf{n}(l)\,\mathbf{n}(l)^H$ has a complex Wishart distribution $\sim \mathcal{W}_M^C(T, \mathbf{P}_{\mathbf{n}}(t))$ with T degrees of freedom [37]. The expectation of $\mathbf{W}_t$ is $T\mathbf{P}_{\mathbf{n}}(t)$ [38]. Hence the expectation of $\mathbf{W}$ is

$$\begin{aligned}\mathrm{E}\{\mathbf{W}\} &= \sum_{t=1}^{N} c_t \frac{\mathrm{E}\{\mathbf{W}_t\}}{T} - \sum_{t=1}^{N} c_t \mathbf{P}_{\mathbf{n}}(t) \\ &= \sum_{t=1}^{N} c_t \frac{T\mathbf{P}_{\mathbf{n}}(t)}{T} - \sum_{t=1}^{N} c_t \mathbf{P}_{\mathbf{n}}(t) \\ &= 0 .\end{aligned} \tag{7.37}$$

The variance of the $\{i, j\}$-th element of $\mathbf{W}_t$ is $\mathrm{var}\left\{W_{t,i,j}\right\} = P_{n,i,i}P_{n,j,j}$ [38]. Hence the summation of the variances of all the elements of $\mathbf{W}_t$ is

$$\begin{aligned}\sum_{i,j=1}^{M} \mathrm{var}\left\{W_{i,j}\right\} &= \sum_{i,j=1}^{M} \mathrm{var}\left\{\sum_{t=1}^{N} c_t \frac{W_{t,i,j}}{T}\right\} \\ &= \sum_{i,j=1}^{M} \sum_{t=1}^{N} \frac{c_t^2}{T^2} \mathrm{var}\left\{W_{t,i,j}\right\} \\ &= \sum_{i,j=1}^{M} \sum_{t=1}^{N} \frac{c_t^2}{T^2} P_{n,i,i}P_{n,j,j} \\ &= \sum_{t=1}^{N} \frac{c_t^2}{T^2} \left[\mathrm{tr}\left(\mathbf{P}_{\mathbf{n}}(t)\right)\right]^2 \\ &\geq \frac{1}{T^2} \left[\sum_{t=1}^{N} c_t \mathrm{tr}\left(\mathbf{P}_{\mathbf{n}}(t)\right)\right]^2 ,\end{aligned} \tag{7.38}$$

where the equality holds when $c_1 \mathrm{tr}\left(\mathbf{P_n}(1)\right) = c_2 \mathrm{tr}\left(\mathbf{P_n}(2)\right) = \cdots = c_N \mathrm{tr}\left(\mathbf{P_n}(N)\right)$. Since $\mathrm{tr}\left(\mathbf{P_n}(t)\right) = M\left(\phi_\gamma(t) + \phi_v\right)$, we can choose $c_t = \frac{1}{\phi_\gamma(t)+\phi_v}$ to minimize the variances of the error matrix.

The MVJD method is summarized in Algorithm 3.

Algorithm 8: MVJD

```
Input: Estimated $\hat{\mathbf{P}}_\mathbf{x}(t)$ $\phi_\gamma(t)$ and $\phi_v$,for $t = 1,\cdots,N$,
Output: $\mathbf{A}$ and $\mathbf{P}(t)$ for $t = 1,\cdots,N$,
Estimate $\mathbf{U}$ and $\mathbf{\Sigma}$ from EVD of $\sum_{t=1}^{N} \frac{1}{\phi_\gamma(t)+\phi_v} \hat{\mathbf{P}}_\mathbf{x}(t)$.
Construct new matrices $\mathbf{P_w}(t)$ for $t = 1,\cdots,N$ using Eq. (7.28).
Estimate $\mathbf{V}$ and $\tilde{\mathbf{P}}(t)$ for $t = 1,\cdots,N$ using the Jacobi-like algorithm [36].
Estimate $\tilde{\mathbf{A}}$ with $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ using Eq. (7.32).
Estimate $\mathbf{A}$ and $\mathbf{P}(t)$ using $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{P}}(t)$.
```

7.4. EXPERIMENTS

In this section, we evaluate the estimation performance of our proposed method in various simulated acoustic scenarios using multiple microphones. We compare our method to both the SOBI based method introduced in Section 7.3.2 and the SCFA method [13] that we will introduce in Section 7.4.1. In Section 7.4.2, we evaluate the different methods using performance measures for the estimation accuracy, the predicted speech quality and the predicted speech intelligibility. Finally, the performances including the computational complexity of all methods are presented and discussed in Sections 7.4.3 and 7.4.4.

7.4.1. REFERENCE METHODS

In addition to the SOBI method introduced in Section 7.3.2, we include another state-of-the-art method for comparison, which is the simultaneous confirmatory factor analysis (SCFA) method [13]. The SCFA method is based on the maximum likelihood cost function:

$$\min \sum_{t=1}^{N} \log\left|\mathbf{P_y}(t)\right| + \mathrm{tr}\left(\hat{\mathbf{P}}_\mathbf{y}(t)\,\mathbf{P}_\mathbf{y}^{-1}(t)\right) . \tag{7.39}$$

Specifically, the following non-convex optimization problem is formalized in [13]

$$\begin{aligned}
\underset{\substack{\mathbf{P}(t),\mathbf{A}\\ \phi_\gamma(t),\phi_v}}{\arg\min}\ & \sum_{t=1}^{N} \log\left|\mathbf{P}_\mathbf{y}(t)\right| + \mathrm{tr}\left(\hat{\mathbf{P}}_\mathbf{y}(t)\,\mathbf{P}_\mathbf{y}^{-1}(t)\right)\\
\text{s.t. } & \mathbf{P}_\mathbf{y}(t) = \mathbf{A}\mathbf{P}(t)\mathbf{A}^H + \phi_\gamma(t)\boldsymbol{\Gamma} + \phi_v\mathbf{I},\\
& \mathbf{P}(t) = \mathrm{diag}\left[\phi_1(t),\cdots,\phi_R(t)\right],\\
& a_{1r} = 1, \phi_r(t) \geq 0, \phi_\gamma(t) \geq 0, \phi_v \geq 0,\\
& \text{for } t = 1,\cdots,N; r = 1,\cdots,R\ .
\end{aligned} \tag{7.40}$$

Note that the signal model assumed here is the same as our proposed model in Eq. (7.8). According to [13], a local minimum for Eq. (7.40) can be found by iteratively reducing the cost function value. At each iteration, a non-linear constrained optimization problem needs to be solved to update the parameters. The number of required iterations is very large (e.g. in the order of 500) due to the non-convexity of the problem and the high dimension of the parameters. Therefore, the SCFA method has a relatively high computational cost.

7.4.2. EVALUATION MEASURES

ESTIMATION ACCURACY

Since the main goal of this work is to find accurate estimates of the parameters of interest, we first introduce the estimation accuracy measures for the different parameters.

For the RTF matrix, to evaluate the alignment of the estimated RTF with the ground-truth RTF, we calculate the Hermitian angle by means of

$$E_\mathbf{a} = \frac{\sum\limits_{\beta=1}^{B}\sum\limits_{k=1}^{K/2+1}\sum\limits_{r=1}^{R} \arccos\left(\frac{\left|\mathbf{a}_r(\beta,k)^H\hat{\mathbf{a}}_r(\beta,k)\right|}{\|\mathbf{a}_r(\beta,k)\|_2\|\hat{\mathbf{a}}_r(\beta,k)\|_2}\right)}{BR(K/2+1)}\ , \tag{7.41}$$

where the error has been averaged over different sources, frequency bins and the number of time segments B. For the PSD of the r-th source ϕ_r and the PSD of the late reverberation ϕ_γ, we use the symmetric log-error distortion measure [39]

$$E_i = \frac{10\sum\limits_{t,k\in\mathcal{Q}}\left|\log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right|}{|\mathcal{Q}|}\ , \tag{7.42}$$

for $i = r$ or γ, where the index set $\mathcal{Q}$ is used to discard zero PSDs, as used in [40] and $|\mathcal{Q}|$ is the cardinality of $\mathcal{Q}$. For the errors of the source PSDs, we use E_s to denote the average value of them, i.e., $E_s = \frac{\sum_{r=1}^{R} E_r}{R}$. Note that the error in Eq. (7.42) can be seen as the summation of the overestimation error and the underestimation error, which are

$$E_i^{\mathrm{ov}} = \frac{10\sum\limits_{t,k\in\mathcal{Q}}\left|\min\left\{0,\log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right\}\right|}{|\mathcal{Q}|}\ , \tag{7.43}$$

and

$$E_i^{\mathrm{un}} = \frac{10 \sum\limits_{t,k\in\mathcal{Q}} \max\left\{0, \log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right\}}{|\mathcal{Q}|} \ , \tag{7.44}$$

respectively. In a noise reduction method, under or overestimates of target source PSDs or noise/interference PSDs have each its own effect. When the target source PSDs have large underestimation errors or the noise or interference PSD has large overestimation errors, the target source obtained by a noise reduction algorithm using these estimates typically has large distortions. On the other hand, if the estimate of the noise or interference PSD has a large underestimation error, the reconstructed signal often comes with musical noise [41]. Therefore, we will also present in detail the underestimation errors and overestimation errors in the experiments.

PREDICTED QUALITY AND INTELLIGIBILITY

Since the estimated parameters are commonly used in noise reduction algorithms, we use the estimates in the well-known multi-channel Wiener filter (MWF) [42] and use the MWF outputs to reconstruct each point source signal. For estimating the r-th signal, the MWF can be expressed as a combination of a minimum variance distortionless response (MVDR) beamformer [43] and a single-channel Wiener filter, which is

$$\hat{\mathbf{w}}_r = \frac{\hat{\phi}_r}{\hat{\phi}_r + \hat{\mathbf{w}}_{r,\mathrm{MVDR}}^H \hat{\mathbf{R}}_{r,nn} \hat{\mathbf{w}}_{r,\mathrm{MVDR}}} \hat{\mathbf{w}}_{r,\mathrm{MVDR}} \ , \tag{7.45}$$

where $\mathbf{w}_{r,\mathrm{MVDR}}$ is the MVDR beamformer

$$\hat{\mathbf{w}}_{r,\mathrm{MVDR}} = \frac{\hat{\mathbf{R}}_{r,nn}^{-1} \hat{\mathbf{a}}_r}{\hat{\mathbf{a}}_r^H \hat{\mathbf{R}}_{r,nn}^{-1} \hat{\mathbf{a}}_r} \ , \tag{7.46}$$

and

$$\hat{\mathbf{R}}_{r,nn} = \sum_{i=1, i\neq r}^{R} \hat{\phi}_i \hat{\mathbf{a}}_i \hat{\mathbf{a}}_i^H + \hat{\phi}_\gamma \hat{\boldsymbol{\Gamma}} + \hat{\phi}_v \mathbf{I} \ . \tag{7.47}$$

Note that the permutation ambiguity exists after estimating the RTF matrix and the sources PSDs (i.e., we cannot determine which column of $\mathbf{A}$ belongs to which source for different frequency bins). This problem is beyond the scope of this work and methods on this topic, to name a few, were investigated in [35], [44], [45]. In the experiments of this work, we use the oracle RTF matrix as guidance to permute the columns of the estimated RTF matrix per time-frequency tile.

The predicted speech quality of each reconstructed signal is evaluated by calculating the segmental-signal-to-noise-ratio (SSNR) [46] and the perceptual evaluation of speech quality (PESQ) measure [47]. The predicted speech intelligibility performance is evaluated by the speech intelligibility in bits (SIIB) measure [48], [49]. Alternately, we select one of the R sources as the target and the remaining $R-1$ sources as interferers. We than average all measures we used in the experiments over these R different setups.

7.4.3. EXPERIMENTS WITH SIMULATED RIRS

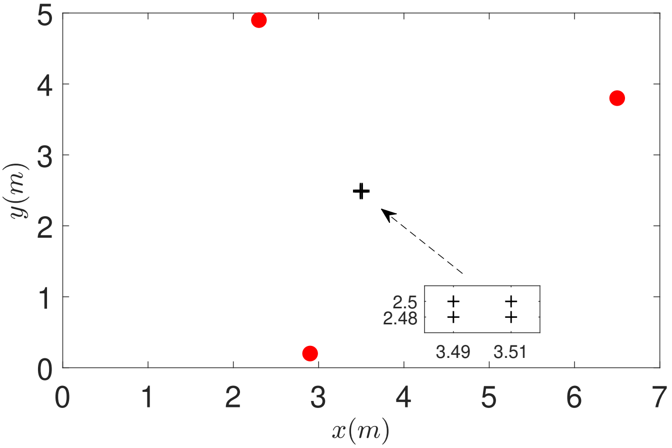


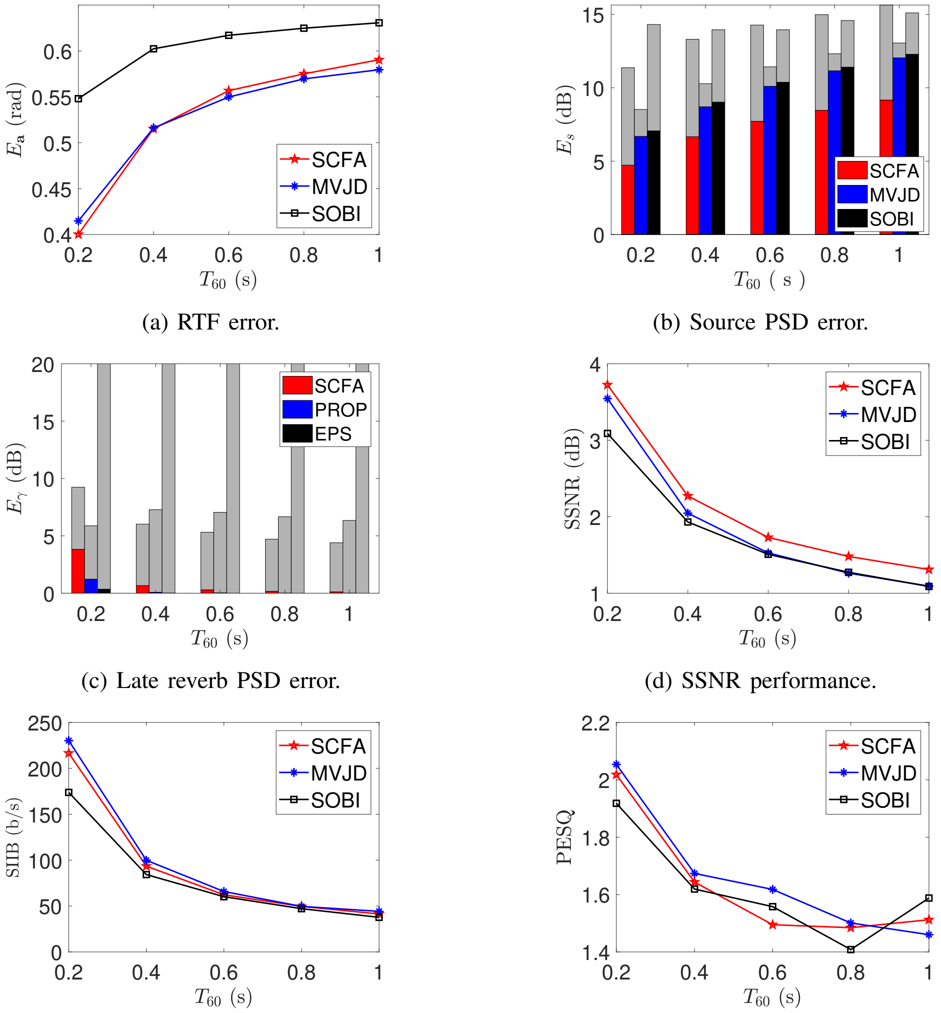
Figure 7.3: Top view of the acoustic scene with a zoom-in of microphones.

The acoustic scene of the first experiment is shown in Fig. 7.3, where four microphones and three sources are placed in the room with a dimension of $7 \times 5 \times 4$m. The speech signals are downloaded from the TIMIT database [50]. To simulate the reverberant signal recorded by each microphone, we convolve the speech signals (with a duration of 33 s) with the room impulse responses (RIRs) generated by the image source method [51]. The microphones are omnidirectional and the RIRs have a duration of 1 s. Then, to synthesize the noisy microphone signals, we add independently generated white Gaussian noise to each reverberant signal. The variance of the noise is fixed at a value calculated from given signal-to-noise ratios (SNRs). The SNR value is the ratio between the overall energy of the direct and early reflections of the first speech signal at position $(3.49, 2.5)$ and the energy of the noise component at the first microphone.

The microphone signals are sampled at a frequency of $f_s = 16$ kHz after which they are transformed to the frequency domain by the STFT procedure, in which the 50 % overlapping square-root Hann window with a length of 32 ms and the FFT length of 512 are used. Note that the window length is the same as the sub-time frame length and also equals the early part of the RIRs. Each time frame has $T = 20$ overlapping sub-time frames and thus a duration of 0.32 s. Note that this duration can be longer than the actual speech source stationary period and the PSDs can be seen as the averages of the PSDs over each time frame.

PERFORMANCE COMPARISON

In Fig. 7.4, we present the performance comparison among our proposed method and the two reference methods, where we adjust the reverberation time from 0.2 s to 1 s. The number of time frames per segment is 8 and the SNR is fixed at 30 dB. We first



(a) RTF error.

(b) Source PSD error.

(c) Late reverb PSD error.

(d) SSNR performance.

(e) SIIB performance.

(f) PESQ performance.

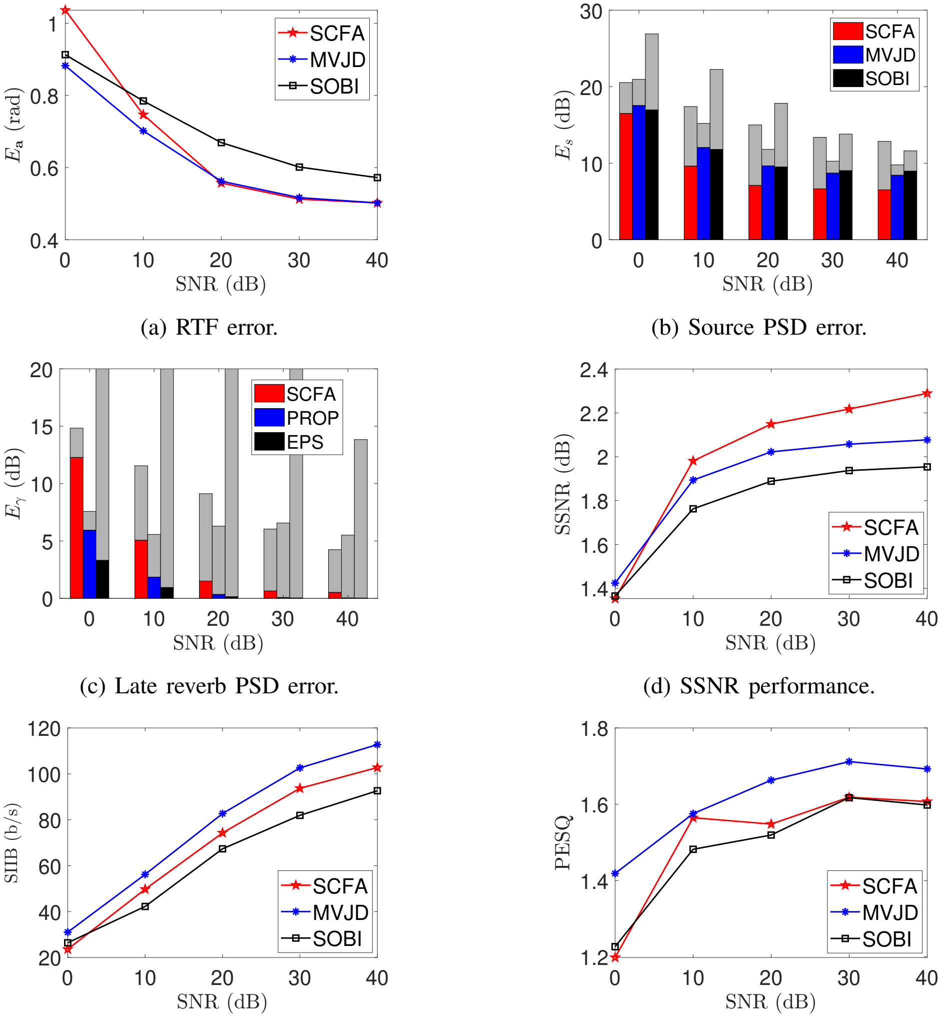
Figure 7.4: Performance vs the late reverberation time. In Figs. b and c, the top gray bars indicate the underestimation errors, and the bottom colored bars indicate overestimation errors.

show the RTF estimation error calculated by Eq. (7.41) in Fig. 7.4a. The error for each method increases as the room becomes more reverberant. Our proposed MVJD method has similar performance compared to SCFA, which both outperform the SOBI method. In Figs. 7.4b and 7.4c, we show the PSDs estimation error calculated by Eq. (7.42). For each bar (each overall error), we also show the overestimation error using the bottom colored bar and the underestimation error using the top gray bar. In Fig. 7.4b, we show the source PSD estimation errors, where the errors also become larger when the reverberation time increases. Our proposed method has the smallest error compared to SOBI and a slightly larger overestimation error compared to SCFA. In particular, the underestimation error (gray bar) of our proposed method outperforms the other two methods. In Fig. 7.4c, the late reverberation PSD errors are presented. For visibility, parts of the bars over 20 dB are not shown. Note that we use our proposed late reverberation estimator in both SOBI and our proposed MVJD. The 'EPS' method in Fig. 7.4c refers to replacing the negative estimates from Eq. (7.12) with ε, the machine precision, as used in [12]. Our proposed estimator has similar errors compared to SCFA, both of which are much smaller than EPS. Note that the overestimation errors of our method are smaller than SCFA. In Figs. 7.4d to 7.4f, it is shown that our proposed method and SCFA outperform SOBI in general regarding to the predicted speech quality and speech intelligibility evaluated by SSNR, PESQ and SIIB. Note that our proposed method has better predicted intelligibility and predicted quality in terms of PESQ but a worse SSNR than SCFA.

In Fig. 7.5, we compare all the methods while changing the noise level by increasing the SNR from 0 dB to 40 dB. The number of time frames per segment is again eight and the reverberation time is fixed at 0.4 s. The RTF estimation error is first shown in Fig. 7.5a. For all the methods, the RTF error is relatively small for high SNR. Our proposed MVJD method has the best performance, which outperforms SCFA at low SNR values and outperforms SOBI at high SNR values. In Fig. 7.5b, the source PSD estimation errors also reduces when the SNR increases. Our proposed method has the smallest underestimation errors, while the SCFA method has the smallest overestimation errors. In Fig. 7.5c, the late reverberation PSD errors are compared, where our proposed late reverberation estimator and SCFA have much smaller errors compared to using the ε procedure. For visibility, parts of the bars over 20 dB are again not shown. Note that the overestimation errors of our method are smaller than SCFA, both which decreases when the SNR increases. In Figs. 7.5d to 7.5f, it is also shown that our proposed method and SCFA outperform SOBI in general regarding to the predicted speech quality and speech intelligibility.

Fig. 7.6 shows the performance comparison for different time segment durations (i.e., different numbers of time frames per segment). For visibility, parts of the bars over 20 dB in Fig. 7.6c are not shown. Our proposed method and the SCFA method still outperform the SOBI method in estimation errors, speech quality and speech intelligibility.

To evaluate the impact of the three robustification steps we proposed for the late reverberation estimator, we compared the estimation errors using different steps, where we fixed reverberation time at 0.4 s, SNR at 30 dB and the number of time frames at 8.



(a) RTF error.

(b) Source PSD error.

(c) Late reverb PSD error.

(d) SSNR performance.

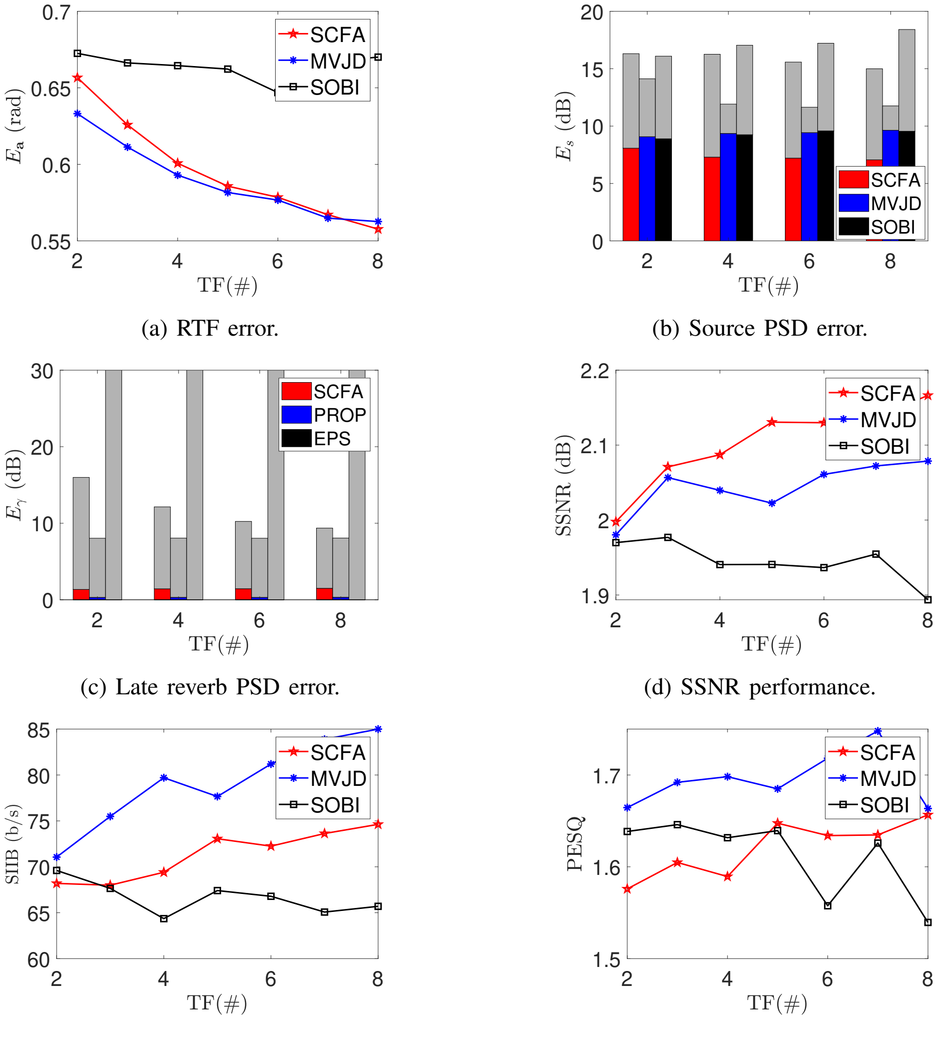
(e) SIIB performance.

(f) PESQ performance.

Figure 7.5: Performance vs SNR. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

Table 7.1: Estimation errors using different steps.

	E_a	E_s	E_γ	E_s^{un}	E_s^{ov}	E_γ^{un}	E_γ^{ov}
Step 1	0.52	10.19	43.50	1.31	8.89	43.48	0.02
Step 1+2	0.52	10.62	7.01	1.99	8.62	6.87	0.14
Step 1+2+3	0.52	10.32	7.40	1.60	8.72	7.33	0.07



(a) RTF error.

(b) Source PSD error.

(c) Late reverb PSD error.

(d) SSNR performance.

(e) SIIB performance.

(f) PESQ performance.

Figure 7.6: Performance vs the number of time frames per segment. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

We can see from the table that by adding step 2 after step 1, the late reverberation PSD error is reduced a lot. Adding step 3 after step 2 shows slight reduction on the error of the source PSDs. The reason is that without step 3, the covariance matrices for the sources might not have R positive eigenvalues, which will likely lead to negative source PSD estimates that will be replaced by small positive value like eps, resulting in a huge underestimation error of the source PSDs for that frequency bin. However, the overall improvement is not big as step 3 is only executed for some time-frequency bins. The average iteration number of the frequency bins executing step 3 in this experiment is 0.05 (with the total number of frequency bins 257).

7.4.4. EXPERIMENTS WITH RECORDED RIRS

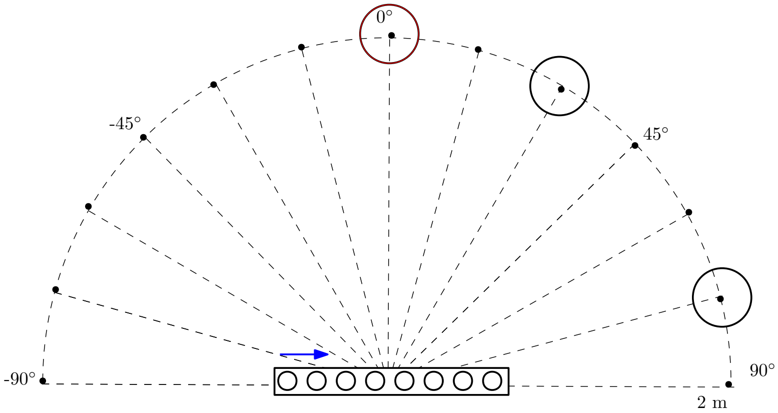
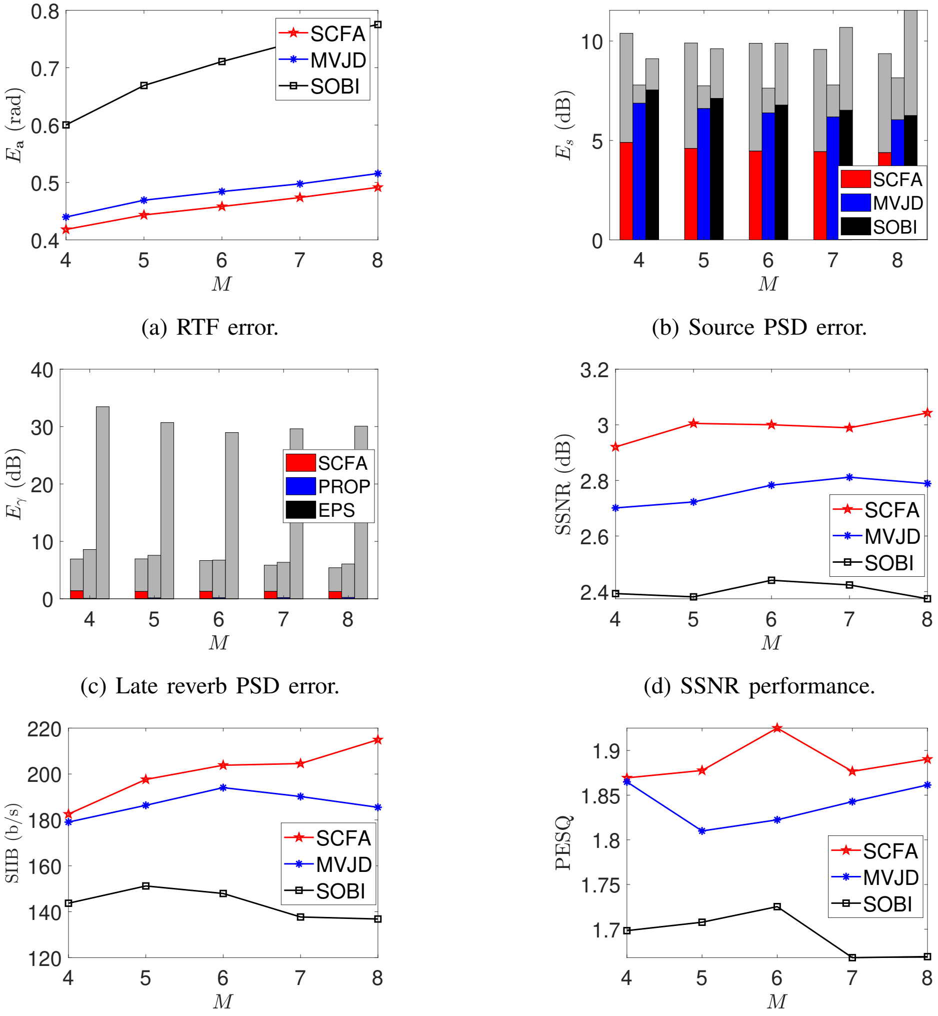


Figure 7.7: Geometric setup of the acoustic scene [52] with big red circles representing the positions of sources. From left to right, as shown by the blue arrow, the first M microphones are used with M changing from 4 to 8.

SETUP

In this section, we use RIRs recorded in a real room with dimension $6 \times 6 \times 2.4$ m [52]. We consider two scenarios in this experiment. For the first scenario with three sources, the geometric positions of the sources and the microphones are shown in Fig. 7.7. The microphones form a uniform linear array with 8 cm interdistance. The data base in [52] contains RIRs measured at a 2 m distance from the microphone array center at different angles. We convolve the RIRs at 0°, 30° and 75° with different speech signals. We also add white Gaussian noise to simulate the microphone self-noise. For the second scenario with two sources, where one source is fixed at an angle of 15° and the other source is placed at different angles ranging from 0° to 90° in steps of 15°. We use the same STFT procedure as we used in the first experiment to transfer the time domain signals to the frequency domain.



(a) RTF error.

(b) Source PSD error.

(c) Late reverb PSD error.

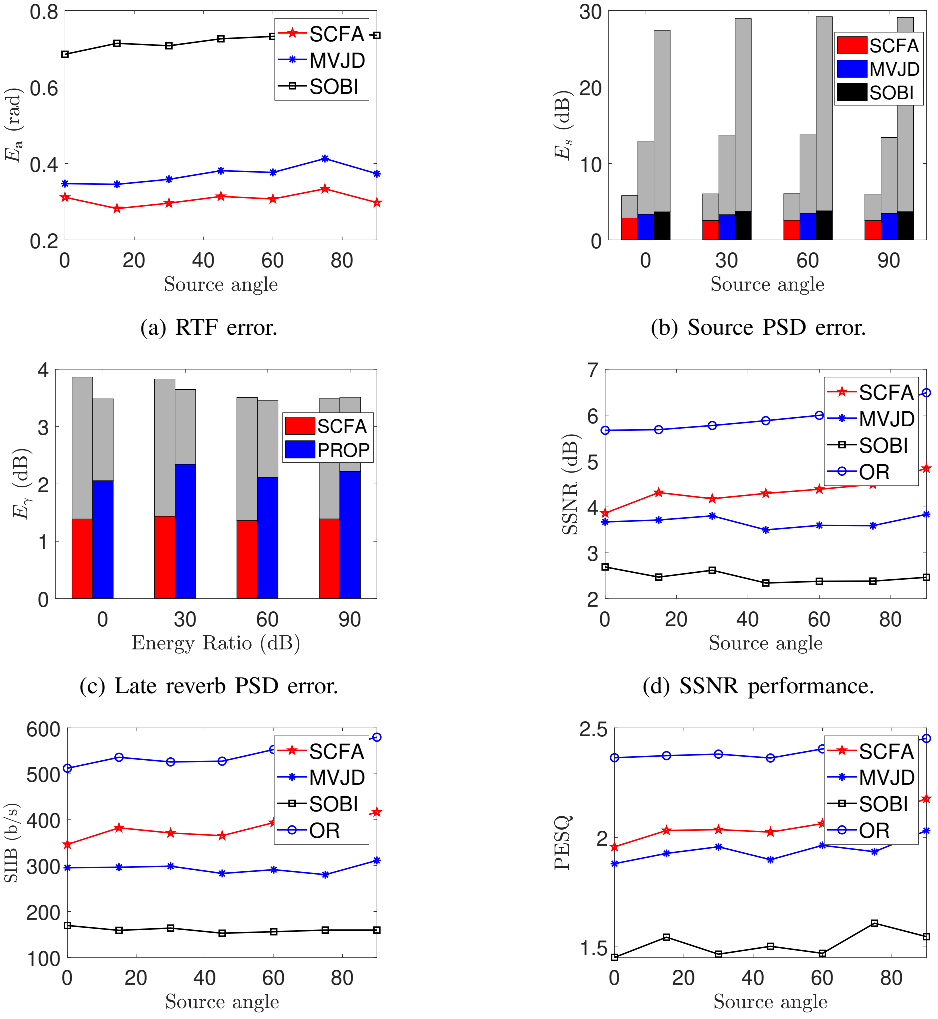
(d) SSNR performance.

(e) SIIB performance.

(f) PESQ performance.

Figure 7.8: Performance vs the number of microphones. In Figs. b and c, the top gray bars indicate the underestimation errors, the bottom colored bars indicate overestimation errors.

In Fig. 7.8, we show the performance comparison for three sources for all methods as a function of the number of microphones. The SNR is 30 dB and the reverberation time is 0.36 s. Note that when using a larger number of microphones, the theoretical spatial coherence matrix calculated by Eq. (7.7) can be close to singular, particularly for low frequency bins as observed in our experiments. To avoid numerical issues, we regularize such matrices by $\mathbf{\Gamma} = \mathbf{\Gamma} + \mu\mathbf{I}$ with $\mu = 10^{-3}$ in this experiment. In terms of RTF errors in Fig. 7.8a, MVJD and SCFA show similar performance and both outperform SOBI. In terms of the source PSD errors in Fig. 7.8b, SCFA has a lower underestimation error than MVJD, but MVJD has lower overestimation errors than SCFA. In terms of the predicted speech quality and intelligibility performance as shown in Figs. 7.8d to 7.8f, our proposed method has performances close to the SCFA method, while both outperform the SOBI method.



(a) RTF error.

(b) Source PSD error.

(c) Late reverb PSD error.

(d) SSNR performance.

(e) SIIB performance.

(f) PESQ performance.

Figure 7.9: Performance vs source position. In Figs. b and c, the top gray bars indicate the underestimation errors, and the bottom colored bars indicate overestimation errors.

In Fig. 7.9, we show the performance comparison for two sources for all methods for different positions of the second source ranging from 0° to 90° by every 15°. It is shown that the performances, on both estimation errors and predicted speech quality and intelligibility, do not change much with different positions of the second source. MVJD shows comparable performance with SCFA, which both greatly outperforms SOBI. We also show the predicted speech quality and intelligibility performance when using oracle parameters to calculate the MWF, which is referred to as 'OR' in Fig. 7.9.

In previous experiments, the maximum number R of sources per time segment and frequency is assumed known. In practice, it needs to be estimated using methods such as [26], [27]. The estimated $\hat{R}$ can be smaller, equal or larger than R. To evaluate this problem, we show in Table 7.2 the predicted speech quality and intelligibility performance of our proposed method for $\hat{R}-R$ being -1, 0 and 1. We considered two sources placed at 0° and 60°. It is shown that our proposed method with overestimated $\hat{R}=R+1$ is similar to the case of $\hat{R}=R$. However, the performance with underestimated $\hat{R}=R-1$ is much worse than the other cases.

Table 7.2: Predicted speech quality and intelligibility comparison.

$\hat{R}-R$	-1	0	1
SSNR	0.92	3.93	3.80
SIIB	163.92	316.60	318.26
PESQ	1.00	2.04	2.01

Finally, we show the computation time using MATLAB for processing the microphone signals with a duration of 33 s using different methods and different number of microphones in Table 7.3. We can see that the SCFA method needs the longest run time, which increases as the number of microphones increases. Our proposed method and the SOBI method have a similar run time, which is in the order of 700 times faster than SCFA. For our proposed method, The computation time of the late reverberation PSD estimator mainly comes from iterative steps with EVD. The average iteration number for each frequency bin is less than 1 as observed in our experiments. In practice, it depends on the accuracy of given noise PSDs. The computation cost of the RTF matrix and source PSDs estimator mainly comes from the joint diagonalization algorithm, which is in the order of R^3. Based on the analysis, if the overall iteration number for the late reverberation PSD estimator is I, the computational complexity of the proposed algorithm is in the order of IM^3+R^3.

Table 7.3: Computation time (in s) comparison.

M	4	5	6	7	8
SCFA	3523	3708	4362	5169	5768
MVJD	9	10	11	8	8
SOBI	8	11	10	9	9

7.4.5. EXPERIMENTS WITH REAL RECORDINGS

SETUP

In this section, we used signals recorded by four microphones mounted on a dummy head in the BRUDEX Database [53], i.e., including natural reverberation. We considered two sources, speaker 1 at 0 ° and speaker 2 at 60 ° with medium reverberant condition in [53]. The sampling frequency is 48000 Hz in this experiment and the FFT length is 2048. The other settings for the STFT procedure are the same as the previous two experiments. Note that besides the real recordings, the RIRs were also measured in [53], with which we can simulate source components such as the late reverberation. For the spatial coherence matrix of the late reverberation, we calculate it using the simulated late reverberation component by

$$\mathbf{\Gamma}_{i,j}(k) = \frac{\sum\limits_{l} d_i(l,k)\, d_j(l,k)^*}{\sqrt{\sum\limits_{l} |d_i(l,k)|^2}\sqrt{\sum\limits_{l} |d_j(l,k)|^2}} \, . \tag{7.48}$$

For the noise component, we assume a spatially white (spectrally non-white) model and use the first second recordings (speech absent duration) to measure the noise PSD for each frequency bin. Note that in this experiment, we added another reference method, ARMA-FastMNMF [54] as a comparison to a state-of-the-art speech enhancement method. For ARMA-FastMNMF, we used the following parameters: number of speech: 2, speech model: NMF, number of noise: 0, tap length of the MA model $L_{MA} = 8$, tap length of the AR model $L_{AR} = 4$, delay of the late reverberation $\Delta = 1$ and the Iterative Source Steering (ISS) algorithm was used. Note that all methods were run in a device with Intel(R) Core(TM) i7-10610U CPU @ 1.80GHz 2.30 GHz without using GPU. Notice that ARMA-FastMNMF does not estimate the underlying parametric model (as the proposed method and SCFA), but directly performs the source separation.

PERFORMANCE COMPARISON

In Fig. 7.10, we evaluate the predicted speech quality and intelligibility performance of all methods. As shown in the figures, our proposed method outperforms SOBI in all measures and outperforms ARMA-FastMNMF in PESQ and SIIB. We also show the computation time normalized by the time it takes for MVJD in Table 7.4. We can see that although SCFA has the best performance in this experiment, its computation time is again very high compared to MVJD. Also, MVJD is about 150 times faster than the ARMA-FastMNMF method.

Table 7.4: Computation time comparison.

Methods	SCFA	MVJD	SOBI	ARMA-FastMNMF
Normalized run time	843.56	1	0.89	154.41

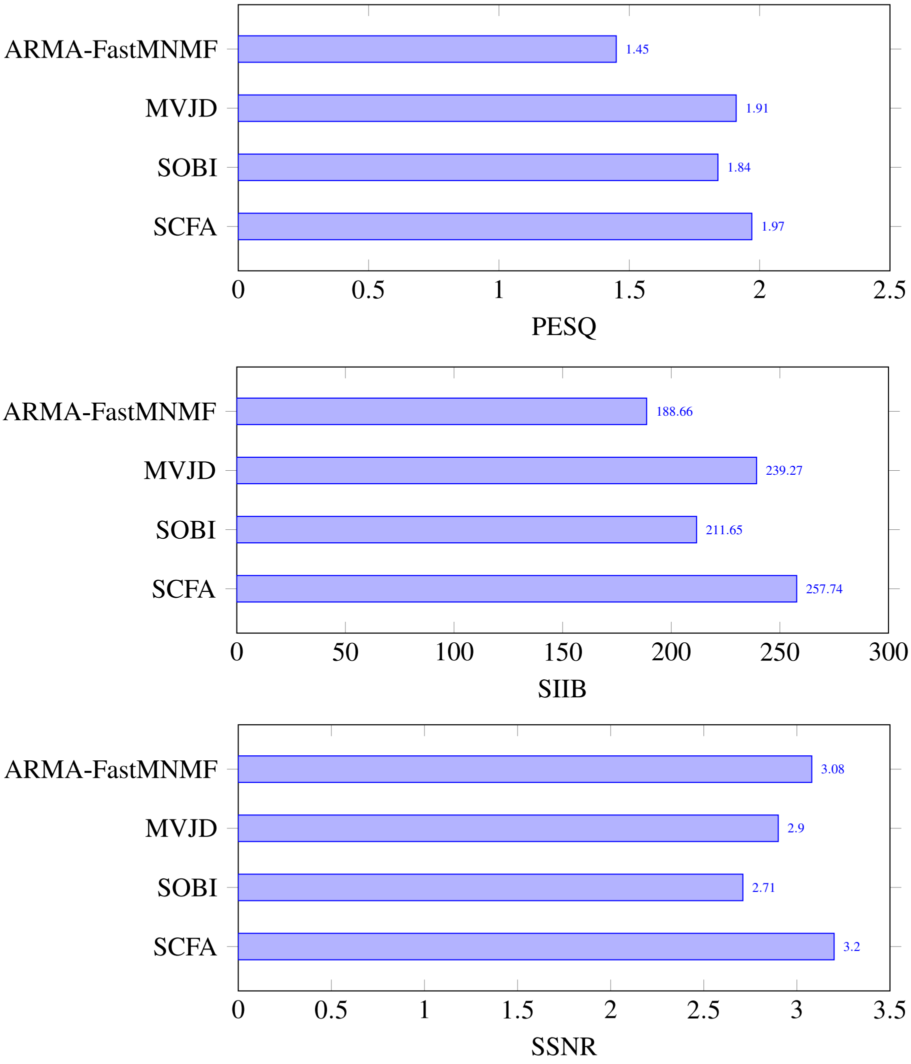


Figure 7.10: Predicted speech quality and intelligibility performance comparison.

7.5. CONCLUDING REMARKS

In this chapter we considered the complex scenario where multiple sources, late reverberation and noise exist concurrently. For this scenario, we proposed a joint estimator of the parameters include the RTFs of the sources and the PSDs of the sources and the late reverberation. We first proposed a late reverberation PSD estimator that does not require the knowledge of the RTFs. Then we proposed the minimum variance joint diagonalization (MVJD) method to estimate the RTFs and the PSDs of the sources. The proposed MVJD method is more robust than the existing joint-diagonalization SOBI method, since we considered an optimal linear combination of a set of covariance matrices instead of only the first one as done with SOBI. The optimality is obtained by minimizing the variances of the error matrix of the linearly combined sample covariance matrices. Experiments demonstrated that our proposed method outperforms the SOBI method in terms of estimation errors, the predicted speech quality and the speech intelligibility. The results also show that our proposed method achieves similar performance compared to the state-of-the-art SCFA method but has a significantly lower computational complexity.

REFERENCES

[1] K. L. Payton, R. M. Uchanski, and L. D. Braida, “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing”, *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1581–1592, Mar. 1994.

[2] J. Xia, B. Xu, S. Pentony, J. Xu, and J. Swaminathan, “Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners”, *J. Acoust. Soc. Amer.*, vol. 143, no. 3, pp. 1523–1533, Mar. 2018.

[3] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[4] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement*. Springer Nature, 2022.

[5] E. Vincent, R. Gribonval, and M. D. Plumbley, “Oracle estimators for the benchmarking of source separation algorithms”, *Signal Process.*, vol. 87, no. 8, pp. 1933–1950, 2007.

[6] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources”, *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, 2000.

[7] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[8] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[9] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[10] I. Kodrasi and S. Doclo, “Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[11] I. Kodrasi and S. Doclo, “Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[12] M. Tammen, S. Doclo, and I. Kodrasi, "Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[13] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[14] J. Zhang, R. Heusdens, and R. C. Hendriks, "Relative acoustic transfer function estimation in wireless acoustic sensor networks", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1507–1519, 2019.

[15] Y. Laufer and S. Gannot, "Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[16] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square Root-Based Multi-Source Early PSD Estimation and Recursive RETF Update in Reverberant Environments by Means of the Orthogonal Procrustes Problem", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 755–769, 2020.

[17] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[18] C. Li, J. Martinez, and R. C. Hendriks, "Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 695–705, 2023.

[19] C. Li and R. C. Hendriks, "Alternating least-squares-based microphone array parameter estimation for a single-source reverberant and noisy acoustic scenario", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3922–3934, 2023.

[20] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms", *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[21] D. Cherkassky and S. Gannot, "Successive Relative Transfer Function Identification Using Blind Oblique Projection", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 474–486, 2020.

[22] H. Gode and S. Doclo, "Covariance Blocking and Whitening Method for Successive Relative Transfer Function Vector Estimation in Multi-Speaker Scenarios", in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, 2023, pp. 1–5.

[23] Y. Laufer, B. Laufer-Goldshtein, and S. Gannot, "ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in Rank-Deficient Noise Field", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 619–634, 2020.

[24] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1056–1071, 2018.

[25] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics", *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.

[26] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis", *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6458–6473, 2018.

[27] H. Sun, P. Samarasinghe, and T. Abhayapala, "Blind source counting and separation with relative harmonic coefficients", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[28] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech", *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, 2001.

[29] C. Li and R. C. Hendriks, "Adaptive time segmentation for improved signal model parameter estimation for a single-source scenario", in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, 2023, pp. 1106–1111.

[30] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator", in *Proc. EURASIP Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[31] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields", *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

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

[33] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[34] B. F. Cron and C. H. Sherman, "Spatial-correlation functions for various noise models", *J. Acoust. Soc. Amer.*, vol. 34, no. 11, pp. 1732–1736, 1962.

[35] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping Separated Frequency Components by Estimating Propagation Model Parameters in Frequency-Domain Blind Source Separation", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1592–1604, 2007.

[36] J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization", *SIAM J. Mat. Anal. Appl.*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[37] N. R. Goodman, "Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction)", *Ann. Math. Stat.*, vol. 34, no. 1, pp. 152–177, 1963, ISSN: 00034851.

[38] D. Maiwald and D. Kraus, “On moments of complex Wishart and complex inverse Wishart distributed matrices”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 5, 1997, 3817–3820 vol.5.

[39] R. C. Hendriks, J. Jensen, and R. Heusdens, “DFT domain subspace based noise tracking for speech enhancement”, in *Proc. Interspeech*, 2007, pp. 830–833.

[40] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[41] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay”, *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[42] H. L. V. Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., Mar. 2002.

[43] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

[44] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Frequency-domain blind source separation of many speech signals using near-field and far-field models”, *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[45] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, “Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207, 2009.

[46] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[47] I.-T. Recommendation, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, *Rec. ITU-T P. 862*, 2001.

[48] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An instrumental intelligibility metric based on information theory”, *IEEE Signal Process. Lett.*, vol. 25, no. 1, pp. 115–119, 2017.

[49] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, “An evaluation of intrusive instrumental intelligibility metrics”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2153–2166, 2018.

[50] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[51] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics”, *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[52] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments”, in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sep. 2014.

[53] D. Fejgin, W. Middelberg, and S. Doclo, “BRUDEX Database: Binaural Room Impulse Responses with Uniformly Distributed External Microphones”, in *Speech Commun.; 15th ITG Conference*, 2023, pp. 126–130.

[54] K. Sekiguchi, Y. Bando, A. A. Nugraha, M. Fontaine, K. Yoshii, and T. Kawahara, “Autoregressive moving average jointly-diagonalizable spatial covariance analysis for joint source separation and dereverberation”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2368–2382, 2022.

8

ADAPTIVE TIME SEGMENTATION FOR IMPROVED SIGNAL MODEL PARAMETER ESTIMATION FOR A SINGLE-SOURCE SCENARIO

This chapter is based on the article published as: C. Li and R. C. Hendriks. “ADaptive Time Segmentation for Improved Signal Model Parameter Estimation for a Single-Source Scenario”. In: Proc. IEEE Asilomar Conf. Signals, Syst., Comput. 2023, pp. 1106–1111.

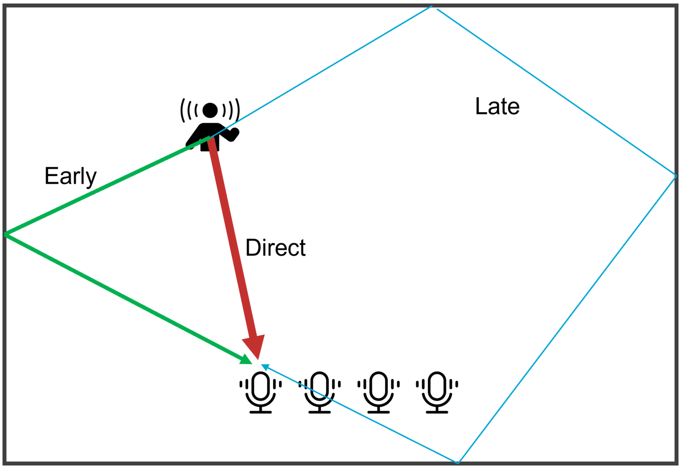


Figure 8.1: Illustration of a single source, reverberant scenario.

In this chapter, we will answer the research question 3 presented in Fig. 1.4 by proposing an adaptive time segmentation method. This method can be combined with estimators such as JMLE from Chapter 3 or JALS from Chapter 5. We will consider the combination with JMLE under the single source reverberant scenario, as shown in Fig. 8.1, although the combination with other estimators can be developed similarly.

Estimating the parameters that describe the acoustic scene is very important for many microphone array applications. For example, consider the power spectral densities (PSDs) or relative acoustic transfer functions (RTFs) that are required when estimating a particular sound source using multi-microphone noise reduction. State-of-the-art algorithms estimate the parameters per segment, where each segment consists of a fixed number of time frames. These algorithms exploit the assumption that PSDs are constant per time frame, and RTFs are constant per segment. However, in practice, sound sources will move relative to the microphone array. Improved performance is therefore expected when the actual time frames that are used to form the segments are adapted such that time frames all share the same (unknown) RTF. In this chapter, we therefore present an algorithm to obtain an optimal adaptive time segmentation and combine this with our previously published joint maximum likelihood estimator (JMLE) for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment.

8.1. INTRODUCTION

In hand-free speech communication applications such as hearing aids and mobile phones, microphone arrays are commonly used to enhance the quality and intelligibility of the target signal as the microphone signals are typically corrupted by late reverberation and ambient noise. Typically, this is done using spatial filtering techniques. However, these techniques depend on acoustic scene-related parameters such as the relative transfer function (RTF) of the target signal and the power spectral densities (PSDs) of the target signal, the late reverberation and the ambient noise, which are typically unknown in practice. Therefore, it is essential to estimate these parameters.

Speech signals are non-stationary in nature, but can be assumed stationary for a very short duration of about 10$\sim$30 ms. This results in the fact that the PSD of each acoustic component is constant for only a short duration. However, the RTF can be assumed constant as long as the sound source does not move relative to the microphone array. Typically, the duration that the source is static (defined here as a time segment) is longer than the duration that the speech source is stationary (defined here as a time frame). Hence, each time segment might contain multiple time frames that share the same RTF.

Recently, several estimation methods have been proposed to estimate the RTF and the PSDs using multiple time frames [1], [2] instead of a single time frame [3]–[10]. The methods using multiple time frames always outperform the methods using a single time frame as long as the sound source is indeed static during the time segment. However, if the source or array changes position or the room acoustics change, the methods using time segments during which the RTF is time-varying have worse estimation performance than when the time segment would be selected such that the underlying RTF is time-invariant. Therefore, in this chapter, we present an algorithm to obtain an adaptive time segmentation and combine this with our previously published joint maximum likelihood estimator (JMLE) [2] for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment. Notice that the use of an adaptive time segmentation in the speech enhancement context has been proposed before, e.g., [11], for improved estimation of the PSDs used in single-microphone noise reduction algorithm. In the current work, we present a different segmentation algorithm for the multi-microphone context based on the inner product of a sequence of initial RTF estimates. In combination with the recently proposed JMLE algorithm, this leads to improved estimates of the RTF and PSDs.

8.2. PRELIMINARIES

We consider a single acoustic point source observed by a microphone array in a reverberant environment. The source changes to new positions at unknown moments, which means it is spatially fixed for unknown time durations. The time duration that the source does not move will be referred to as a time segment indexed by β. The β-th time segment consists of one or multiple time frames from t_β to $\left(t_\beta + T_\beta - 1\right)$, where T_β is the number of time frames for the β-th time segment. Within a time frame, the speech source is assumed to be stationary. The time frame will be indexed by t. Each time frame t contains multiple overlapping sub-time frames. See Fig. 8.2 for a visual interpretation. We use the short-time Fourier transform (STFT) to transfer the signal

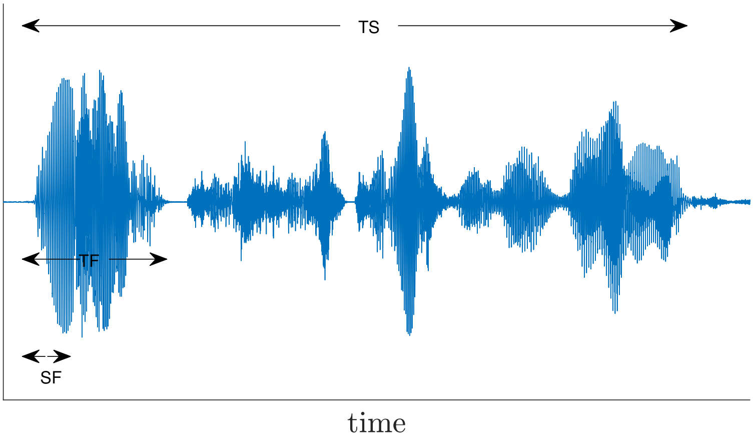


Figure 8.2: Visualisation of the definition of time segment (TS), time frames (TF) and sub frames (SF).

received at the m-th microphone into the frequency domain, leading to

$$y_m(l,k) = x_m(l,k) + r_m(l,k), \tag{8.1}$$

with l the sub-time frame index, k the frequency bin index, and m the microphone index. In Eq. (8.1), x_m denotes the direct and early reflections of the source. Variable r_m denotes the late reverberation, which is the sum of all the late reflections of the source. Using the relative transfer function (RTF) between the microphones, we can model x_m as

$$x_m(l,k) = a_m(l,k)\, s(l,k), \tag{8.2}$$

with $s(l,k)$ the direct component and early reflections at the reference microphone (the 1^{st} microphone in this work) and $a_m(l,k)$ the RTF of the source from the reference microphone to the m-th microphone. In vector form, all M microphone signals in the STFT domain can be expressed as

$$\mathbf{y}(l,k) = \mathbf{a}(l,k)\, s(l,k) + \mathbf{r}(l,k) \in \mathbb{C}^{M\times 1}. \tag{8.3}$$

Assuming that the early reflections and the late reverberation that fall in one sub-time frame are uncorrelated and zero-mean, we can write the covariance matrix of $\mathbf{y}(l,k)$ as

$$\mathbf{P_y}(l,k) \triangleq \mathbf{P_x}(l,k) + \mathbf{P_r}(l,k), \tag{8.4}$$

where $\mathbf{P_y} \triangleq E\left\{\mathbf{y}\mathbf{y}^H\right\}$ with $E\{\cdot\}$ the expectation operator. Matrices $\mathbf{P_x}$ and $\mathbf{P_r}$ are defined in the same way as $\mathbf{P_y}$. For $\mathbf{P_x}(l,k)$, we have

$$\mathbf{P_x}(l,k) = \phi_s(l,k)\, \mathbf{a}(l,k)\, \mathbf{a}^H(l,k), \tag{8.5}$$

where $\phi_s(l,k) \triangleq E\left\{|s(l,k)|^2\right\}$ is the PSD of the source at the reference microphone. For the late reverberation, we assume a spatially homogeneous sound field model

$$\mathbf{P_r}(l,k) = \phi_\gamma(l,k)\boldsymbol{\Gamma}(k), \tag{8.6}$$

where $\phi_\gamma(l,k)$ is the unknown time-varying PSD of the late reverberation and $\boldsymbol{\Gamma}(k)$ is the known time-invariant spatial coherence matrix, which can be calculated using the microphone array geometry [12].

8.2.1. PROBLEM FORMULATION

By using Eqs. (8.5) and (8.6), we formulate the noisy covariance matrix as

$$\mathbf{P_y}(t,k) = \phi_s(t,k)\mathbf{a}(\beta,k)\mathbf{a}^H(\beta,k) + \phi_\gamma(t,k)\boldsymbol{\Gamma}(k), \tag{8.7}$$

where we assumed the microphone signals are stationary over a time frame t consisting of the L_s sub-time frames indexed by $l = 1+(t-1)L_s$ till $l = tL_s$ and the RTF stays constant over a time segment β consisting of the time frames indexed by $t = t_\beta$ till $t = t_\beta + T_\beta - 1$. Based on the stationarity assumption, we can estimate $\mathbf{P_y}(t,k)$ using the sample covariance matrix $\hat{\mathbf{P}}_\mathbf{y}(t,k) = 1/L_s \sum_{l=1+(t-1)L_s}^{tL_s} \mathbf{y}(l,k)\mathbf{y}^H(l_s,k)$. Assuming that the RTF is constant for all time frames in a time segment (i.e., $t \in [t_\beta, t_\beta + T_\beta - 1]$), we can use the set $\{\mathbf{P_y}(t,k)\}_{t_\beta}^{t_\beta+T_\beta-1}$ jointly to estimate $\mathbf{a}(\beta,k)$.

The aim of this work is to estimate the time segment indices $\{t_\beta, T_\beta\}$ using the fact that the true but unknown RTFs are the same in the time frames from a single segment. Note that $t_1 = 1$ and the last time frame of the $(\beta-1)$-th time segment should be followed by the first time frame of the β-th time segment (i.e.,$t_\beta = t_{\beta-1} + T_{\beta-1}$). Since we determine the time segments sequentially, we know $\{t_{\beta-1}, T_{\beta-1}\}$ when determining the β-th time segment. Therefore, t_β is known as well and we only need to estimate T_β.

8.3. JMLE

We first present the algorithm for joint MLE of the parameters $\mathbf{a}$ and $\{\phi_s(t), \phi_\gamma(t)\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ for a given time segment $[t_\beta, t_\beta + T_\beta - 1]$. Note that this is based on our work recently published in [2]. In the next section, we will then propose the algorithm to determine $\{T_\beta\}$. Note that only in this section, we omit the frequency indexes for the simplicity of notation.

With the assumption that all the time frames are independent and the STFT coefficients are complex Gaussian distributed, we can write the negative log-likelihood function of the STFT coefficients (up to a constant and scale) as

$$L = -\sum_{t=t_\beta}^{t_\beta+T_\beta-1} \left[\log|\mathbf{P_y}(t)| + \mathrm{tr}\left(\hat{\mathbf{P}}_\mathbf{y}(t)\mathbf{P}_\mathbf{y}^{-1}(t)\right)\right]. \tag{8.8}$$

Then, using the reparameterization $\tilde{\mathbf{a}} = \frac{\mathbf{L}^{-1}\mathbf{a}}{\sqrt{\mathbf{a}^H\boldsymbol{\Gamma}^{-1}\mathbf{a}}}$ and $\tilde{\phi}_s(t) = \phi_s(t)\,\mathbf{a}^H\boldsymbol{\Gamma}^{-1}\mathbf{a}$, we reformulate the covariance matrix in Eq. (8.7) as

$$\mathbf{P}_{\mathbf{y}}(t) = \mathbf{L}\left(\tilde{\phi}_s(t)\,\tilde{\mathbf{a}}\tilde{\mathbf{a}}^H + \phi_\gamma(t)\,\mathbf{I}\right)\mathbf{L}^H, \tag{8.9}$$

where $\mathbf{L}$ is the Cholesky factor of $\boldsymbol{\Gamma}$ (i.e. $\boldsymbol{\Gamma} = \mathbf{L}\mathbf{L}^H$). The MLE cost function then becomes [2]

$$\begin{aligned}\underset{\tilde{\phi}_s(t),\tilde{\mathbf{a}},\phi_\gamma(t)}{\arg\min} \sum_{t=t_\beta}^{t_\beta+T_\beta-1} & \log\left[\left(\tilde{\phi}_s(t) + \phi_\gamma(t)\right)\left(\phi_\gamma(t)^{M-1}\right)\right] \\ & + \operatorname{tr}\left(\phi_\gamma(t)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) \\ & - \frac{\phi_\gamma(t)^{-2}\tilde{\phi}_s(t)}{1+\phi_\gamma(t)^{-1}\tilde{\phi}_s(t)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}},\end{aligned} \tag{8.10}$$

where $\hat{\mathbf{P}}_{\mathbf{w}} = \mathbf{L}^{-1}\hat{\mathbf{P}}_{\mathbf{y}}\mathbf{L}^{-H}$.

To solve the above optimization, we first find initial estimates of the parameters by considering each time frame independently (as explained in Section 8.3.1). This initialisation step does thus not require a segmentation algorithm as it works on the individual time frames. After the initialisation, alternating estimation between $\mathbf{a}$ and $\left\{\phi_s(t), \phi_\gamma(t)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ is performed (see Section 8.3.2), which thus can benefit from a correct segmentation.

8.3.1. INITIALISATION

When considering a single time frame, the cost function reduces to

$$\begin{aligned}\underset{\tilde{\phi}_s(t),\tilde{\mathbf{a}},\phi_\gamma(t)}{\arg\min}\; & \log\left[\left(\tilde{\phi}_s(t) + \phi_\gamma(t)\right)\left(\phi_\gamma(t)^{M-1}\right)\right] \\ & + \operatorname{tr}\left(\phi_\gamma(t)^{-1}\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) \\ & - \frac{\phi_\gamma(t)^{-2}\tilde{\phi}_s(t)}{1+\phi_\gamma(t)^{-1}\tilde{\phi}_s(t)}\tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}},\end{aligned} \tag{8.11}$$

where only the last term depends on $\tilde{\mathbf{a}}$ with a negative coefficient. Hence, the MLE-optimal $\tilde{\mathbf{a}}$ based on a single time frame is the solution to

$$\underset{\tilde{\mathbf{a}}}{\arg\max}\, \tilde{\mathbf{a}}^H\hat{\mathbf{P}}_{\mathbf{w}}(t)\,\tilde{\mathbf{a}}, \tag{8.12}$$

which is the principal eigenvector of $\hat{\mathbf{P}}_{\mathbf{w}}(t)$. Note that the initialization step does not use the prior information that all time frames in a time segment share the same RTF. Let T denote the maximum size of a segment. For the T time frames that could potentially form the β-th time segment, we will have T different estimates of $\tilde{\mathbf{a}}$ at this step. These are denoted by $\left\{\hat{\tilde{\mathbf{a}}}(t)\right\}_{t=t_\beta}^{t_\beta+T-1}$. These estimates will be used for the time segmentation algorithm to find the actual length T_β of the β-th time segment in Section 8.4.

With the estimated RTF $\hat{\tilde{\mathbf{a}}}(t)$, we can find the optimal estimates of $\tilde{\phi}_s(t)$ and $\tilde{\phi}_\gamma(t)$ by substituting $\hat{\tilde{\mathbf{a}}}(t)$ into Eq. (8.11) [2], that is,

$$\hat{\tilde{\phi}}_s(t) = \frac{M\lambda_{\max}(t) - \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right)}{M-1}, \tag{8.13}$$

$$\hat{\phi}_\gamma(t) = \frac{\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) - \lambda_{\max}(t)}{M-1}, \tag{8.14}$$

where $\lambda_{\max}(t)$ is the principal eigenvalue of $\hat{\mathbf{P}}_{\mathbf{w}}(t)$.

8.3.2. ALTERNATING ESTIMATION

The initial estimates of the PSDs $\left\{\hat{\tilde{\phi}}_s(t), \hat{\phi}_\gamma(t)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ can be substituted into the cost function in Eq. (8.10) to estimate the RTF using all time frames in the time segment jointly

$$\arg\max_{\tilde{\mathbf{a}}} \sum_{t=t_\beta}^{t_\beta+T_\beta-1} \left(\frac{\hat{\tilde{\phi}}_s(t)}{\hat{\phi}_\gamma(t) + \hat{\tilde{\phi}}_s(t)} \frac{1}{\hat{\phi}_\gamma(t)} \tilde{\mathbf{a}}^H \hat{\mathbf{P}}_{\mathbf{w}}(t)\, \tilde{\mathbf{a}} \right), \tag{8.15}$$

which is the principal eigenvector of

$$\sum_{t=t_\beta}^{t_\beta+T_\beta-1} \frac{\hat{\tilde{\phi}}_s(t)}{\hat{\phi}_\gamma(t) + \hat{\tilde{\phi}}_s(t)} \frac{1}{\hat{\phi}_\gamma(t)} \hat{\mathbf{P}}_{\mathbf{w}}(t). \tag{8.16}$$

Then, with the estimated RTF $\hat{\mathbf{a}}$, we can estimate the PSDs using [2]

$$\hat{\tilde{\phi}}_s(t) = \frac{M\hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{w}}(t)\, \hat{\tilde{\mathbf{a}}} - \operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right)}{M-1} \tag{8.17}$$

and

$$\hat{\phi}_\gamma(t) = \frac{\operatorname{tr}\left(\hat{\mathbf{P}}_{\mathbf{w}}(t)\right) - \hat{\tilde{\mathbf{a}}}^H \hat{\mathbf{P}}_{\mathbf{w}}(t)\, \hat{\tilde{\mathbf{a}}}}{M-1}. \tag{8.18}$$

Note that $\hat{\phi}_\gamma(t)$ is positive but $\hat{\tilde{\phi}}_s(t)$ can be negative [2], while $\tilde{\phi}_s(t)$ is positive. We therefore replace the negative estimates $\hat{\tilde{\phi}}_s(t)$ with the initial estimates from Eq. (8.13).

We alternatingly estimate $\tilde{\mathbf{a}}$ and $\left\{\tilde{\phi}_s(t), \phi_\gamma(t)\right\}_{t=t_\beta}^{t_\beta+T_\beta-1}$ untile a certain number of iterations are executed. Finally, the RTF vector and the PSD of the target source have to be compensated for the reparameterization, and are given by $\hat{\mathbf{a}} = \frac{\mathbf{L}\hat{\tilde{\mathbf{a}}}}{\mathbf{L}\hat{\tilde{\mathbf{a}}}\mathbf{e}_1}$ and $\hat{\phi}_s = \frac{\hat{\tilde{\phi}}_s}{\hat{\mathbf{a}}^H \mathbf{\Gamma}^{-1} \hat{\mathbf{a}}}$, where $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$.

8.4. TIME SEGMENTATION

In this section, we present the proposed algorithm for an adaptive time segmentation, where the number of time frames in a segment depends on how time-varying the RTF is. Due to the latency consideration, we consider that a maximum of T time frames can form a time segment. Hence, the maximum length of a segment is T. The minimum size is a single time frame. After the initialization step in the JMLE (Section 8.3.1), we have estimated reparameterized RTF vectors $\left\{\hat{\bar{\mathbf{a}}}(t,k)\right\}_{t=t_\beta}^{t_\beta+T-1}$ for the T time frames that could potentially be part of segment β, which will start at time frame t_β (i.e., the first time frame after the previous time segment $\beta-1$). By analyzing the distance between these roughly estimated RTF vectors, we can find all time frames that should fall into the β-th time segment.

As the RTF estimation error is often expressed using the Hermitian angle, it would be natural to use this as a metric to determine the distance between two RTF vectors $\left\{\hat{\bar{\mathbf{a}}}(i,k),\hat{\bar{\mathbf{a}}}(j,k)\right\}$. We consider the i-th TF and j-th TF to belong to the same time segment if the Hermitian angle satisfies

$$\arccos\left(\frac{\left|\hat{\bar{\mathbf{a}}}(i,k)^H\hat{\bar{\mathbf{a}}}(j,k)\right|}{\left\|\hat{\bar{\mathbf{a}}}(i,k)\right\|_2\left\|\hat{\bar{\mathbf{a}}}(j,k)\right\|_2}\right)<c_h, \tag{8.19}$$

where c_h is a given constant threshold.

Alternatively, we could also construct a $M\times 2$ matrix $\mathbf{A}(i,j,k)=\left[\hat{\bar{\mathbf{a}}}(i,k),\hat{\bar{\mathbf{a}}}(j,k)\right]$ and analyze its second largest singular value $\sigma_2(i,j,k)$. In the ideal case that $\hat{\bar{\mathbf{a}}}(i,k)$ and $\hat{\bar{\mathbf{a}}}(j,k)$ are estimates of the same RTF vector without any errors, $\mathbf{A}(i,j,k)$ has rank 1 and $\sigma_2(i,j,k)=0$. Hence, we consider the i-th TF and the j-th TF to belong to the same time segment if

$$\sigma_2(i,j,k)<c_s, \tag{8.20}$$

where c_s is 1:1 related to c_h as we will show below.

We now show that these two methods are equivalent. Since $\arccos(\cdot)$ is a monotonous decreasing function, Eq. (8.19) (i.e., the Hermitian angles) is equivalent to

$$\left|\hat{\bar{\mathbf{a}}}(i,k)^H\hat{\bar{\mathbf{a}}}(j,k)\right|>c, \tag{8.21}$$

where $\arccos(c)=c_h$ and we used the fact that $\left\|\hat{\bar{\mathbf{a}}}(t,k)\right\|_2=1$ for all t and k. For the singular value method, the second largest singular value of $\mathbf{A}(i,j,k)$ is the square root of the second largest eigenvalue of

$$\begin{aligned}&\mathbf{A}(i,j,k)^H\mathbf{A}(i,j,k)\\&\qquad=\begin{bmatrix}1 & \hat{\bar{\mathbf{a}}}(i)^H\hat{\bar{\mathbf{a}}}(j)\\ \hat{\bar{\mathbf{a}}}(j)^H\hat{\bar{\mathbf{a}}}(i) & 1\end{bmatrix},\end{aligned} \tag{8.22}$$

which is given by $\sigma_2 = \sqrt{1 - \left|\hat{\mathbf{a}}(i,k)^H \hat{\mathbf{a}}(j,k)\right|}$. Hence, Eq. (8.20) is also equivalent to Eq. (8.21) with $\sqrt{1-c} = c_s$.

Note that if the i-th TF and the j-th TF belong to the same time segment, the inner products $\left|\hat{\mathbf{a}}(i,k)^H \hat{\mathbf{a}}(j,k)\right|$ are close to one for all frequencies. Therefore, we average the inner products for all frequency bins to express the similarities between the i-th and the j-th time frames with a single quality, that is,

$$B(i,j) = \frac{\sum_{k=1}^{K} \left|\hat{\mathbf{a}}(i,k)^H \hat{\mathbf{a}}(j,k)\right|}{K}. \tag{8.23}$$

Furthermore, by assuming the source does not change to other positions in between the i-th TF and the j-th TF when $B(i,j)$ is sufficiently large, the time frames in between them should also belong to the same time segment.

We assume the time segments before the β-the time segment have been determined (i.e., $\{t_i, T_i\}_{i=1}^{\beta-1}$ is known). Since $t_\beta = t_{\beta-1} + T_{\beta-1}$, t_β is known. We only need to estimate the length T_β of the β-th time segment. We first calculate $B\left(t_\beta, t_\beta + j - 1\right)$ for $j = 1, \cdots, T$. Then, we find T_β by

$$\max\left\{ j | B\left(t_\beta, t_\beta + j - 1\right) > c, j = 1, \cdots, T \right\}. \tag{8.24}$$

We then execute the JMLE algorithm using time frames from t_β to $t_\beta + T_\beta - 1$ jointly to estimate the RTF vector for the β-th time segment and the PSDs for time frames from t_β to $t_\beta + T_\beta - 1$.

The JMLE method combined with the adaptive time segmentation method is summarized in Algorithm 1.

8.5. EXPERIMENTS

To evaluate the performance of the proposed method, we simulate the microphone signals by convolving the speech signal from the TIMIT data base [13] with the recorded room impulse responses (RIRs) from [14]. The setup for recording the RIRs is shown in Fig. 8.3, where 8 microphones are placed in a line with inter distance of 8 cm. The sound source is placed at a distance of 2 m from the center of the microphone array at different angles. We also add white Gaussian noise to the reverberant signals to simulate the microphone self noise even though the used JMLE method assumes the signals are noise free. The target signal-to-self noise ratio (SNR) is set to 50 dB, which is calculated over the whole time duration since the target signal is non-stationary. The noisy microphone signals are sampled at a rate of 16 kHz and processed by the STFT procedure. That is, we use the square-root Hann window with 50% overlap between adjacent sub-time frames and an FFT, both with a length of 512 samples (32 ms). Note that each time frame consists of L_s = 40 overlapping sub-time frames and thus has a duration of 0.64 s. The speed of sound is set to 344 m/s. The reverberation time is 0.61s. The threshold c is set to 0.6 in the experiments based on some initial experiments.

```
Algorithm 9: TS-JMLE
  Input: $\{\hat{\mathbf{P}}_{\mathbf{y}}(t)\}_{t=t_\beta}^{t_\beta+T-1}$, $\hat{\mathbf{\Gamma}}$,$c$,$IterN$
  Output: $T_\beta$,$\hat{\mathbf{a}}(\beta,k)$ and $\{\hat{\phi}_s(t,k),\hat{\phi}_\gamma(t,k)\}_{t=t_\beta}^{t_\beta+T_\beta-1}$
  for all $k$, $t=t_\beta:t_\beta+T-1$ do
      Estimate $\hat{\hat{\mathbf{a}}}(t,k)$, $\hat{\hat{\phi}}_s(t)$ and $\hat{\phi}_\gamma(t)$ using Eqs. (8.12) to (8.14).
  Calculate $B(t_\beta,j)$ for $j=t_\beta:t_\beta+T-1$ using Eq. (8.23);
  Estimate $T_\beta$ by Eq. (8.24).
  for all $k$ do
      for iter=1:IterN do
          Calculate $\mathbf{P}(\beta)=\sum_{t=t_\beta}^{t_\beta+T_\beta-1}\frac{\hat{\hat{\phi}}_s(t)}{\hat{\phi}_\gamma(t)+\hat{\hat{\phi}}_s(t)}\frac{1}{\hat{\phi}_\gamma(t)}\hat{\mathbf{P}}_{\mathbf{w}}(t)$
          Estimate $\hat{\hat{\mathbf{a}}}(\beta)$ using the principal eigenvector of $\mathbf{P}(\beta)$.
          Estimate $\hat{\hat{\phi}}_s(t)$ and $\hat{\phi}_\gamma(t)$ for $t=t_\beta,\cdots,t_\beta+T_\beta-1$ using Eqs. (8.17)
           and (8.18).
      Estimate $\hat{\mathbf{a}}(\beta)$ and $\hat{\phi}_s(t)$ by $\hat{\mathbf{a}}=\frac{\mathbf{L}\hat{\hat{\mathbf{a}}}}{\mathbf{L}\hat{\hat{\mathbf{a}}}\mathbf{e}_1}$ and $\hat{\phi}_s=\frac{\hat{\hat{\phi}}_s}{\hat{\mathbf{a}}^H\mathbf{\Gamma}^{-1}\hat{\mathbf{a}}}$.
```

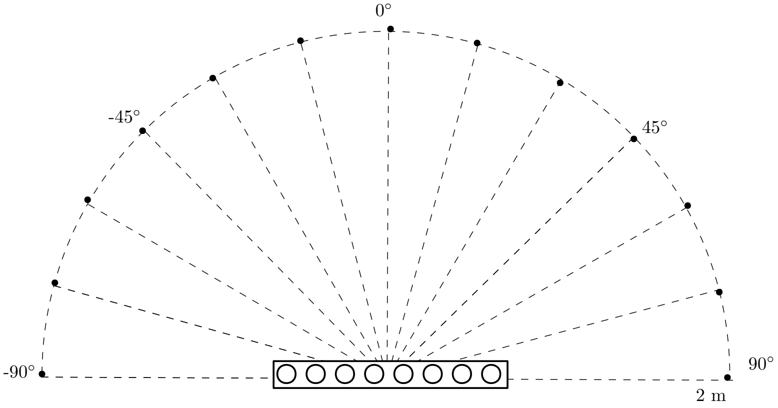


Figure 8.3: Geometric setup for the real RIRs.

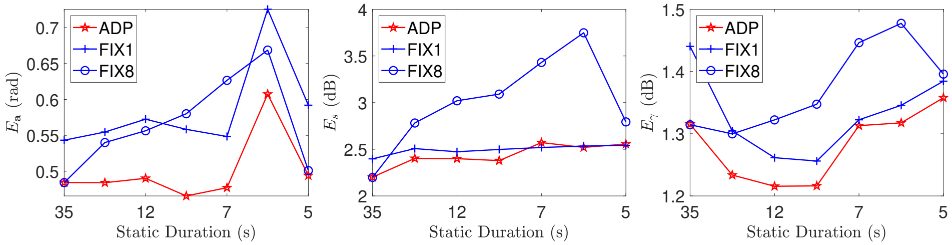
To compute the performance between the proposed and reference methods, we use the averaged Hermitian angle error (in rad) [15]

$$E_{\mathbf{a}}=\frac{\sum_{t=1}^{N}\sum_{k=1}^{K/2+1}\operatorname{acos}\left(\frac{\left|\mathbf{a}^H(t,k)\hat{\mathbf{a}}(t,k)\right|}{\|\mathbf{a}^H(t,k)\|_2\|\hat{\mathbf{a}}(t,k)\|_2}\right)}{N(K/2+1)}, \tag{8.25}$$

with N the total number of time frames across all segments. Note that we average the errors over time frames instead of over time segments because the estimated time segments might have different duration. For the PSDs estimates, we use the symmetric log-error distortion measure [16]

$$E_i = \frac{10 \sum_{t=1}^{N} \sum_{k=1}^{K/2+1} \left|\log\left(\frac{\phi_i(t,k)}{\hat{\phi}_i(t,k)}\right)\right|}{N\,(K/2+1)}, \tag{8.26}$$

with $i \in \{s, \gamma\}$.



(a) RTF estimation error vs static duration at each position.

(b) Target source PSD estimation error vs static duration at each position.

(c) Late reverberation PSD estimation error vs static duration at each position.

Figure 8.4: Performance comparison of the JMLE method combined with adaptive time segmentation and fix time segmentation.

In Fig. 8.4, we show the estimation performance comparison of the JMLE method combined with different time segmentation strategies. 'ADP' denotes our proposed adaptive time segmentation method. The maximum size T of a time segment is set to 8. 'FIX1' denotes considering a single time frame (TF) as a time segment (TS), and 'FIX8' denotes considering every 8 TFs as a TS. The speech signal has a duration of 35 s. In the experiment, we simulate the time varying RTF by changing the source position from 0° to $(k-1)\times 15^\circ$ by 15° every $\frac{35}{k}$ seconds. As k increases from 1 to 7, the duration with which the source stays at the same position thus decreases from 35 s to $\frac{35}{7} = 5$ s along the x-axis in the graphs in Fig. 8.4. For the RTF estimation error, the proposed ADP has the smallest error, which is about 0.1 rad smaller than FIX1 for different static time durations. The error for FIX8 fluctuates, but is always larger than ADP except for a static duration of 35 s and 5 s when ADP and FIX8 are approximately equal. For the source being static at 0° for 35s, all TFs share the same RTF. Therefore, ADP gives us the maximum size of the time segment, which is equal to FIX8 (considering 8 TFs as a TS) and much better than FIX1 (considering 1 TF as a TS). For the source staying at each position for 5 s, since the duration of a TF is 0.64 s, each TS contains about $\frac{5}{0.64} \approx 8$ TFs. therefore, the error of FIX8 is also close to ADP in this case. For the errors of the target PSD and the late reverberation PSD, ADP also has the best performance.

8.6. CONCLUSIONS

We presented an algorithm to obtain an optimal adaptive time segmentation and combined this with our previously published joint maximum likelihood estimator (JMLE) for jointly estimating the RTF, source PSD and late reverberation PSD of a single source in a reverberant environment. We proved that comparing the Hermitian angle of two RTF estimates to a threshold is equivalent to comparing the second largest singular value of the matrix combining these two RTF estimates or their inner product. We thus provided a thresholding method based on averaged inner products over all frequency bins. The JMLE combined with our adaptive time segmentation outperforms the JMLE combined with fixed time segmentation.

REFERENCES

[1] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, “Robust joint estimation of multimicrophone signal model parameters”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[2] C. Li, J. Martinez, and R. C. Hendriks, “Joint Maximum Likelihood Estimation of Microphone Array Parameters for a Reverberant Single Source Scenario”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 695–705, 2023.

[3] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood PSD estimation for speech enhancement in reverberation and noise”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, 2016.

[4] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 544–548.

[5] B. Schwartz, S. Gannot, and E. A. Habets, “Two model-based EM algorithms for blind source separation in noisy environments”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2209–2222, 2017.

[6] I. Kodrasi and S. Doclo, “Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, 2018.

[7] I. Kodrasi and S. Doclo, “Joint Late Reverberation and Noise Power Spectral Density Estimation in a Spatially Homogeneous Noise Field”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 441–445.

[8] M. Tammen, S. Doclo, and I. Kodrasi, “Joint Estimation of RETF Vector and Power Spectral Densities for Speech Enhancement Based on Alternating Least Squares”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 795–799.

[9] Y. Laufer and S. Gannot, “Scoring-Based ML Estimation and CRBs for Reverberation, Speech, and Noise PSDs in a Spatially Homogeneous Noise Field”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 61–76, 2020.

[10] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, “Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming”, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[11] R. C. Hendriks, R. Heusdens, and J. Jensen, “Adaptive time segmentation for improved speech enhancement”, *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2064–2074, 2006.

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

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993.

[14] E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments”, in *Proc. IEEE Int. Workshop Acoust. Signal Enhanc.*, Sep. 2014.

[15] R. Varzandeh, M. Taseska, and E. A. P. Habets, “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”, in *Proc. IEEE Hands-free Speech Commun. Microphone Arrays*, 2017, pp. 11–15.

[16] R. C. Hendriks, J. Jensen, and R. Heusdens, “DFT domain subspace based noise tracking for speech enhancement”, in *Proc. Interspeech*, 2007, pp. 830–833.

9 DISCUSSION, CONCLUSION AND FUTURE WORK

In this chapter, we first draw the conclusions of the dissertation, which summarize the main results and contributions, after which we draw the conclusions by reflecting on the research questions from Chapter 1. After that, we will also discuss some possible future research problems and suggestions on how to approach these.

9.1. CONCLUSION

An essential problem of microphone array signal processing, is to estimate the parameters that describe the scene. These parameters include, for example, ATFs from the (point) source to the microphones, source PSDs, the late reverberation PSDs, etc. With these parameters estimated, further processing such as noise reduction, source separation and dereverberation can be achieved. As in practice the parametric model describing the scenario can come in different levels of complexity, we addressed in this work different acoustic scenarios, from a single-source reverberant scenario, to a multi-source reverberant and noisy scenario. For each assumed scenario, at least one novel estimator has been proposed. The estimation performance of each proposed method is evaluated. The computational cost, which is an important factor for any application, is also investigated by providing the theoretical computational complexity or real run time. In addition, given estimated parameters, we apply linear filters such as the multichannel Wiener filter, to extract the target signal and show the performance on noise reduction, predicted speech quality and predicted speech intelligibility.

The first main contribution of this dissertation is to use multiple time frames that share the same RTFs for parameter estimation. In most previous existing works on parameter estimation, these methods use each time frame independently, even when the source is assumed static and the RTF is constant for consecutive time frames. In [1], it was also proposed to use multiple time frames in the SCFA method. It was shown that the more time frames are used, the better estimation performance is obtained. Therefore, for most methods proposed in this dissertation, which use a single time frame, we always provide their extensions of using multiple time frames.

In the following, we provide the specific results and contributions of this dissertation for each scenario assumed below

- A single reverberant source (Chapter 3)
  For the simplest scenario in this dissertation, the parameters we aimed to estimate include the RTF and the PSD of the source, and the PSD of the late reverberation. For this, we considered the maximum likelihood estimator when using a single time frame, although the solution was already given in [2], we provide an alternative proof. With this proof, we could easily extend the estimator to using multiple time frames. In experiments, we showed that the proposed estimator, JMLE, has similar performance compared to SCFA when the SNR is relatively high. The proposed algorithm outperforms the other reference methods that consist of a combination of several existing state-of-the-art methods. With respected to the computational cost, JMLE shows much lower computational complexity then SCFA.

- A single reverberant and noisy source (Chapter 4, 5)
  The proposed method in Chapter 3 does not work well when the SNR is low. Therefore, an estimator for a single reverberant and noisy source is in need. In Chapter 4 we first focus on the RTF estimator. We proposed here an estimator that uses only off-diagonal elements of the simplified covariance matrix. This estimator is not affected by the late reverberation and the noise PSD that only appear on the diagonal elements.

  In Chapter 5, we then consider the joint estimator of the RTFs, the PSDs of the source, the PSDs of the late reverberation, and the PSDs of the ambient noise. We first improved an existing alternating least square (ALS) based method that uses a single time frame and then extend it to use multiple time frames. Furthermore, we found out that the ALS based methods might have negative estimates during the iteration steps while the PSDs are by definition non-negative. To solve this issue, unlike replacing the negative estimates with a machine precision small number, we used estimated PSDs of previous time frames to constrain the PSDs of current time frames. We also proposed robust upper bounds of the estimates to avoid large overestimation errors.

  The experiments demonstrate that the estimation performance of both proposed methods is similar to SCFA, which outperforms the other reference methods. The computational cost of our proposed methods is significantly lower than SCFA.

- Multiple reverberant and noisy sources (Chapter 6, 7)
  For scenarios with multiple sources, we showed in Chapter 7 that we need multiple time frames to get unique estimates of the RTFs. In Chapter 6, we assume that the environment is close to non-reverberant and noiseless. We modified the well-known SOBI method to estimate the RTFs. We proposed to average covariance matrices of different number of time frames that have been sorted based on rough model mismatch errors at the first step of SOBI. Therefore, we eventually get several estimates

for each RTF. We then select the estimates related to the optimal cost function as the final estimates. This method has satisfying performance for low reverberation time and high SNR. However, for high reverberation time or low SNR, we need a more robust estimator.

In Chapter 7, we therefore address the multi-source reverberant and noisy scenario and proposed an estimator for the RTFs of the sources and the PSDs of the sources and the late reverberation. We first proposed a late reverberation PSD estimator without the knowledge of RTFs. Then, we estimate the RTFs and PSDs of the sources with the estimated late reverberation PSD. Similar to Chapter 6, we modify the first step of SOBI. Unlike using the first covariance matrix in SOBI or averaging some of the covariance matrices as in Chapter 6, we analyzed the variances of the error matrix of possible linear combinations of these covariance matrices and found the optimal one by minimizing the variance. The robustness and effectiveness of our proposed method has been demonstrated in the experiments. The computational cost of the proposed method is about 850 time faster than SCFA, with both methods achieving similar estimation performance.

The last contribution of this dissertation is the adaptive time segmentation method we proposed in Chapter 8. The prior information on which time frames share the same RTF in the estimators that use multiple time frames is unknown in practice for a single non-static source. We proposed an algorithm to obtain an optimal adaptive time segmentation and combine this method with our proposed JMLE method from Chapter 3 for a single reverberant non-static source. We proposed a thresholding method after proving that comparing the Hermitian angle or their inner product of two RTFs to a threshold is equivalent to comparing the second largest singular value of the matrix combining these two RTFs. In the experiments, it has been shown that using the adaptive time segmentation outperforms using a fixed time segmentation.

In summary, we answered all the research questions presented in Fig. 1.4. For RQ 1.1, we proposed the JMLE method in Chapter 3 to estimate the RTF and the PSD of the source and the PSD of the late reverberation using multiple time frames. For RQ 1.2, we proposed a RTF estimator which is insensitive to noise PSD errors in Chapter 4. For RQ 1.3, we proposed the JALS method in Chapter 5 to estimate the RTF of the source and the PSDs of the source, the late reverberation and the noise using multiple time frames. For RQ 2.1, we proposed an RTF estimator which works in a low reverberant and high-SNR environment in Chapter 6. For RQ 2.2, in Chapter 7, we proposed a robust late reverberation PSD estimator and a joint estimator of the RTFs and the PSDs of the sources (called MVJD) which is robust to reverberant and noisy environment. Finally, for RQ 3, we proposed an adaptive time segmentation method in Chapter 8.

9.2. FUTURE RESEARCH

In this section, we share some research directions that are worth for further investigation. For some of them, we give our suggestions on possible solutions.

SOURCE NUMBER ESTIMATION

One essential problem we did not address in this dissertation, is the estimation of the number of sources. Although we assumed the number of sources is known, in practice this is usually unknown. The estimation of the number of sources is therefore an important problem. This problem is challenging compared to the traditional array signal processing problem. The reason is the approximations we made in the signal model. For ease of analysing, we have made several approximations for the STFT domain signal model. For the direct and early component, we used the MTF approximation such that the signal is a multiplication of the sound source and the acoustic transfer function (ATF) vector. The covariance matrix of this component is of rank 1. However, the actual signal model is an inter-frame and inter-band convolution, of which the covariance matrix is not necessarily rank 1. Also, for the late reverberation component, an ideal diffuse noise field is assumed, which might be violated in practice. These model inaccuracies hinders the way to estimate the number of sources.

LESS MICROPHONES THAN SOURCES

In this dissertation, we assumed that the number of sources is less than the number of microphones. In practice, it is also possible that the number of sources exceeds the number of microphones. In such cases, when using a sufficient number of time frames, the covariance matrix is of full rank, with the rank being less than the number of RTFs. Some of the estimators proposed in this dissertation will no longer work. For instance, in Chapter 7, we estimated the late reverberation PSD based on the $M-R$ smallest eigenvalues of the covariance matrix, with M the number of microphones and R the number of sources. If $M<R$, $M-R$ is negative, and the proposed estimator in Chapter 7 will not work. For this problem, we suggest to use methods based on tensor decomposition, where the original covariance matrix can be extended to a tensor by considering the time index.

PERMUTATION ALIGNMENT

For multiple sources, we have to deal with the permutation ambiguity after estimating the RTF matrix and the source PSDs, i.e., for one specific source, it is unknown which column of the RTF matrix belongs to which source, neither does the PSD. Some solutions to this problem were investigated in [3], [4]. In the experiments of this work, we used the oracle RTF matrix as guidance to permute the columns of the estimated RTF matrix per time-frequency tile. In practice, we need to select an existing method or develop an algorithm to solve this problem.

ADAPTIVE TIME SEGMENTATION FOR MULTIPLE NON-STATIC SOURCES

We proposed an adaptive time segmentation method for a single source. The next problem we would naturally consider is how to obtain an optimal time segmentation for multiple non-static sources such that each time segment contains as many time frames that share the same RTF matrix as possible. In this problem, it is also of interest to know, at the changing point, which source changes position.

REFERENCES

[1] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multimicrophone signal model parameters", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1136–1150, 2019.

[2] P. Hoang, Z.-H. Tan, J. M. de Haan, and J. Jensen, "Joint Maximum Likelihood Estimation of Power Spectral Densities and Relative Acoustic Transfer Functions for Acoustic Beamforming", in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6119–6123.

[3] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation of many speech signals using near-field and far-field models", *EURASIP J. Adv. Signal. Process.*, vol. 2006, pp. 1–13, 2006.

[4] D. Nion, K. N. Mokios, N. D. Sidiropoulos, and A. Potamianos, "Batch and adaptive PARAFAC-based blind separation of convolutive speech mixtures", *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1193–1207, 2009.

ACKNOWLEDGEMENTS

In my previous opinion, the way to get a PhD should be unpleasant and full of suffering. Luckily, I enjoyed a lot during the past years, owing to the persons who helped me and the work-life balance culture in this green land. Expressing my gratitude to all of them is challenging with this short page and my poor memory, but I will try my best.

First, I would like to thank my promoters, Prof. Richard Hendriks and Prof. Alle-Jan van der Veen. Richard, every meeting with you gave me motivation and energy to continue my research works. Thank you for listening to me with patience and giving me valuable suggestions whenever I felt a bit lost. You guided me in the right direction. You gave me valuable and very timely feedback. Thank you for being my supervisor and friend! Alle-Jan, thank you for all the suggestions, criticism and guidance. Thank you for being such a wonderful group leader! Our group will not have this inclusive environment without you.

I owe many thanks to Dr. Jorge Martinez Castaneda. Jorge, I am lucky to have you as my daily supervisor for the early period of my PhD. I am more than lucky to have you as my friend. Thank you for supporting me in my research and life. The discussions with you are very valuable to me. I will always remember the sunny days when we had all kinds of discussions in your office, from which I learned a lot!

Thank you my colleagues in the CAS/SPS group, for being so nice to me. Thank you to all the experienced researchers in the group. Prof. Geert Leus. Thank you for all the inspiring discussions! Thank you, Dr. Jie Zhang. Thank you for telling me how great it is to do a PhD in the CAS group and for helping me in all aspects! Thank you Prof. Richard Heusdens and Dr. Qiongxiu Li, for the unforgettable memories during the two ICASSPs I attended. Thank you, my office mates in EWI, Miao, Giovanni, Jordi, Yujie and Silverio, for making 17.240 an enjoyable working space! Thank you, my office mates in the two offices of building 28. Thank you my first office mates, Anu, Alberto and Costas, for all the laughing moments! My English was not good enough to understand some of the jokes you made, but I enjoyed those laughing moments with you. Thank you, my second office mates, Didem, Aybuke, Hanie, and Metin, for being so nice to me that I felt it's safe to share everything with you. Thank you, Shuoyan, Sinian, and Chen, for taking care of Mickey whenever we asked for your help, and those happy mahjong games. Also, thank you, the young fellows, Sofia, Seline, Yanbin, Ellen, Ids, Ruben, Zhonggang, Peiyuan... for making SPS such a wonderful group! Last but not least, thank you, Laura, for supporting the whole group professionally and warmly!

Yanbo, thank you for all the support and trust. You have many admirable qualities from which I can continuously learn. Thank you and your wife, Yansu, for treating us like your family. I also want to say thank you to Matthieuw, Liming, Sen, and Yadan, for those joint memories in Delft.

Finally, I would like to express my deepest gratitude to my parents, Chuanxin Li and Shiling Song. Thank you for supporting me with all you have for my whole life. Thank you for your endless and unconditional love. I also would like to thank my brother, Peiheng Li, and my sister-in-law, Jieshu He, for taking care of everything at home. At last, I want to express my sincere gratitude to my wife Li (Karlie) Liu. Thank you for accompanying me in this place so far from home, with your courage and full trust in me. Thank you for enlightening my life with your cheerful personality and the cute cat Mickey. Thank you for all the support and love.

CURRICULUM VITæ

Changheng Li

Changheng Li was born in Henan, China, in 1996. He received the B.Sc. degree in Applied Mathematics (the School of the Gifted Young) and the M.Sc. degree in Electrical Engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2017 and 2020, respectively. He is currently a PhD candidate with the Signal Processing Systems (SPS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology (TUD). His current research interest focuses on microphone array signal processing and speech enhancement.