Low Latency Convolutional Blind Source Separation

by

Jiawen Chua

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Engineering and Computer Science.

Victoria University of Wellington
2019

Abstract

In most real-time systems, particularly for applications involving system identification, latency is a critical issue. These applications include, but are not limited to, blind source separation (BSS), beamforming, speech dereverberation, acoustic echo cancellation and channel equalization. The system latency consists of an algorithmic delay and an estimation computational time. The latter can be avoided by using a multi-thread system, which runs the estimation process and the processing procedure simultaneously. The former, which consists of a delay of one window length, is usually unavoidable for the frequency-domain approaches. For frequency-domain approaches, a block of data is acquired by using a window, transformed and processed in the frequency domain, and recovered back to the time domain by using an overlap-add technique.

In the frequency domain, the convolutive model, which is usually used to describe the process of a linear time-invariant (LTI) system, can be represented by a series of multiplicative models to facilitate estimation. To implement frequency-domain approaches in real-time applications, the short-time Fourier transform (STFT) is commonly used. The window used in the STFT must be at least twice the room impulse response which is long, so that the multiplicative model is sufficiently accurate. The delay constraint caused by the associated blockwise processing window length makes most the frequency-domain approaches inapplicable for real-time systems.

This thesis aims to design a BSS system that can be used in a real-time scenario with minimal latency. Existing BSS approaches can be integrated into our system to perform source separation with low delay without affecting the separation performance. The second goal is to design a BSS system that

can perform source separation in a non-stationary environment.

We first introduce a subspace approach to directly estimate the separation parameters in the low-frequency-resolution time-frequency (LFRTF) domain. In the LFRTF domain, a shorter window is used to reduce the algorithmic delay of the system during the signal acquisition, e.g., the window length is shorter than the room impulse response. The subspace method facilitates the deconvolution of a convolutive mixture to a new instantaneous mixture and simplifies the estimation process.

Second, we propose an alternative approach to address the algorithmic latency problem. The alternative method enables us to obtain the separation parameters in the LFRTF domain based on parameters estimated in the high-frequency-resolution time-frequency (HFRTF) domain, where the window length is longer than the room impulse response, without affecting the separation performance.

The thesis also provides a solution to address the BSS problem in a non-stationary environment. We utilize the “meta-information” that is obtained from previous BSS operations to facilitate the separation in the future without performing the entire BSS process again. Repeating a BSS process can be computationally expensive. Most conventional BSS algorithms require sufficient signal samples to perform analysis and this prolongs the estimation delay. By utilizing information from the entire spectrum, our method enables us to update the separation parameters with only a single snapshot of observation data. Hence, our method minimizes the estimation period, reduces the redundancy and improves the efficacy of the system.

The final contribution of the thesis is a non-iterative method for impulse response shortening. This method allows us to use a shorter representation to approximate the long impulse response. It further improves the computational efficiency of the algorithm and yet achieves satisfactory performance.

Declaration

The contents of this thesis are the results of original research and have not been submitted to any other university or institution for a higher degree. Most of the contents have been published either in journal papers or peer review conference proceedings. These publications are:

1. J. Chua, G. Wang, W. B. Kleijn, “Convolutional blind source separation with low latency”. In *Acoustic Signal Enhancement (IWAENC), 2016 IEEE International Workshop on* (pp. 1-5). IEEE. (Nominated for Best Student Paper Award)
2. J. Chua, W. B. Kleijn, “Non-iterative impulse response shortening method for system latency reduction”. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 581-585). IEEE.
3. J. Chua, W. B. Kleijn, “A Low Latency Approach for Blind Source Separation”. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, accepted.

The research in this thesis has been performed jointly with Prof. W. Bastiaan Kleijn. The substantial majority of this work is my own.

Acknowledgments

First and foremost, I would like to express my utmost gratitude to my supervisor, Prof. Bastiaan Kleijn for guiding, assisting and supporting me throughout the studies. His ways of thinking about and solving problems inspired me and I enjoyed great benefits from that. Without his help, the thesis would not have reached the current stage and I greatly appreciate for precious time and effort he committed to this thesis.

I am grateful to Victoria University of Wellington and GN ReSound for providing me with a scholarship. I am also grateful to the Communication and Signal Processing (CaSP) group for preparing seminars and presentations that greatly benefited me over the last four years.

I also thank my flatmates, colleagues and friends for their encouragement and continuous support during hard times.

Finally, I would like to thank my parents, brothers and my partner Shirley Chua, who have provided me with moral and emotional support.

Contents

1	Introduction	1
1.1	Problem Statement and Potential Solutions	1
1.2	Limitations of Conventional Approaches	3
1.3	Contribution of the Thesis	6
1.3.1	Algorithmic latency	7
1.3.2	Estimation delay	8
1.4	Organization of the Thesis	8
2	Background and Literature Review	11
2.1	Possible BSS Applications	11
2.2	System Model	13
2.2.1	The HFRTF-Domain Representation	14
2.2.2	The LFRTF-Domain Representation	15
2.3	Existing Blind Source Separation Approaches	17
2.3.1	Independent Component Analysis	17
2.3.2	Joint Diagonalization	20
2.3.3	Generalized Eigenvalue Decomposition	21
2.3.4	Blind Channel Identification	23
2.3.5	Sparsity in the Time-Frequency Domain	26
2.3.6	Clustering	27
2.3.7	Nonnegative Matrix Factorization	30
2.4	Permutation and Scaling Ambiguity	31
2.5	Existing Beamforming Techniques	33

2.5.1	Capon's Method and LCMV beamformer	34
2.5.2	Generalized Sidelobe Canceller	36
2.5.3	Robustness Improvement	38
2.6	Discussion	40
3	Subspace Method for Blind Source Separation With Convo-	
	lutive Transfer Function Model	43
3.1	Introduction	44
3.2	Blind Source Separation Model	45
3.3	Conventional Subspace Methods	46
3.4	Proposed Subspace Method	50
3.4.1	Problem Formulation	50
3.4.2	Principle of the Proposed Method	52
3.4.3	Summary of the Proposed Algorithm	54
3.5	Experimental Results	55
3.5.1	Experimental Setup and Procedure	55
3.5.2	Experimental Results	58
3.5.3	Discussion	60
3.6	Conclusion	63
4	Low-Latency Approach for Convolutional Blind Source Sepa-	
	ration	69
4.1	Introduction	70
4.2	Time-Frequency Transform	72
4.2.1	Analysis and Synthesis Operator of the STFT	72
4.2.2	Time-Domain Signal Model	74
4.2.3	TF-Domain Signal Representations	74
4.2.4	Instantaneous Approximation	75
4.3	Proposed Approach	77
4.3.1	Blind Source Separation Model	79
4.3.2	Time-domain Approach	80

4.3.3	Low-Frequency-Resolution Time-Frequency (LFRTF)-Domain Approach	87
4.4	Results	92
4.4.1	Experimental Setup and Procedure	92
4.4.2	Experimental Results	95
4.5	Conclusion	100
5	Blind Source Separation Method for Non-Stationary Scenarios Using Meta-Information	103
5.1	Introduction	104
5.2	Problem Formulation	105
5.3	Time-Invariant Channel	107
5.4	Time-Varying Channel	111
5.4.1	Model of a Moving Source	112
5.4.2	Proposed Method	114
5.4.3	Robustness Improvement	116
5.5	Results	117
5.5.1	Experimental Setup and Procedure	118
5.5.2	Experimental Results for a Static Interfering Source . .	120
5.5.3	Experimental Results for a Moving Interfering Source .	122
5.6	Conclusion	125
6	Non-Iterative Impulse Response Shortening Method	127
6.1	Introduction	127
6.2	Impulse Response Shortening	129
6.2.1	System Identification Model	129
6.2.2	Proposed Method	131
6.3	Blind Source Separation Application	132
6.3.1	Problem Formulation of BSS	132
6.3.2	Estimating the Demixing Operator	133
6.4	Cross-talk cancellation application	135
6.5	Results	136

6.5.1	BSS Experimental Setup	136
6.5.2	BSS Simulation Results	137
6.5.3	Cross-talk Cancellation Experimental Setup	138
6.5.4	Cross-talk Cancellation Simulation Results	138
6.6	Conclusion	140
7	Conclusions and Future Work	143
7.1	Conclusions	143
7.1.1	Algorithmic latency	143
7.1.2	Estimation delay	145
7.2	Future Research Directions	146

List of Figures

1.1	A two-threaded system for a blind source separation application	5
1.2	Overall structure of the thesis contribution.	7
2.1	The block diagram of a SIMO system [62].	24
2.2	A scatter plot of an instantaneous mixture of two sparse sources, where the red and green lines represent the mixing vectors of the first source and the second source, respectively. The x-axis (x1) and y-axis (x2) denote the amplitudes of the first observation and the second observation, respectively.	28
2.3	A histogram of probability, for which the observation has the same direction as $\hat{a}(\theta)$ based on the data in Fig. 2.2.	29
2.4	A block diagram of the GSC beamformer [57].	37
3.1	An illustrative example of the operations from (3.16) to (3.18).	52
3.2	The positions of the microphones and the sources	56
3.3	The separation performance (SIR) of various techniques, where the demixing operators were computed using the first 120 TF observation samples.	59
3.4	The separation performance (SDR) of various techniques, where the demixing operators were computed using the first 120 TF observation samples.	60

- 3.5 The separation performance (SIR) of various techniques, where the demixing operators were computed using the first 1.92 s observation samples. 65
- 3.6 The separation performance of various techniques, where the demixing operators were computed using the first 1.92 s observation samples. 66
- 3.7 The separation performance (SIR) of various techniques, where the demixing operators were computed using the observations with different length and the room reverberation time was 0.15 s. 67
- 3.8 The separation performance (SDR) of various techniques, where the demixing operators were computed using the observations with different length and the room reverberation time was 0.15 s. 68
- 4.1 A matrix representation of a synthetic room impulse response $\Phi \ddot{\Lambda}^{ml} \Phi^*$ in the time domain using the model in (4.27) with $m = 1$, $l = 1$, $N_x = 256$, $N_h = 32$, $N_w = 64$ and $N_l = 16$. The y-axis and x-axis represent the rows and the columns of the matrix $\Phi \ddot{\Lambda}^{ml} \Phi^*$, respectively. The white coloured parts indicate negative infinity values. 81
- 4.2 The relationship between $\underline{n}_p$, $\bar{n}_p$, $\underline{N}_p$ and $\bar{N}_p$ as shown in Fig. 4.1, where $p = 7$, $N_f = 16$ and $N_w = 64$ 84
- 4.3 A matrix representation of a synthetic room impulse response $\Psi^* \hat{\mathbf{H}}^{ml} \Psi$ in the LFRTF domain using the model in (4.44) with $m = 1$, $l = 1$, $N_x = 256$, $N_h = 32$, $N_w = 64$, $N_l = 16$, $N_b = 16$ and $N_a = 8$. The y-axis and x-axis represent the rows and the columns of the matrix $\Psi^* \hat{\mathbf{H}}^{ml} \Psi$, respectively. The white coloured parts indicate negative infinity values. . . 88
- 4.4 The separation performance (SIR) of various techniques with different reverberation times, where the number of microphones $M = 8$ and the algorithmic latency was 16 ms. 95

4.5	The separation performance (SDR) of various techniques with different reverberation times, where the number of microphones $M = 8$ and the algorithmic latency was 16 ms.	96
4.6	Performance of the proposed methods (SIR) against different number of microphones M at different algorithmic latency, where the reverberation time is $\beta = 0.2$ s.	97
4.7	Performance of the proposed methods (SDR) against different number of microphones M at different algorithmic latency, where the reverberation time is $\beta = 0.2$ s.	98
4.8	Performance of the proposed methods (SIR in (a) and SDR in (b)) against algorithmic latency using six microphones, where the reverberation time is $\beta = 0.2$ s.	98
4.9	Performance of the proposed methods (SIR in (a) and SDR in (b)) against algorithmic latency using six microphones, where the room size is $9 \text{ m} \times 6 \text{ m} \times 2.7 \text{ m}$ and the reverberation time is $\beta = 0.5$ s.	99
5.1	The GSC-like structure of the BSS problem for extracting the l^{th} desired source.	111
5.2	The positions of the microphones and the desired sources. The black crosses indicate the microphones, the blue dot indicates the first desired source while the red dot indicates the second desired source.	119
5.3	The positions of the microphones indicated by black crosses, desired sources indicated by red and blue dots as well as the moving interference signal, where the green dot indicates the initial position while the green line indicates the moving path.	120
5.4	Separation performance of various methods when a new static interfering source is introduced to the BSS system.	121

5.5	Separation performance of various methods when a new slow-moving interfering source (1 m/s) is introduced to the BSS system.	122
5.6	Separation performance of various methods when a new fast-moving interfering source (10 m/s) is introduced to the BSS system.	123
5.7	Partial results obtained in the first trial of experiments for a slow-moving interfering source.	124
5.8	Partial results obtained in the first trial of experiments for a fast-moving interfering source.	125
6.1	Cross-talk cancellation performance of the proposed approach, where $Q = 4$ and the pre-filter length was 128.	140
6.2	Performance comparison with different pre-filter length.	140

List of Tables

3.1	The average SIR performance and computational time of the first set of the experiments by using various techniques. . . .	61
3.2	Summary of the computational cost for each BSS method. . .	63
6.1	Performance comparison between various approaches low-latency for BSS using the ICA method.	138
6.2	Performance comparison between various low-latency approaches for BSS using the Modified-TIFROM approach.	139

Chapter 1

Introduction

This chapter provides an introduction to the thesis. It first defines a statement of the problem that the thesis aims to solve and next it briefly discusses the limitations of the conventional approaches. It concludes with the organization of the thesis.

1.1 Problem Statement and Potential Solutions

Human brains can extract a particular sound source of interest even though the environment is noisy. This phenomenon is so-called “the cocktail party effect”, which is the ability to focus one’s auditory attention on an interesting stimulus while filtering out unwanted stimuli [193]. The same functionality can be particularly useful in many communication devices, for instance, hearing aids, mobile phones, video calling, smart speakers and so on. For hearing aids, mobile phones and video calling, it can help users to comprehend speech better in a noisy environment by delivering a source of interest and suppressing unwanted signals. It can also help devices such as smart speakers to improve the speech recognition performance by filtering out noise.

A basic approach to improve signal reception is the beamforming tech-

nique [99]. A physical beamformer is designed such that it forms a response pattern with higher sensitivity in desired directions using a sensors array [187]. This can be achieved by filtering the observation signals, such that the signals from desired directions experience constructive interference while others experience destructive interference. In a free-field (non-reverberant) scenario, the source separation problem can often be solved in a satisfactory manner with beamforming techniques. However, in a reverberant scenario (e.g., a classroom), the effective signal consists of contribution via the direct path and early reflections from arbitrary directions [23]. The early reflections can compensate for background noise and increase intelligibility [58]. Hence, if beamforming is used to capture the direct path signal, much useful information will be lost as the relative strength of the desired signal will be low. Moreover, the arrangement of the sensors must be known explicitly.

Blind source separation (BSS) is another family of approaches to solving the cocktail party problem. The aim of BSS is to separate the original source signals from the mixture without any knowledge of the source signals and the mixing process. It makes no assumption on spatial locations of the original sources. Thus, in a reverberant environment, the recovered sources may contain direct path components and also reflections. Since early reflections can compensate for background noise [58], signals with higher signal-to-noise-ratio (SNR) level can be recovered. Moreover, BSS relaxes the restriction of beamforming, which relies on knowledge of the array structure of the sensors and the location of the sources to extract interested sources. Hence, if BSS is used the sensors can be placed at arbitrary locations. These two factors make BSS a better and more space-saving solution than beamforming.

The topic of BSS has been widely studied over the last two decades. An enormous number of approaches were proposed for BSS problems [1, 20, 142, 198]. However, most methods are not applicable to practical real-world applications because of the system latency issue, especially for the BSS approaches that operate in the frequency domain. They generally require a window length that is longer than the mixing filter, so that the system model

is sufficiently accurate. This lengthens the algorithmic delay of a BSS system. Hence, these methods are limited to off-line computation.

In some situations, the environment may vary. The changes can be due to time-varying mixing channels, moving sources or dynamic source number. Hence, the separation parameters in the BSS system must be updated frequently to ensure good separation performance. Recently, a few methods were proposed to address the BSS problem for changing environments [128,169,175]. These methods often perform BSS from scratch without utilizing the “meta-information” that is available from previous operations. This affects the efficiency of a real-time system. Moreover, they also suffer from the problem caused by the algorithmic delay.

The overall goal of this thesis is to develop real-world BSS techniques that can be integrated with existing BSS approaches to perform source separation in real-time with minimal latency. The methods are beneficial to many real-time audio appliances, for instance, hearing aids, mobile phones, video calling, smart speaker and so on. Moreover, they must be efficient while operating in non-stationary environments. The methods reduce the estimation delay, so that the separation parameters can be updated promptly to guarantee good separation performance at all times.

1.2 Limitations of Conventional Approaches

In this section, we first provide a brief introduction of the blind source separation (BSS) technique and its potential applications. Next, we discuss the limitations of conventional BSS approaches while implementing in practical real-world applications.

Blind source separation aims to extract original sources from their mixtures without any prior knowledge of the mixing channels and the sources. The technique is particularly useful in a multi-talker acoustic environment as it is able to extract a targeted source signal and remove the other signals. For instance, the BSS algorithm can be applied in hearing instruments to help

users to concentrate on a particular talker. BSS can also be implemented in devices such as mobile phones and smart speakers, so that cleaner speech signals can be obtained. This can facilitate further operations, for instance, noise reduction and speech recognition.

In many applications, the mixing process is convolutive. This makes the problem more complicated as the mixing channels contain memory. To facilitate the estimation and improve the computational efficacy, the convolutive BSS problem is often transformed into the time-frequency (TF) domain [129]. If the windows used are sufficiently long, then the time-domain convolutive mixture becomes a set of instantaneous mixtures in the TF domain, which can then be solved by using an instantaneous BSS algorithm for every frequency bin, such as the conventional independent component analysis (ICA) approach [20, 36, 69].

Unfortunately, most existing TF-domain BSS techniques cannot be applied directly in a real-time system. The methods share a latency problem. We can decompose the delay into an algorithmic delay and an estimation delay.

The algorithmic delay is the delay inherent to the algorithm, which cannot be solved by devising programming or using faster processors. This makes it of great importance for real-time implementations. Most real-time systems operate on a block-by-block basis [111]. That is, they first capture a block of mixture samples using a short-time (ST) window, perform a discrete Fourier transform (DFT) and conduct the separation process in the frequency domain. They then transform the modified data into the time-domain, using an overlap-add technique [6] to synthesize the signals. Thus, an algorithmic delay of at least one ST window length is necessary.

To ensure that the instantaneous BSS model in the TF domain is sufficiently accurate, the length of the ST window has to be at least twice the length of the mixing filters [10, 146]. In the real world, the effective mixing filters are generally long. For instance, a regular classroom with a dimension of $9\text{ m} \times 6\text{ m} \times 2.7\text{ m}$ has a reverberation time of about 0.5 s [23]. Hence,

the algorithmic delay of the conventional BSS algorithms will be at least 1 s. This is not acceptable in the live-communication applications, such as hearing instruments, where the acceptable range for the delay is less than 30 ms [56]. For voice over internet protocol (VOIP) applications, the recommendation of the maximum delay for one-way transmission is 150 ms [51].

Figure 1.1: A two-threaded system for a blind source separation application

The estimation delay is the time taken for computing the demixing operators. The problem caused by the estimation delay can be reduced by employing a two-threaded system [121,122,162,172] as depicted in Fig. 1.1. The first thread is responsible for the separation process. This involves transforming input data into the frequency domain, multiplying the separation parameters and finally transforming the processed data into the time domain for rendering for the user. The second thread is responsible for the estimation process. In this thread, the separation parameters will be estimated after acquiring a sufficient amount of input data. By using a two-threaded system, the separation and estimation processes can be performed in parallel. The separation parameters are first initialized and then updated after every iteration of the

estimation process. During intervals between the estimation updates, the separated output signals are processed by the same separation parameters. Hence, the two-threaded system can only be useful in a stationary scenario or in an environment that is changing slightly.

The BSS performance during the intervals between the estimation processes in a two-threaded system can be affected significantly in a non-stationary environment. An example case where a new, relatively loud, source is introduced into the system, where the new source can either be stationary or moving. Most existing BSS approaches require a sufficient number of samples to perform source separation, so that satisfactory performance can be obtained, such as the conventional ICA method. This prolongs the estimation period and degrades the user experience as the source signals cannot be recovered well until the new separation parameters are updated.

This thesis aims to solve the latency issue caused by the conventional frequency-domain BSS approaches, so that they will be applicable for real-time applications. The details of the contribution of the thesis are discussed in the next section.

1.3 Contribution of the Thesis

The overall structure of the thesis contribution is illustrated in Fig. 1.2. We focus on the latency problem of the BSS approaches, which is the major limitation factor that makes most of the existing BSS approaches inapplicable in real-time systems. Our aim is to design a BSS system that can operate in real-time with minimal latency.

In the thesis, we distinguish the system latency into two types, which are the algorithmic latency and the estimation delay. In a two-threaded BSS system, the algorithmic latency is related to the separation process in the first thread while the estimation delay is related to the estimation process in the second thread. We want to reduce the algorithmic delay of our BSS system should, so that it falls within the acceptable range of live-

Figure 1.2: Overall structure of the thesis contribution.

communication applications, i.e., below 30 ms [56]. We also want to minimize the computational cost of the BSS algorithm to reduce the estimation delay, so that the separation parameters can be updated promptly. We now briefly discuss our contributions for each of the latency problems.

1.3.1 Algorithmic latency

We propose two methods to reduce the algorithmic delay of the BSS system. The first approach is presented in Chapter 3. The approach directly operates in the low-frequency-resolution time-frequency (LFRTF) domain. In the LFRTF, the window length is shorter than the room impulse response. The proposed approach reformulates the convolutive problem into a multiplicative one and utilizes a subspace method to identify the subspace of the solutions. The final separation parameters can then be estimated using the existing BSS approaches. However, the computational cost of this approach may be high due to long mixing channels and an increasing number of observations. The second approach, which is presented in Chapter 4, avoids this

problem and it is more applicable to real-world systems. It first performs the BSS operation using existing BSS methods in the high-frequency-resolution time-frequency (HFRTF), where the window length is longer than the room impulse response, so that the instantaneous model is sufficiently accurate. Then it designs the separation parameters in the LFRTF, so the separation procedure can be performed with minimal latency.

1.3.2 Estimation delay

In Chapter 5, we also consider the non-stationary scenario, where a new source is introduced into the system. By using the “meta-information” that is preserved from the previous BSS operations, our BSS system can adapt to the changing environment as soon as possible to find the optimal solution for separating the new source. Thus, the system can be updated promptly without re-performing the BSS algorithm from scratch.

Furthermore, we propose a non-iterative method for impulse response shortening in Chapter 6. This method enables us to approximate the long impulse responses with shorter representatives to reduce the computational cost. We exploit that the scaling ambiguity has a minimal effect on the BSS separation performance. Thus, we apply scaling factors to smoothen the frequency response and reduce the length of the impulse response.

1.4 Organization of the Thesis

The remainder of the thesis is organized as follows. Chapter 2 provides the literature review of the existing BSS approaches and the related work. The main contributions of the thesis are presented in Chapter 3-6. Chapter 7 draws conclusions and discusses possible future works. We now present a brief overview of each chapter:

Chapter 2: This chapter provides a literature review and background of the thesis. We first discuss the system model of a cocktail party problem.

Next, we study various techniques of the existing BSS and beamforming algorithms and discuss the advantages and the disadvantages of each method. The main problems of the TF-domain BSS approaches, which are so-called permutation and scaling ambiguities, are also discussed. Finally, we provide a brief summary of the existing approaches and discuss the important aspects that this thesis aims to solve.

Chapter 3: In this chapter, we provide a method to directly estimate the separation parameters in the low-frequency-resolution time-frequency (LFRTF) domain. This allows a short window length to capture signal samples, and hence the system latency can be reduced. The window length can be shorter than the impulse response. The mixing model remains a convolutive model in the LFRTF domain but with a shorter memory. This makes the problem tractable and facilitates the estimation procedure for finding the separation parameters in the BSS system.

Chapter 4: In this chapter, we provide a post-processing method that facilitates the reduction of the algorithmic latency of a system by using existing TF-domain BSS algorithms. This method allows the estimating process to be performed in the high-frequency-resolution time-frequency (HFRTF) domain while the separation is performed in the LFRTF domain. In the HFRTF domain, the window length is longer than the impulse response, which makes the instantaneous mixing model in the frequency domain sufficiently accurate. Hence, the mixing or demixing parameters can be estimated using existing BSS algorithms. We then compute the separation parameters in the LFRTF domain using the parameters that are estimated in the HFRTF domain.

Chapter 5: This chapter aims to solve the BSS problem in a non-stationary scenario, especially for the case when a new, relatively loud, source is introduced into the system. The proposed method allows the system to update the separation parameters promptly by using “meta-information” that is preserved from previous BSS operations. By using information from the entire spectrum, we can update the separation parameters within only one block

of data. This minimizes the estimation delay and improves the system efficacy. We can also avoid redundant operations, such as repeating the BSS algorithm, which is often time-consuming.

Chapter 6: In this chapter, we provide a non-iterative method for impulse response shortening. The idea originates from the fact that zero-padding a time-domain signal makes the spectrum of the signal smoother in the frequency domain. Moreover, the separation performance is not affected significantly by a scaling of the frequency bins. Thus, we are able to apply optimized complex scaling factors to the frequency-domain mixing channels, so that the spectrum becomes smoother across frequencies, which implies that shorter responses are obtained. The computational complexity can be reduced by using shorter representatives of the impulse responses, and hence, the estimation delay can be decreased.

Chapter 7: This chapter summarizes the thesis and draws general conclusions of the research topic. Possible future research directions are also discussed.

Chapter 2

Background and Literature Review

In this chapter, we provide a literature review and background on the topic of blind source separation (BSS), which is a common technique used to address the cocktail party problem. First, we discuss possible applications that the BSS technique can be implemented. Second, we discuss a BSS model that is used for acoustic signals. Third, we study various existing approaches for BSS and discuss their basic principles. This is followed by a discussion about the permutation and scaling ambiguities that presented in the BSS approaches, which are operating in the time-frequency domain. Then, we study the existing approaches to deal with this problem. We also provide a brief overview of the beamforming techniques. Finally, the chapter concludes with a brief summary about the drawbacks of the studied approaches.

2.1 Possible BSS Applications

Blind source separation (BSS) aims to simultaneously separate mixtures into each independent signal of interest. It has been widely used in many different fields. These applications include but are not limited to image [88, 117], biomedical [74, 77, 115], Sonar [38, 44], seismic [103, 173], telecommunications

[31, 179], and audio [106, 190] applications. In the thesis, we are interested in acoustics applications, and hence in the remaining of the chapter all the discussions and the derivations are based on acoustics implementations.

In audio applications the BSS technique can be implemented in many types of devices, such as hearing aids, mobile phones and smart speakers. It enables users to focus on the source that they are only interested in. BSS systems are particularly useful in a noisy environment, where a high level of concentration is needed.

BSS systems can also be implemented in a classroom setting to improve students learning by helping them to stay focused in class. This can be done by placing multiple microphones at different arbitrary locations in a classroom. When people talk, the speech signals are captured by the microphones and transmitted to a central processor. Then, BSS is performed on the observation signals by the processor and the separated sources are delivered to the hearing devices such as hearing aids and earphones.

In many real-world applications, a low latency is required for BSS; this is particularly true for hearing aids to keep the algorithmic latency between inputs and outputs as short as possible. Large delay can affect the user experience due to the loss of lip synchronization [4]. The details of proposed methods for addressing the algorithmic latency in existing BSS methods are discussed in Chapter 3 and 4.

In some circumstances, users want to focus on more than one audio stream. For instance, a new loud source may suddenly appears. The new source can be a moving bus on a street or a walking waiter in a cafeteria. The BSS strategy can help to decompose raw mixtures into independent original sources and the desired sources are then recombined into a meaningful mixture. In this case, the BSS system must be able to adapt promptly. To address this problem, we propose an approach that allows separation parameters in the BSS system to be updated within a single snapshot of observation data. The details will be discussed in Chapter 5. We also provide an approach to reduce estimation latency of the BSS system in Chapter 6.

We have briefly discussed the potential BSS implementations for acoustics applications. In the next section, we provide a problem formulation of the BSS system model.

2.2 System Model

In this section, we discuss a convolutive BSS model that can be used for acoustic signals. In the derivation, we assume a linear time-invariant (LTI) system in a noiseless scenario.

Denote by M the number of sensors and L the number of original sources. Let $\mathbf{A}_\tau$ denote a matrix of the unknown mixing filters or the channel impulse responses at time τ , for instance, $\mathbf{A}_\tau^{ml}$ (the entry of $\mathbf{A}_\tau$ in row m and column l) indicates the coefficient of the impulse response between observation m and source l at time τ . The mixing of the original signals in a LTI system without noise can be represented as [139]

$$\mathbf{x}_t = \sum_{\tau=0}^{N_h-1} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \in \mathbb{R}^{M \times 1}, \quad (2.1)$$

in discrete form, where $\mathbf{x}_t = [x_t^1 \ \cdots \ x_t^M]^\top$ is a vector of the observation signals, $\mathbf{s}_t = [s_t^1 \ \cdots \ s_t^L]^\top$ is a vector of the original sources and N_h is the length of the impulse response. If $M > L$ the problem is overdetermined, it is determined if $M = L$ and it is underdetermined if $M < L$ [7]. In this research, we will focus on the first scenario, i.e., $M > L$.

Consider a block consists of K samples instead of individual samples at time t , where $K > N_h$, Eq. (2.1) can be expressed as

$$\mathbf{X}_t = \mathbf{A} \mathbf{S}_t, \quad (2.2)$$

where

$$\mathbf{X}_t = [\mathbf{x}_t^\top \ \mathbf{x}_{t-1}^\top \ \cdots \ \mathbf{x}_{t-K+1}^\top]^\top \in \mathbb{R}^{MK \times 1}, \quad (2.3)$$

$$\mathbf{S}_t = [\mathbf{s}_t^\top \ \mathbf{s}_{t-1}^\top \ \cdots \ \mathbf{s}_{t-K-N_h+1}^\top]^\top \in \mathbb{R}^{L(K+N_h-1) \times 1}, \quad (2.4)$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_0 & \cdots & \mathbf{A}_{N_h-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_0 & \cdots & \mathbf{A}_{N_h-1} \end{bmatrix} \in \mathbb{R}^{MK \times L(K+N_h-1)}. \quad (2.5)$$

Note that $\mathbf{A}$ has a block-Toeplitz structure [24] as the mixing filters are not varying in a LTI system.

To separate the original sources from the mixtures, we try to find a demixing matrix $\mathbf{W}$ in

$$\mathbf{Y}_t = \mathbf{W}^T \mathbf{X}_t \in \mathbb{R}^{L(K+N_h-1) \times 1}, \quad (2.6)$$

where $\mathbf{Y}_t$ is an estimate of the original sources $\mathbf{S}_t$. However, this requires operations with huge Toeplitz matrices, which is time-consuming and computationally expensive. Therefore, it is often transformed into the time-frequency (TF) domain using a short-time Fourier Transform (STFT) [48].

In this thesis, we use two types of TF-domain representations to describe the time-domain signals. They are high-frequency-resolution time-frequency domain (HFRTF) and low-frequency-resolution time-frequency domain (LFRTF). We now provide the details and the mathematical definitions of these two TF-domain representations.

2.2.1 The HFRTF-Domain Representation

By the Convolution Theorem [39], a circular convolutive process in the time domain is equivalent to a multiplicative process in the frequency domain. The circular convolution describes the linear convolution accurately after zero-padding [133]. Without zero-padding, there is a discrepancy in the approximation of a linear convolution by using a circular convolution due to sampling the continuous frequency representation of a finite length signal [152]. This artifact is known as a circularity effect. In the TF domain, the circularity effect can be minimized by using a window that is longer than the mixing filters, e.g., $N_w > 2N_h$ [10, 145], where N_w indicates the window length, while performing the STFT. We refer this TF domain as high-frequency-resolution time-frequency (HFRTF) domain.

Let p indicates the time-block index and k indicates the frequency bin index. In the HFRTF domain, by using an appropriate window length, i.e., $N_w > 2N_h$, the observations $\mathbf{x}_{p,k}$ can be approximated as [29]

$$\mathbf{x}_{p,k} \approx \mathbf{A}_k \mathbf{s}_{p,k}, \quad (2.7)$$

where $\mathbf{A}_k \in \mathbb{C}^{M \times L}$ is the mixing matrix in frequency bin index k and $\mathbf{s}_{p,k}$ indicates the original sources in the TF domain. Hence, the observation signals can be approximated by a series of instantaneous BSS models in every frequency bin k , which simplifies the estimation process. The original sources in every frequency bin index k can then be separated by using the demixing matrix $\mathbf{W}_k$, e.g.,

$$\mathbf{y}_{p,k} = \mathbf{W}_k^H \mathbf{x}_{p,k}, \quad (2.8)$$

where $\cdot^H$ denotes a Hermitian transpose.

2.2.2 The LFRTF-Domain Representation

Recall that the algorithmic latency of the real-time system is related to the window length. To reduce the system latency, the window length has to be kept as short as possible, e.g. less than 30 ms. However, the reverberation time of the common room is generally larger than 30 ms. For instance, a regular classroom has a reverberation time of about 0.5 s [23]. Hence, the instantaneous BSS model is no longer valid. By using a short window, which is generally shorter than the room impulse responses, the mixing model in the TF domain is still convolutive but with a shorter memory samples. We refer this TF domain as low-frequency-resolution time-frequency (LFRTF) domain.

Let denote by p and by p' the time-block index of the signal and the mixing filter, respectively, in the LFRTF domain. Let k represents the frequency bin index and k' represents the neighbour frequency bin index. The observation

$\mathbf{x}_{p,k}$ in the LFRTF domain can be expressed as [14]:

$$\mathbf{x}_{p,k} = \sum_{p'=0}^{T-1} \sum_{k'=0}^{N_w-1} \mathbf{A}_{p',k,k'} \mathbf{s}_{p-p',k'}, \quad (2.9)$$

where T is the memory length of the mixing filters in the LFRTF domain, N_w is the length of the window, $\mathbf{s}_{p,k}$ denotes the original sources at TF block index (p, k) in the LFRTF domain, while $\mathbf{A}_{p',k,k'}$ for $k = k'$ indicates a band-to-band mixing matrix and for $k \neq k'$ indicates a crossband mixing matrix at memory block index p' . The details for the computation of the mixing filters $\mathbf{A}_{p',k,k'}$ are described in Sec. 4.3.3.

Generally, it is difficult to directly estimate the demixing parameters for the blind source separation problem in the LFRTF domain. Instead, we can perform the estimation in the HFRTF domain since many HFRTF-domain BSS methods have been proposed in last two decades. Then, we compute the LFRTF-domain demixing parameters based on the HFRTF-domain estimation. The details will be discussed in Chapter 4.

In the next section, we investigate different existing HFRTF-domain BSS methods that aim to find either the mixing matrix $\mathbf{A}_k$ or the demixing matrix $\mathbf{W}_k$. For the BSS methods that aim to find the mixing matrix $\mathbf{A}_k$, it requires an extra operation to estimate the demixing matrix $\mathbf{W}_k$. The straightforward solution, which for the overdetermined case that we consider, is to compute the pseudo-inverse of the mixing matrix $\mathbf{A}_k$:

$$\mathbf{W}_k = \mathbf{A}_k (\mathbf{A}_k^H \mathbf{A}_k)^{-1}. \quad (2.10)$$

An alternative solution is to find each demixing vector $\mathbf{w}_k^l$ ($\mathbf{w}_k^l$ is the l^{th} column of $\mathbf{W}_k$) independently, where the demixing vector $\mathbf{w}_k^l$ lives in the null space of the subspace that is spanned by the mixing vectors of the other sources. The details of the method can be found in Sec. 4.3.2 and Sec. 4.3.3. The separated signals can finally be recovered using (2.8).

2.3 Existing Blind Source Separation Approaches

Over the last two decades, many techniques have been developed to address the BSS problem in the HFRTF domain. An overview of approaches can be found in [71, 142]. These methods consider different aspects of solving the BSS problem. These attributes include time-invariant filters, statistical properties of original sources and activity periods of original sources. Most BSS approaches can be categorized into two main categories: (i) Category 1 intends to find the demixing parameters and (ii) Category 2 aims to estimate the mixing channels.

In following subsections, we investigate various existing BSS approaches that are suitable for our applications. Then, we discuss their basic principles.

2.3.1 Independent Component Analysis

The independent component analysis (ICA) technique falls into Category 1. It aims to identify independent components from a linear time-invariant mixture without any prior knowledge of the mixing system and the original sources by utilizing the statistical properties of the original sources. It is higher-order statistics based as it assumes that all the original sources are independent to each other. Thus, different independence criteria can be used as a cost function for optimization to estimate the original sources, e.g., means of higher order moments or cumulants (fourth order) [72, 163, 197] kurtosis [70, 147, 170, 177] and mutual information [8, 16, 20, 68].

The ICA was first introduced by Jutten in 1987 [79], then it was formalized for linear mixtures in 1991 by Comon [34, 35]. A family of new contrast functions for ICA was developed in [67], the computation speed and the robustness were then further improved by the approach proposed in [68], which is known as Fast-ICA. A version of Fast-ICA to handle complex signals can be found in [20]. With the progress of times, more and more ICA

approaches have been proposed by using different criteria as contrast functions for estimating independent sources [36, 71]. A refinement of the ICA approach, which is known as independent vector analysis (IVA), has been proposed [83, 84] to avoid permutation issues, which will be discussed in Sec. 2.4 in the time-frequency domain. It considers multivariate components, for instance, all the frequency bins, while computing the cost function. The IVA method is further improved by [130–132] to have a faster convergence rate.

Due to the simplicity and efficiency of the complex Fast-ICA algorithm [20], it is employed as our initial separation method and also used as a comparison method in Chapter 3, 4 and 6.

We provide a brief overview of the Fast-ICA approach [20]. By the Central Limit Theorem, the mixture will follow Gaussian distribution even though the distribution of each original source is non-Gaussian [127]. Thus, a demixing matrix can be estimated by maximizing the non-Gaussianity of every estimated source. Hence, negentropy can be used a contrast function for the optimization problem. Negentropy is a measure of the distance for a given distribution to normality. It is vanished if the signal is Gaussian distributed.

Denote by $\mathbf{w}_k^l$ the demixing vector for the l^{th} original source, i.e., the l^{th} column vector in the demixing matrix $\mathbf{W}_k$ in frequency bin index k . To prevent estimating the same source in every iteration, the Fast-ICA algorithm always makes $\mathbf{w}_k^l$ to be orthogonal to $\mathbf{w}_k^i$, where $i = 1, \dots, l-1$. This is done by projecting the final estimate of $\mathbf{w}_k^l$ into the null space of the subspace that is spanned by $\{\mathbf{w}_k^1, \dots, \mathbf{w}_k^{l-1}\}$. Hence, the estimate of the demixing matrix $\mathbf{W}_k$ will be an orthogonal matrix by using the Fast-ICA algorithm. However, in practice the mixing matrix $\mathbf{A}_k$ is not necessarily an orthogonal matrix.

To implement Fast-ICA algorithm with real-world data, a pre-processing technique the so-called whitening process must first be applied to the observation data. This transformation makes the processed data have an identity covariance matrix by minimizing the correlation between principal components [81]. Hence, the mixing matrix of the processed data is an orthogonal

Algorithm 1 Pseudocode of the Fast-ICA algorithm.

Input: $\mathbf{x}_{p,k}$ $\triangleright (p, k)$ denotes TF block index in the TF domain.

Output: $\mathbf{y}_{p,k}, \mathbf{W}_k$

```

1: for all  $k$  do
2:   Compute the whitening matrix  $\mathbf{Q}_k$  based on  $\mathbf{x}_{p,k}$ .
3:   Computed the whitened data  $\mathbf{x}_{p,k}^Q$ .
4:   for all  $l \in [1, L]$  do  $\triangleright L$  indicates a number of original sources.
5:     Estimate  $\mathbf{w}_k^l$  by maximizing the negentropy using the data  $\mathbf{x}_{p,k}^Q$ .
6:     Make  $\mathbf{w}_k^l$  to be orthogonal to  $\mathbf{w}_k^i$ , where  $i = 1, \dots, l - 1$ .
7:   end for
8:   Perform dewhitening process on  $\mathbf{W}_k$  using  $\mathbf{Q}_k$ .
9:   Compute the estimate the original sources  $\mathbf{y}_{p,k}$  by using (2.8).
10: end for

```

matrix. The pseudocode of the Fast-ICA algorithm is provided in Algorithm 1.

ICA is a probabilistic approach, which means a sufficient number of samples is needed to identify statistically independent sources with sufficient accuracy. Moreover, ICA uses high-order statistics, hence it requires more number of samples than methods using second order statistics to make the statistics sufficient. The BSS system with the ICA approach can be useful for static sources or environments that slightly vary but not the non-stationary scenarios. If the environment is changing, e.g., a new source is added into the system, the BSS system has to collect a sufficient number of samples to re-estimate the separation parameters as the previous estimated parameters are no longer useful. This implies a BSS system with long adaptation period is undesirable.

2.3.2 Joint Diagonalization

Wu and Principe [194] proposed a joint diagonalization method, which belongs to Category 1, for performing blind source separation by assuming that original sources are uncorrelated and stationary over time. For instance, all the sources are active and not moving during the estimation period. Then, the demixing matrix can be estimated by simultaneously minimizing the cross-correlation between the separated signals that are obtained in two time-segments during the estimation period. This method is then improved by [139] through minimizing multiple cross-correlations simultaneously in the time-frequency (TF) domain. Note that the joint diagonalization method is second-order statistics based. Therefore, it is less sensitive to noise and outliers [181], and hence fewer samples are required for analysis compared to ICA, which is higher-order statistics based.

We assume that the original source $\mathbf{s}_{p,k}$ has zero mean. The second-order statistics of the estimated sources among N_q blocks in frequency bin k can be represented by

$$\begin{aligned}\mathbf{R}_{\mathbf{y}\mathbf{y},k}(p) &= \frac{1}{N_q} \sum_{q=p-N_q+1}^p \mathbf{y}_{q,k} \mathbf{y}_{q,k}^H \\ &= \mathbf{W}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k}(p) \mathbf{W}_k,\end{aligned}\tag{2.11}$$

where $\mathbf{R}_{(\cdot)}$ is the covariance matrix. If all the estimated sources are uncorrelated, the off-diagonal entries in $\mathbf{R}_{\mathbf{y}\mathbf{y},k}(p)$ should vanish.

The optimum $\mathbf{W}_k^*$ can be estimated by simultaneous minimizing the off-diagonal entries in $\mathbf{R}_{\mathbf{y}\mathbf{y},k}(p)$ for some p :

$$\mathbf{W}_k^* = \arg \min_{\mathbf{W}_k} \sum_p \|\mathbf{R}_{\mathbf{y}\mathbf{y},k}(p) - \mathbf{\Lambda}_{\mathbf{y}}(p)\|_{\mathcal{F}}^2,\tag{2.12}$$

where $\mathbf{\Lambda}_{\mathbf{y}}(p)$ is the diagonal matrix of $\mathbf{R}_{\mathbf{y}\mathbf{y},k}(p)$ and $\|\cdot\|_{\mathcal{F}}$ is a Frobenius norm. To avoid a trivial estimate with all zero elements in $\mathbf{W}_k$, a unity norm constraint is often added, e.g., $\|\mathbf{w}_k^l\| = 1, l \in \{1, \dots, L\}$.

Similar to ICA approaches, the joint diagonalization method is also a probabilistic approach. A sufficient number of samples is still needed to ensure the separation performance. This affects the adaptation time of the BSS system due to non-stationary scenarios as the previous estimated separation parameters are no longer useful.

2.3.3 Generalized Eigenvalue Decomposition

In 2004, Parra and Sajda [138] first proposed the source separation can be carried out by means of a generalized eigenvalue decomposition (GEVD) of two covariance matrices. However, this method does not explain how to choose practically these two covariance matrices. This was then carried on by [149, 150] to extract intermittent sources by estimating null spaces of the active sources. This method also belongs to the Category 1 as the null spaces are equivalent to the demixing vectors. It utilizes both the statistical properties and the activity periods of the original sources. We provide a brief discussion on the GEVD method in this subsection.

Consider a linear time-invariant instantaneous mixing model at time-block index p and in frequency bin index k with the same number of observations $\mathbf{x}_{p,k}$ and original sources $\mathbf{s}_{p,k}$, such as $M = L$. We assume that the original sources have zero mean. Then, the covariance matrix of all the samples in the mixtures $\mathbf{x}_{p,k}$ at frequency bin index k can be written as

$$\mathbf{R}_{\mathbf{xx},k} = \mathbf{A}_k \mathbf{R}_{\mathbf{ss},k} \mathbf{A}_k^H \approx \sum_{l=1}^L \sigma_k^{l^2} \mathbf{a}_k^l \mathbf{a}_k^{lH}, \quad (2.13)$$

given that all the sources are uncorrelated to each other, where $\sigma_k^{l^2}$ is the power spectral density of the l^{th} source and $\mathbf{a}_k^l$ is the l^{th} column of the mixing matrix $\mathbf{A}_k$, i.e., $\mathbf{A}_k = \begin{bmatrix} \mathbf{a}_k^1 & \cdots & \mathbf{a}_k^L \end{bmatrix}$.

To find the vectors $\mathbf{a}_k^l$, we first assume the first L_1 sources are silent at time interval τ without losing generality (e.g., $\mathbf{s}_{p,k}^l = 0$, where $1 \leq l \leq L_1$ and $p \in \tau$). Hence, the covariance matrix of the mixtures during the time

interval τ can be written as

$$\mathbf{R}_{\mathbf{xx},k}(\tau) = \mathbf{A}_k \mathbf{R}_{\mathbf{ss},k}(\tau) \mathbf{A}_k^H \approx \sum_{l=L_1+1}^L \sigma_k^l(\tau)^2 \mathbf{a}_k^l \mathbf{a}_k^{lH}. \quad (2.14)$$

Provided that $\mathbf{W}_k$ is the demixing matrix ($\mathbf{W}_k^H \mathbf{A}_k = \mathbf{A}_k^H \mathbf{W}_k = \mathbf{I}$), if we multiply Eq. (2.14) with $\mathbf{W}_k$, this yields

$$\begin{aligned} \mathbf{R}_{\mathbf{xx},k}(\tau) \mathbf{W}_k &= \mathbf{A}_k \mathbf{R}_{\mathbf{ss},k}(\tau) \mathbf{A}_k^H \mathbf{W}_k = \mathbf{A}_k \mathbf{R}_{\mathbf{ss},k} \mathbf{R}_{\mathbf{ss},k}^{-1} \mathbf{R}_{\mathbf{ss},k}(\tau) \\ &= \mathbf{A}_k \mathbf{W}_k^H \mathbf{R}_{\mathbf{xx},k} \mathbf{W}_k \mathbf{R}_{\mathbf{ss},k}^{-1} \mathbf{R}_{\mathbf{ss},k}(\tau) = \mathbf{R}_{\mathbf{xx},k} \mathbf{W}_k \mathbf{\Lambda}_k(\tau), \end{aligned} \quad (2.15)$$

where $\mathbf{\Lambda}_k(\tau) = \mathbf{R}_{\mathbf{ss},k}^{-1} \mathbf{R}_{\mathbf{ss},k}(\tau)$, which is a diagonal matrix due to the independence of the sources. Note that $\mathbf{R}_{\mathbf{xx},k}(\tau)$ and $\mathbf{R}_{\mathbf{xx},k}$ represent the overall covariance matrix and the covariance matrix during time interval τ , respectively, at frequency bin index k .

Eq. (2.15) constitutes a generalized eigenvalue equation. The eigenvalues, which are the diagonal entries of $\mathbf{\Lambda}_k(\tau)$, represent the ratio of the individual source statistics measured by the diagonals of $\mathbf{R}_{\mathbf{ss},k}$ and $\mathbf{R}_{\mathbf{ss},k}(\tau)$ at time interval τ . Since the first L_1 sources are silent at time interval τ , the first L_1 diagonal entries in $\mathbf{\Lambda}_k(\tau)$ will be zero. The corresponding eigenvectors, which are the first L_1 columns of $\mathbf{W}_k$, define a demixing subspace $\varepsilon_{0,k}$ that is orthogonal to the subspace that is spanned by the mixing vectors of the other original sources, e.g., $\{\mathbf{a}_k^{L_1+1}, \dots, \mathbf{a}_k^L\}$. Therefore, if a generalized eigenvalue decomposition of the pair $(\mathbf{R}_{\mathbf{xx},k}(\tau), \mathbf{R}_{\mathbf{xx},k})$ is performed, this yields two distinct eigenvalues. They are either zeros, where the associated generalized eigenvectors lie in the demixing subspace $\varepsilon_{0,k}$, or nonzero, where the associated generalized eigenvectors span a mixing subspace $\varepsilon_{1,k}$ that is orthogonal to the demixing subspace $\varepsilon_{0,k}$ in the M -dimensional space [151].

By performing a GEVD on $(\mathbf{R}_{\mathbf{xx},k}(\tau), \mathbf{R}_{\mathbf{xx},k})$, we are able to detect if the sources $\{\mathbf{s}_{p,k}^1, \dots, \mathbf{s}_{p,k}^{L_1}\}$ vanish. We can also estimate the demixing subspace $\varepsilon_{0,k}$, which is the null space of the mixing subspace $\varepsilon_{1,k}$ that is spanned by the mixing vectors of the other active sources $\{\mathbf{s}_{p,k}^{L_1+1}, \dots, \mathbf{s}_{p,k}^L\}$, e.g., $\{\mathbf{a}_k^{L_1+1}, \dots, \mathbf{a}_k^L\}$. When $\{\mathbf{s}_{p,k}^1, \dots, \mathbf{s}_{p,k}^{L_1}\}$ are active, where $p \notin \tau$, they can

be extracted by projecting the observations $\mathbf{x}_{p,k}$ onto $\varepsilon_{0,k}$. Note that the extracted sources will still be a combination mixture of the intermittent sources $\{\mathbf{s}_{p,k}^1, \dots, \mathbf{s}_{p,k}^{L_1}\}$. To separate all the sources, we have to assume that every source is inactive during a certain time interval within the estimation period. More details of the GEVD method can be found in [149, 150].

The GEVD method [149, 150] requires every source or different combination of sources to be inactive during certain time intervals to compute the separation parameters. This requirement can be hard to be fulfilled in some circumstances, for instance, in a noisy cafeteria. Furthermore, selecting an appropriate duration to acquire sufficient signal samples can also be a difficult task.

2.3.4 Blind Channel Identification

Instead of finding the demixing filter, blind channel identification (BCI) aims to identify the channel impulse responses (CIR) of the system without any prior knowledge of the input signals by using only observation signals. Hence, it belongs to the approaches in Category 2. The BCI strategy relies only on the assumption that the channels are time-invariant during analysis periods. This method does not suffer from the amplitude and scaling ambiguities. The original sources can be recovered by directly inverting the estimated CIR using the multiple input/output inverse theorem (MINT) [119].

The innovative idea of self-recovering (blind) adaptive equalization was first proposed by Sato [153], then further developed by Godard [50], Treichler and Agee [182], Benveniste and Goursat [18] as well as Shalvi and Weinstein [161]. These works are based on optimization criteria involving high-order statistics (cumulants) of the observation. This leads to slow convergence and many local optima attractors [179]. In 1994, Tong et al. proposed that the communication channel can be blindly identified based on second-order statistics [179]. These works are designed particularly for the telecommunication signals but not for acoustic signals [43, 63].

Figure 2.1: The block diagram of a SIMO system [62].

Lin, Gaubitch and Naylor [98] proposed a method to perform blind channel identification (BCI) of single-input-multiple-output (SIMO) systems with a common zero. The multiple-input-multiple-output (MIMO) version of the blind channel identification was done in [65]. The logic is to divide the MIMO system into multiple SIMO systems. The SIMO model solution can then be used to solve the MIMO problem. According to [64], a multi-channel system can be blindly identified mainly because of the channel diversity. For example, if all channels are identical, the system reduces to a single-channel case and the individual channels are, therefore, unidentifiable. In this section, we assume this assumption holds when discussing the BCI problem formulation of a SIMO system and the principle of the method.

The block diagram of a SIMO system is illustrated in Fig. 2.1. In the time domain, the i^{th} observation $x_i(n)$ is the result of a linear convolution between the source signal $s(n)$ and the corresponding CIR h_i , corrupted by an additive noise $b_i(n)$:

$$x_i(n) = h_i(n) * s(n) + b_i(n), \quad i = 1, 2, \dots, M, \quad (2.16)$$

where M is the number of channels.

When the input signal is unknown, the autocorrelation between the sensor outputs can be exploited to estimate the CIRs. This is because

$$x_i(n)*h_j(n) = s(n)*h_i(n)*h_j(n) = x_j(n)*h_i(n), \quad i, j = 1, 2, \dots, M, \quad i \neq j, \quad (2.17)$$

if noise is absent. Rewriting (2.17) in vector form,

$$\mathbf{x}_i^T(n)\mathbf{h}_j = \mathbf{x}_j^T(n)\mathbf{h}_i, \quad i, j = 1, 2, \dots, M, \quad i \neq j, \quad (2.18)$$

where

$$\mathbf{x}_i(n) = [x_i(n) \quad x_i(n-1) \quad \dots \quad x_i(n-N_h+1)]^T, \quad (2.19)$$

$$\mathbf{h}_i = [h_i(0) \quad h_i(1) \quad \dots \quad h_i(N_h-1)]^T, \quad (2.20)$$

and N_h is the length of the CIR. An error signal will be produced when we take the difference between the left-hand side and the right-hand side of (2.18). The error signal based on the i^{th} and j^{th} observations at time n can be written as

$$e_{ij}(n) = \begin{cases} \mathbf{x}_i^T(n)\mathbf{h}_j - \mathbf{x}_j^T(n)\mathbf{h}_i, & i \neq j \quad i, j = 1, 2, \dots, M, \\ 0, & i = j \quad i, j = 1, 2, \dots, M. \end{cases} \quad (2.21)$$

We have $(M-1)M/2$ distinct error signals $e_{ij}(n)$, which excludes the case $e_{ii}(n) = 0$ and count the $e_{ij}(n) = -e_{ji}(n)$ pair only once. Assuming that these error signals are equally important, summing them together defines a cost function as

$$J(n) = \sum_{i=1}^{M-1} \sum_{j=i+1}^M \frac{e_{ij}^2(n)}{\|\mathbf{h}\|^2}, \quad (2.22)$$

where $\mathbf{h} = [\mathbf{h}_1^T \quad \dots \quad \mathbf{h}_M^T]^T$ is a concatenated representative of the M channel impulse responses and set $\|\mathbf{h}\| = 1$ as a constraint to avoid a trivial estimate with all zero elements in $\mathbf{h}$. Therefore, the desired solution for $\mathbf{h}$ can be determined by minimizing the cost function $J(n)$ which is defined in (2.22), such as

$$\mathbf{h}^* = \arg \min_{\mathbf{h}} E \{J(n)\}, \quad (2.23)$$

where $E\{\cdot\}$ denotes an expectation operator.

The basic principle tells us how to estimate the channel impulse responses in a SIMO system. The BCI approach is commonly used in telecommunication area. However, we are interested in acoustic channels in a reverberant environment. In practice, these channels are more correlated than the telecommunication channels due to reverberation, thus making them difficult to identify [64]. Moreover, the computational cost is extremely high for a SIMO system, and more so for a MIMO system.

2.3.5 Sparsity in the Time-Frequency Domain

The mixture of speech signals is often “sparse” in the time-frequency (TF) domain. In this context, “sparse” means most of the time-frequency bins in the spectrum will contain only one active talker. By successfully detecting the only one active source, we can then identify the corresponding mixing vector. Hence, this approach usually makes no assumption on the statistical distribution of the original sources, unlike the ICA approaches. The approach belongs to Category 2 as it aims to find the mixing parameters by utilizing the activity period of the original sources.

In 2004, Yilmaz and Rickard [198] proposed a degenerate unmixing estimation technique (DUET) algorithm to solve the BSS problem by using time-frequency maskings based on the sparsity of the mixture. A mask is applied in the TF domain to capture the TF points, where only one source is active while the other TF points are discarded. Attenuation factor and delay parameter of the active source can then be estimated by using the masked TF points. A two-dimensional histogram is plotted by using the attenuation factor and delay parameter. The number of peaks in the histogram represents the number of sources. The mixing parameters can then be estimated by using the peak locations. By knowing the mixing parameters and source number, the original sources can be recovered.

In 2005, Abrard and Deville [2] proposed another sparsity-based BSS

approach, which is called the time-frequency ratio of mixtures (TIFROM) algorithm. The essence of this approach is that if the same source is active during a time interval τ , the relative transfer functions are the same within those TF blocks, i.e.,

$$\frac{\mathbf{x}_{p,k}}{x_{p,k}^1} = \frac{\mathbf{x}_{q,k}}{x_{q,k}^1}, \quad \forall p, q \in \tau, \quad (2.24)$$

where $\mathbf{x}_{p,k}$ and $x_{p,k}^m$ represent the observation vector and the observation signal in the m^{th} microphone, respectively, in TF block index (p, k) . The advantage of this approach is that it requires every source to be active individually within a few adjacent TF blocks, then they may simultaneously active in the remainder TF blocks during the analysis period.

The DUET and TIFROM approaches require the mixture to be sparse in the time-frequency domain, which is generally valid for speech signals in non-reverberant environments. This is not a statistically based approach, hence sample-collecting time can be reduced by selecting appropriate time intervals. However, in a reverberant scenario the performance deteriorates [21].

2.3.6 Clustering

An alternative approach to estimating the mixing matrix from the “sparse” observations is a clustering approach, which is done in [11]. This method is more relaxed to the sparsity assumption compared to DUET [198] and TIFROM [1]. It does not require other signals to vanish while the active signal is on. The clustering approach is one of the strategies in Category 2, which requires time-invariant channels and depends slightly on the activity periods of the original sources.

Fig. 2.2 illustrates a scatter plot of an instantaneous mixture of two sparse sources. The red and green lines represent the mixing vectors of source 1 and source 2, respectively. Denote $\mathcal{A}$ as a set of all possible mixing vectors and each possible mixing vector is computed by using an angle parameter $\theta \in \{0, \pi\}$:

$$\hat{\mathbf{a}}(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \end{bmatrix}^T. \quad (2.25)$$

Figure 2.2: A scatter plot of an instantaneous mixture of two sparse sources, where the red and green lines represent the mixing vectors of the first source and the second source, respectively. The x-axis (x_1) and y-axis (x_2) denote the amplitudes of the first observation and the second observation, respectively.

The angle difference between each observation point and every $\hat{\mathbf{a}}(\theta) \in \mathcal{A}$ is computed. Then, a probability $P(\theta)$, which is proportional to the number of observation points that have small angle difference with $\hat{\mathbf{a}}(\theta)$, can be computed. By repeating this with all possible θ , we will obtain two optimum angles with the two highest probabilities. Hence, the estimated mixing vectors that are computed by using these two optimum angles can best describe the real mixing vectors. Fig. 2.3 illustrates a histogram of $P(\theta)$ based on the data in Fig. 2.2.

The clustering approach is very intuitive in a two-dimensional space. The mixing vectors can be obtained by clustering the observations based on the angle parameter in polar coordinates. This does not require any knowledge about the arrangement of the sensor array. However, as the number of sensors increases, the dimension of the space that the data lies in becomes higher. In

Figure 2.3: A histogram of probability, for which the observation has the same direction as $\hat{a}(\theta)$ based on the data in Fig. 2.2.

N -dimensional space, the Cartesian coordinates x_i can be computed as [22]:

$$\begin{aligned}
 x_1 &= r \cos(\theta_1), \\
 x_i &= r \cos(\theta_i) \prod_{j=1}^{i-1} \sin(\theta_j), \\
 x_N &= r \prod_{j=1}^{N-1} \sin(\theta_j),
 \end{aligned} \tag{2.26}$$

where r is the magnitude. In this context, r is not important as the observation points are clustered according to their directions. Hence, the possible mixing vectors in N -dimensional space can be computed as:

$$\hat{\mathbf{a}} = \begin{bmatrix} \cos(\theta_1) & \cos(\theta_2) \sin(\theta_1) & \cdots & \cos(\theta_{N-1}) \prod_{j=1}^{N-2} \sin(\theta_j) & \prod_{j=1}^{N-1} \sin(\theta_j) \end{bmatrix}^T \tag{2.27}$$

As the number of angle parameters that are used to compute $\hat{\mathbf{a}}$ becomes larger, the computational cost increases significantly. Therefore, the clustering approach does not suit our purpose as we are interested in overdetermined

scenarios, which have more sensors than sources. In that case, the observation data lies in a high-dimensional space, which leads to a high computational effort.

2.3.7 Nonnegative Matrix Factorization

The nonnegative matrix factorization (NMF) method was proposed by Lee and Seung [90], which is based on the works on positive matrix factorization [135, 136]. The goal of NMF is to decompose a matrix into two nonnegative matrices. It has been used for many applications such as image processing [59, 95], document clustering [160, 195], biomedical data classification [26, 82] and spectral data analysis [75, 141]. In recent years, NMF has been applied to the single-channel BSS problem [134, 165, 191]. In audio applications, the spectrogram of signals is used as the nonnegative matrix. Thus, to address multichannel data, it can be done by either concatenating the spectrogram of each channel into a large matrix [140] or considering a nonnegative tensor factorization (NTF), which is also known as a nonnegative parallel factor analysis (PARAFAC), where the spectrograms of multiple channels are modelled by a 3-valence tensor [47, 156]. The NMF method does not fit in the approaches that are either in Category 1 or Category 2, as it directly estimates the original sources neither by determining the mixing channels nor by determining the separation parameters. The NMF method does not suffer from the permutation ambiguity as the analysis is performed using the spectrogram of signals [134, 164].

The main idea of the NMF method in BSS applications is to represent the spectrogram of source signals $\mathbf{V} \in \mathbb{R}^{K \times P}$ by a summation of latent components, which can be seen as a decomposition of one nonnegative matrix into two nonnegative matrices:

$$\mathbf{V} = \mathbf{W}\mathbf{H}. \quad (2.28)$$

The first matrix $\mathbf{W} \in \mathbb{R}^{K \times T}$ describes a dictionary of atoms or hidden features of the sources while the second matrix $\mathbf{H} \in \mathbb{R}^{T \times P}$ describes a set of

activation coefficients across signal blocks. The parameters in the nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ can then be updated in an iterative manner by using a multiplicative update rule with a divergence as a cost function, e.g., Euclidean distance, [140], Kullback-Leibler divergence [47] or Itakura-Saito divergence [45]. Once it converges, the source images can be obtained using the estimated nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$. The proper audio signals are then reconstructed using the source images by Wiener filtering.

The NMF approach can be particularly useful in separating music signals [55, 196] as the atoms in the first matrix can be seen as representing music notes [178]. For speech applications, the convergence rate can be relatively slow and the method can be trapped in local minima [116]. Choosing an appropriate number of latent components K is not obvious. Furthermore, most of the NMF methods for speech applications [17, 166] that achieve satisfactory performance are supervised or semi-supervised, where pretraining is required, which can be less useful in practical real-world applications.

2.4 Permutation and Scaling Ambiguity

In this section, we first discuss the permutation and scaling ambiguity in BSS approaches that operate in the time-frequency (TF) domain. Next, we study the existing approaches for coping with these issues.

To increase the computational efficiency, the convolutive BSS problem is often transformed into the time-frequency domain. Hence, operations with large Toeplitz matrix and computations of the long filter coefficients can be avoided. However, the drawbacks are permutation and scaling ambiguities across the spectrum. That means the order and the scaling factor of the separated sources are unknown in every frequency bin k [125]:

$$\mathbf{y}_{p,k} = \mathbf{W}_k^H \mathbf{x}_{p,k} = \mathbf{P}_k \mathbf{\Lambda}_k \mathbf{s}_{p,k}, \quad (2.29)$$

where $\mathbf{P}_k$ is a permutation matrix and $\mathbf{\Lambda}_k$ is a diagonal matrix. The diagonal entries of the latter matrix represent the scaling factors of the corresponding

sources. If the scaling issue is not resolved, the recovered sources are filtered versions of the original sources but their temporal envelopes are similar to the original ones. Hence, they are still yet intelligible. We can utilize the scaling ambiguity to shorten the impulse response and reduce the computational complexity of the algorithm, where the details are discussed in Chapter 6. Therefore, we concentrate on the permutation problem rather than the scaling issue. We study existing approaches to address the permutation issue.

Various approaches for solving the permutation ambiguity exist. Based on the literature, we classify these approaches into three main categories. The strategies of these categories are (i) based on the recovered signals in the time-frequency domain [9, 80, 124], (ii) based on the direction of the arrival (DOA) of the signals [73, 89, 118, 157] and (iii) based on the mixing matrices in each frequency bin [12, 15, 143, 155, 168].

The first category assumes that if the recovered band-passed signals originate from the same source signal, then they are under the influence of a similar modulation in amplitude [124]. Hence, they solve the permutation based on the correlation between the temporal envelopes of band-passed signals. This strategy employs knowledge of the recovered band-passed signals, which means the demixing matrix $\mathbf{W}_k$ is required. Moreover, the correlation between temporal envelopes is weak when frequencies are apart [80].

The second category employs knowledge of the array structure of the microphones to model the mixing matrix $\mathbf{A}_k$ in each frequency bin with an angle parameter θ_k [73, 89]. Based on the estimated demixing matrix $\mathbf{W}_k$, the method approximates θ_k by computing the inner product between each demixing vector and each mixing vector that is generated using θ_k in every frequency bin k . If the inner products produce null or small values, that means the receiving signals are generated from those directions. However, not every frequency bin produces an ideal null at the exact direction [157]. Moreover, in reality most signal energy comes from other directions in a reverberant environment.

The third category exploits the continuity of the frequency response of

the mixing filter [143]. The study by Baumann et al. [15] assumes that the transfer functions that are results from concatenating all mixing matrices $\mathbf{A}_k$ will be smooth across the spectrum, as in the case of a beamforming preprocessor. That means the difference of the mixing filter coefficients between adjacent frequency bins should be small. This implies that the Hermitian angle [148] of the mixing vectors between neighbour frequency bins is small [12]. Therefore, the minimum distance or the Hermitian angle between the adjacent filter coefficients is used to determine the permutation. However, this approach is sensitive to any misaligned frequency bins [143].

To conclude, solutions for determining the permutation in each frequency bin can be classified into three main categories. The first category uses the recovered sources, the second category employs knowledge of the sensor array and the demixing matrices while the third category exploits the mixing matrices to resolve the permutation ambiguity. The best solution depends on the type of BSS approach that is used. For instance, the first and second categories are more suitable for the BSS approaches in Category 1, which intent to find the demixing matrices. The third category suits the BSS methods in Category 2, which aim to estimate the mixing matrices.

2.5 Existing Beamforming Techniques

Another family of approaches to solving the cocktail party problem is beamforming [3]. These methods utilize spatial information of the observations to design a receiving filter based on the direction of arrival (DOA) of the desired source such that the maximum energy of the desired source is obtained. In contrast with BSS algorithms, beamforming approaches do not suffer from a permutation ambiguity as the directions of the sources are presumably known. Moreover, the array structure of microphones is also known in beamforming approaches while BSS algorithms do not rely on the locations of microphones.

Beamforming techniques have been applied in many other applications,

for instance, sonar [86], seismology [60], biomedical [104] and especially wireless communications [30, 100]. For acoustic applications, the performance of beamforming techniques may degrade due to reverberation, which consists of reflections from hard surfaces [46]. It was shown in [176] the BSS algorithms performed slightly better than beamforming techniques when the number of desired sources increased.

We show that the combination of the BSS algorithm and the beamforming techniques can be useful when new sources are added into the system in Chapter 5. In this section, we briefly discuss the basic principles of the beamformers that are commonly used, such as Capon's beamformer, linearly constrained minimum variance (LCMV) beamformer and the generalized sidelobe canceller (GSC) beamformer.

2.5.1 Capon's Method and LCMV beamformer

Capon's method [27] is well-known in the context of beamforming where it leads to the minimum variance distortionless response (MVDR) beamformer. The MVDR beamformer extracts the desired source by suppressing all the signals that are coming from the direction of the interference and noise.

Using a point source model, the observation signals at TF block index (p, k) can be expressed as [192]:

$$\mathbf{x}_{p,k} = \mathbf{a}_k(\theta)s_{p,k} + \mathbf{i}_{p,k} + \mathbf{n}_{p,k}, \quad (2.30)$$

where $\mathbf{i}_{p,k}$ and $\mathbf{n}_{p,k}$ represent the interference and noise, respectively, and $\mathbf{a}_k(\theta)$ denotes the steering vector of the desired source, where θ is the direction of arrival of the desired source. The DOA can be determined by using the multiple signal classification (MUSIC) technique [158] or the steered-response power phase transform (SRP-PHAT) algorithm [40, 41]. By knowing the geometry of the sensor array and the DOA θ , the steering vector $\mathbf{a}_k(\theta)$ can be computed.

The estimated of the desired source at TF block index (p, k) can be

extracted by

$$y_{p,k} = \mathbf{w}_k^H \mathbf{x}_{p,k} = \mathbf{w}_k^H \mathbf{a}_k(\theta) s_{p,k} + \mathbf{w}_k^H \mathbf{i}_{p,k} + \mathbf{w}_k^H \mathbf{n}_{p,k}, \quad (2.31)$$

where $\mathbf{w}_k$ is the weight vector of the beamformer. Then, the signal-to-interference-plus-noise (SINR) ratio can be calculated as:

$$\text{SINR}_k = \frac{\sigma_k^2 |\mathbf{w}_k^H \mathbf{a}_k(\theta)|^2}{\mathbf{w}_k^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k} \mathbf{w}_k}, \quad (2.32)$$

where σ_k^2 is the power spectral density of the desired source at frequency bin index k and $\mathbf{R}_{\mathbf{i}+\mathbf{n}}$ is the covariance matrix of the interference and noise, which can be expressed as:

$$\mathbf{R}_{\mathbf{i}+\mathbf{n},k} = \frac{1}{N_p} \sum_{p=0}^{N_p-1} (\mathbf{i}_{p,k} + \mathbf{n}_{p,k})^H (\mathbf{i}_{p,k} + \mathbf{n}_{p,k}), \quad (2.33)$$

where N_p denotes the length of the signals in the TF domain.

Instead of maximizing the SINR value, the MVDR beamformer finds the optimal weight vector $\mathbf{w}_k$ by minimizing the denominator of (2.32) subject to $\mathbf{w}_k^H \mathbf{a}_k(\theta) = 1$. The optimization problem can be written as:

$$\underset{\mathbf{w}_k}{\text{minimize}} \quad \mathbf{w}_k^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k} \mathbf{w}_k, \quad (2.34)$$

$$\text{subject to} \quad \mathbf{w}_k^H \mathbf{a}_k(\theta) = 1. \quad (2.35)$$

The constraint (2.35) is to prevent a trivial estimate with all zero values in $\mathbf{w}_k$. By using the Lagrange multiplier method, the optimal weight vector $\mathbf{w}_k$ can be obtained as:

$$\mathbf{w}_k = \frac{\mathbf{R}_{\mathbf{i}+\mathbf{n},k}^{-1} \mathbf{a}_k(\theta)}{\mathbf{a}_k(\theta)^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k}^{-1} \mathbf{a}_k(\theta)}. \quad (2.36)$$

Alternatively, we can extend the MVDR beamformer by adding more linear constraints while minimizing (2.34). For instance, we want to make the weight vector $\mathbf{w}_k$ orthogonal to the steering vectors of interfering signals. The minimization problem can then be reformulated as:

$$\underset{\mathbf{w}_k}{\text{minimize}} \quad \mathbf{w}_k^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k} \mathbf{w}_k, \quad (2.37)$$

$$\text{subject to} \quad \mathbf{C}_k \mathbf{w}_k = \mathbf{f}_k, \quad (2.38)$$

where the constraint (2.38) is defined by a set of linear equations that is specified by $\mathbf{C}_k$ and $\mathbf{f}_k$. This is known as the linearly constrained minimum variance (LCMV) beamformer [49]. The optimum weight vector $\mathbf{w}_k$ can be found by using the Lagrange multiplier method as:

$$\mathbf{w}_k = \mathbf{R}_{\mathbf{i}+\mathbf{n},k}^{-1} \mathbf{C}_k (\mathbf{C}_k^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k}^{-1} \mathbf{C}_k)^{-1} \mathbf{f}_k. \quad (2.39)$$

The conventional MVDR and LCMV beamformers are sensitive to perturbations. If the desired source moves slightly, the performance can be affected. Moreover, the covariance matrix of the interference and noise is usually unavailable in practice, and the sample covariance matrix of the observations $\mathbf{R}_{\mathbf{xx},k}$ is used in (2.34) and (2.37) instead. This also affects the performance of the beamformers unless the covariance matrix of the observations $\mathbf{R}_{\mathbf{xx},k}$ is obtained with very high accuracy [192].

2.5.2 Generalized Sidelobe Canceller

The generalized sidelobe canceller (GSC) was first proposed in [53]. It is an alternative formulation of the LCMV beamformer [49], where the constraints are inherited implicitly in the cost function.

A block diagram of the GSC beamformer is illustrated in Fig. 2.4. The fixed beamformer strengthens the signals that are arriving from the direction of the desired source and the output is denoted as $\hat{y}$. The blocking matrix, which defines of a subspace that is orthonormal to the fixed beamformer, filters out the desired source. Thus, the output after the blocking matrix contains only the interference and noise. Then, the adaptive noise canceller is designed to eliminate the interference and noise in $\hat{y}$. So, the final output y will contain only the desired source.

Define p and k the block index and the frequency bin index, respectively, in the time-frequency (TF) domain. Let denote by $\mathbf{w}_{q,k} \in \mathbb{C}^{M \times 1}$ the weight of the fixed beamformer, $\mathbf{B}_k \in \mathbb{C}^{M \times N}$ the blocking matrix, where $M > N$, and $\mathbf{w}_{a,k} \in \mathbb{C}^{N \times 1}$ the weight of the adaptive noise canceller. The output of

Figure 2.4: A block diagram of the GSC beamformer [57].

the GSC beamformer in the TF block index (p, k) can be expressed as:

$$\mathbf{y}_{p,k} = (\mathbf{w}_{q,k} - \mathbf{B}_k \mathbf{w}_{a,k})^H \mathbf{x}_{p,k}. \quad (2.40)$$

The weight of the fixed beamformer $\mathbf{w}_{q,k}$ and the blocking matrix $\mathbf{B}_k$ can be computed using the constraints that are defined in the LCMV beamformer [93], e.g., $\mathbf{C}_k$ and $\mathbf{f}_k$ in (2.37).

The optimum weight for the noise canceller can be found by solving the following unconstrained optimization problem [53]:

$$\underset{\mathbf{w}_{a,k}}{\text{minimize}} \quad \sum_p |\mathbf{y}_{p,k}|^2. \quad (2.41)$$

The optimization will not suppress the strength of the desired source for any value of $\mathbf{w}_{a,k}$ as the component of the desired source in $\mathbf{B}_k^H \mathbf{x}_{p,k}$ is suppressed by the blocking matrix $\mathbf{B}_k$. Substituting (2.40) into (2.41) yields:

$$\underset{\mathbf{w}_{a,k}}{\text{minimize}} \quad (\mathbf{w}_{q,k} - \mathbf{B}_k \mathbf{w}_{a,k})^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} (\mathbf{w}_{q,k} - \mathbf{B}_k \mathbf{w}_{a,k}). \quad (2.42)$$

By taking the derivative and setting the result to zero, the closed-form solution of (2.42) can be expressed as

$$\mathbf{w}_{a,k} = (\mathbf{B}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{B}_k)^{-1} \mathbf{B}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{w}_{q,k}. \quad (2.43)$$

The conventional GSC approach is sensitive to the DOA of the desired source. When a mismatch occurs, the desired source will be treated as interference and the optimum weight of the noise canceller tends to suppress this component as well as the interference and noise. Hence, the performance can be significantly degraded. We will later discuss the robustness improvement of the beamforming techniques in Section 2.5.3.

2.5.3 Robustness Improvement

Conventional beamforming techniques rely heavily on the estimation of the DOA to obtain an appropriate steering vector of the desired source. In practice, the measurement of DOA can be inaccurate due to noise or non-stationary environments. For instance, the sources may slightly change their positions. This may degrade the performance of the beamformer [94].

Many methods have been developed to improve the robustness of beamforming. The robust adaptive beamforming is the most common one [94]. An overview of this approach can be found in [192]. In this subsection, we briefly discuss two approaches to improve the robustness of the system. The first method is an approach to approximate the covariance matrix while the second method is diagonal loading.

The covariance matrix is a core tool for many signal processing methods, especially for beamforming. In real-world scenarios, the covariance matrix obtained from the raw observation data can be inaccurate due to noise, outliers and an insufficient number of samples [66, 109]. Consider mixtures of L zero-mean and uncorrelated signals $\mathbf{s}$, e.g., $\mathbf{x}_n = \mathbf{A}\mathbf{s}_n$, where n denotes the sample index and $\mathbf{A}$ is a mixing matrix. The covariance matrix of the mixtures without environmental noise can be expressed as:

$$\mathbf{R}_{\mathbf{xx}} = \sum_n \mathbf{x}_n \mathbf{x}_n^T = \sum_n \mathbf{A} \mathbf{s}_n \mathbf{s}_n^T \mathbf{A}^T = \mathbf{A} \mathbf{\Lambda}_s \mathbf{A}^T + \mathbf{E}, \quad (2.44)$$

where $\mathbf{\Lambda}_s$ is a diagonal matrix, whose diagonal entries consist of the variance of every signal, e.g., $\sigma_{s_1}^2, \dots, \sigma_{s_L}^2$, and $\mathbf{E}$ denotes an error matrix, which is

caused by the covariance of the signals, e.g., $\sigma_{s^{i_s j}}, i, j \in [1, L], i \neq j$. The error matrix $\mathbf{E}$ vanishes if we have a sufficient number of samples, otherwise it can be relatively large due to outliers. Hence, the estimation accuracy of the covariance matrix becomes important for the case of a small number of samples.

Tyler's covariance estimator [183] is a well-known approach to approximate the covariance matrix due to its simplicity. It guarantees to converge [199] and it is also robust to outliers, thus the effect of the error matrix $\mathbf{E}$ in (2.44) can be minimized. Tyler proved that given M -dimensional independent and identically distributed (i.i.d.) samples, e.g., $\mathbf{x}_i$, where $i = 1, \dots, N$, which is drawn from an elliptical distribution, a good covariance estimator is defined as the solution to the fixed point equation

$$\mathbf{Q} = \frac{M}{N} \sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{Q}^{-1} \mathbf{x}_i}, \quad (2.45)$$

where M is the dimensional number of the samples while N is the total number of samples. Tyler's covariance estimator is also a minimax robust estimator. In addition, it is distribution-free in the sense that its asymptotic variance does not depend on the parametric form of the underlying distribution [171].

Regularization is a technique to make the solution of an optimization problem more robust to perturbations. It adds a scalar penalty to the norm of the estimated parameters to prevent obtaining a solution with large entries [25]. Among the regularization techniques, the Tikhonov regularization is the most common. A Tikhonov regularized optimization problem can be written as:

$$\underset{\mathbf{y}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{y} - \mathbf{b}\|^2 + \lambda \|\mathbf{y}\|^2, \quad (2.46)$$

where λ is a regularized parameter and $\mathbf{A} \in \mathbb{C}^{M \times L}$ as well as $\mathbf{b} \in \mathbb{C}^{M \times 1}$ are given. Taking the derivatives and setting to zero, we obtain

$$\mathbf{y} = (\mathbf{A}^H \mathbf{A} + \lambda \mathbf{I})^{-1} \mathbf{A}^H \mathbf{b}. \quad (2.47)$$

The regularized parameter can also be added while designing the weight vector in beamformer applications. For instance, the optimization problem of a regularized MVDR beamformer can be written as:

$$\underset{\mathbf{w}_k}{\text{minimize}} \quad \mathbf{w}_k^H \mathbf{R}_{\mathbf{i}+\mathbf{n},k} \mathbf{w}_k + \lambda \|\mathbf{w}_k\|^2, \quad (2.48)$$

$$\text{subject to} \quad \mathbf{w}_k^H \mathbf{a}_k(\theta) = 1. \quad (2.49)$$

The solution of (2.48) subject to (2.49) is:

$$\mathbf{w}_k = \frac{(\mathbf{R}_{\mathbf{i}+\mathbf{n},k} + \lambda \mathbf{I})^{-1} \mathbf{a}_k(\theta)}{\mathbf{a}_k(\theta)^H (\mathbf{R}_{\mathbf{i}+\mathbf{n},k} + \lambda \mathbf{I})^{-1} \mathbf{a}_k(\theta)}. \quad (2.50)$$

Instead of inverting the covariance matrix, we invert the combination of the covariance and a scaled identity matrix. This is called as a diagonal loading technique, which can also be seen as a form of Tikhonov regularization [137]. It improves the robustness of the system and also avoids ill-posed inverse problems as the covariance matrix will always be full rank after diagonal loading.

2.6 Discussion

Most existing BSS approaches, except the NMF method, can be categorized into two categories, where Category 1 intends to estimate the mixing channels while Category 2 aims to find the demixing parameters. All approaches have their own advantages and disadvantages for different scenarios. For instance, the ICA and joint diagonalization approach perform BSS based on the statistical properties of the original sources, hence a sufficient number of samples is required. However, they do not rely on activity behaviours of the sources. The sparsity-based methods, such as DUET and TIFROM, perform BSS by analyzing activity periods of the original sources, thus number of samples can be reduced by selecting appropriate time intervals for analysis. The BCI and NMF methods have an advantage over other approaches that

they do not suffer from the permutation issue. Most BSS methods require time-invariant channel, e.g., stationary sources.

Beamforming techniques can also be used to address the cocktail party problem in far-field scenarios or in environments that have less reverberation. However, they require knowledge about locations of the microphones and the sources for designing steering vectors. The performance can be significantly affected due to any misalignment and reverberation.

Methods in most literatures have a common problem by using a set of instantaneous mixing model in the TF domain to approximate the convolutive mixing model in the time domain. This assumption is sufficiently correct when the window length is at least twice the length of the mixing channel and the window length decides the algorithmic latency of a real-time system. Hence, directly implementing the existing BSS approaches in a real-time system is not applicable, additional processing techniques are required.

In a non-stationary scenario, the adaptation speed of the system becomes crucial. The BSS system should adapt as soon as possible with the environmental changes, which can be due to time-varying channels, moving sources, dynamic source number and noise. Repeating BSS processes affects the efficiency of the system. Hence, “meta-information” that is obtained from previous operations can be preserved to facilitate the future separation. For instance, when a new source is introduced into the system, we do not need to re-perform the estimation from scratch. By using the “meta-information”, only the estimation of the demixing vector or the mixing vector for the new source is required.

The latency issue is important while designing a BSS system that is able to operate in a real-world scenario. In the following chapters, we propose several strategies to reduce the system latency, which can be caused by the algorithmic latency and the estimation delay.

Chapter 3

Subspace Method for Blind Source Separation With Convolutive Transfer Function Model

In this chapter we propose a method that is able to directly estimate demixing parameters of a blind source separation (BSS) system in the low-frequency-resolution time-frequency (LFRTF) domain. The time-frequency (TF) observation signals fall in a LFRTF domain when the length of the window, that is used to perform the short-time Fourier transform, is shorter than the length of the mixing filters. In the LFRTF domain, the mixing process is still convolutive but the length of the mixing filters is short. Our proposed method facilitates the deconvolution of a convolutive mixture to a new instantaneous mixture, so traditional instantaneous TF-domain BSS algorithms can still be adopted. The proposed method has two advantages. First, the demixing parameters are shareable between TF frames. For instance, the estimated demixing parameters of the current frame can be applied to the next incoming frame. Second, it quickens the estimation process as it facilitates the estimation of the demixing parameters with a shorter time-segment observa-

tion samples compared to the traditional BSS algorithms. The experimental results confirm the effectiveness of the proposed approach.

3.1 Introduction

The system latency or delay is a crucial factor that makes most of existing TF-domain BSS algorithms inapplicable in real-time systems. The TF-domain model is commonly adopted to approximate convolutive time-domain mixing model with a series of multiplicative mixing model in the frequency domain. However, the TF-domain model is sufficiently accurate only when the window length is at least twice the length of the impulse response. This leads to a long time lag between inputs and outputs of a system as the window length decides the algorithmic latency.

As an alternative to TF-domain BSS, we utilize the subspace-based methods to facilitate the deconvolution of a convolutive mixture to a new instantaneous mixture. These methods are well established in the telecommunication area [52, 101, 102, 108, 167, 184–186]. However, these approaches are not suitable for audio applications as the impulse responses in these applications are more complex compared to the telecommunication channels due to reverberations. The acoustic impulse response can be relatively long and this makes the computational cost of the algorithm high, and hence it is impracticable. Moreover, the estimated parameters of the systems proposed in [167, 186] cannot be shared between different time intervals even in a time-invariant scenario without further processing, see Sec. 3.3.

To reduce the algorithmic latency caused by the TF-domain BSS approaches, the window length must be reduced. If the window length is shorter than the impulse response, it results in a low-frequency-resolution time-frequency (LFRTF) domain. In the LFRTF domain, the mixing model of the BSS problem is still convolutive but the length is significantly shorter than the time-domain mixing filters [96, 174]. Thus, the problem becomes tractable as the size of the matrices is now smaller compared to the original

problem formulation in the time domain. This facilitates matrix operations, such as singular-value-decomposition and matrix inversion under a certain constraint of computational resources. The aforementioned subspace-based method can then be implemented in every frequency bin in the LFRTF domain.

In this chapter, we propose a subspace approach that enables us to solve the TF-domain BSS problem directly in the LFRTF domain. Our method differs from [167, 186] in that the estimated parameters can directly be used to separate the mixtures of the next incoming block of samples, provided that the environment is time-invariant or changing gradually. The computational complexity is also lower. Compared to the traditional TF-domain BSS approaches, our method is able to perform the estimation with a shorter length of signal samples, thus the separation parameters can be updated quicker.

This chapter first starts with an introduction to TF-domain BSS models that operate in the LFRTF domain. Second, we briefly review the most recent literature approaches. Next, we provide the proposed method that performs the estimation process in the LFRTF domain. Then, we present the experimental results of our proposed method and comparisons with the other literature approaches.

3.2 Blind Source Separation Model

This section outlines a method for the model of blind source separation (BSS) in the time-frequency (TF) domain that uses a convolutive transfer function (CTF) approximation. The model allows the length of the short-time (ST) window to be shorter than the length of the mixing filters. It is particularly useful for real-time systems, which require a short algorithmic latency. In the derivation, we consider a noise-free scenario.

We denote by M the number of observations and by L the number of original sources. Consider a linear time-invariant (LTI) system with finite mixing filters that are longer than the short-time (ST) window. Thus, the

mixing model is still convolutive but the memory length is shorter than the time-domain mixing filters. Then the mixtures in the TF domain can be approximated as [14, 33, 174]:

$$\mathbf{x}_{p,k} \approx \sum_{p'=0}^{T-1} \mathbf{A}_{p',k} \mathbf{s}_{p-p',k}, \quad (3.1)$$

where p and k are the time frame index and the frequency bin index, respectively, where $\mathbf{x}_{p,k} \in \mathbb{C}^{M \times 1}$ is a vector observation signal at TF instant (p, k) . $\mathbf{A}_{p',k} \in \mathbb{C}^{M \times L}$ indicates the mixing filters at TF instant (p', k) while T represents the length of the filters in the TF domain. A vector of the original sources at TF instant (p, k) is represented by $\mathbf{s}_{p,k} \in \mathbb{C}^{L \times 1}$. It is worth noting that the model in (3.1) is a simplified model of (2.9) by considering only the band-to-band mixing filters and neglecting the contributions of the cross-band filters.

Alternatively, (3.1) can be written in a matrix form:

$$\mathbf{x}_{p,k} = \mathbf{A}_k \hat{\mathbf{s}}_{p,k}, \quad (3.2)$$

where

$$\begin{aligned} \mathbf{A}_k &= \begin{bmatrix} \mathbf{A}_{0,k} & \cdots & \mathbf{A}_{T-1,k} \end{bmatrix} \in \mathbb{C}^{M \times LT}, \\ \hat{\mathbf{s}}_{p,k} &= \begin{bmatrix} \mathbf{s}_{p-T+1,k}^T & \cdots & \mathbf{s}_{p,k}^T \end{bmatrix}^T \in \mathbb{C}^{LT \times 1}. \end{aligned}$$

The goal of BSS is to find a demixing matrix $\tilde{\mathbf{W}}_k \in \mathbb{C}^{M \times L}$ for every frequency bin k , so that the original sources can be recovered by:

$$\mathbf{y}_{p,k} = \tilde{\mathbf{W}}_k^H \mathbf{x}_{p,k}, \quad (3.3)$$

where $\mathbf{y}_{p,k}$ indicates the estimated original sources and $[\cdot]^H$ denotes a Hermitian transpose.

3.3 Conventional Subspace Methods

In this section, we briefly review the most recent subspace methodology [167, 186] that facilitates the deconvolution of a convolutive mixture to a

new instantaneous mixture. This allows us to estimate the original sources in the low-latency domain using the traditional instantaneous TF-domain BSS approaches, for instance, the independent component analysis (ICA) approach.

For N time-frequency samples (3.2) can be rewritten in a matrix form:

$$\mathbf{X}_k = \hat{\mathbf{S}}_k \mathbf{A}_k^H, \quad (3.4)$$

where $\mathbf{X}_k = \begin{bmatrix} \mathbf{x}_{T-1,k} & \mathbf{x}_{T,k} & \cdots & \mathbf{x}_{T+N-2,k} \end{bmatrix}^H \in \mathbb{C}^{N \times M}$ and

$$\hat{\mathbf{S}}_k = \begin{bmatrix} \mathbf{s}_{T-1,k}^H & \mathbf{s}_{T-2,k}^H & \cdots & \mathbf{s}_{0,k}^H \\ \mathbf{s}_{T,k}^H & \mathbf{s}_{T-1,k}^H & \cdots & \mathbf{s}_{1,k}^H \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{s}_{T+N-2,k}^H & \mathbf{s}_{T+N-3,k}^H & \cdots & \mathbf{s}_{N-1,k}^H \end{bmatrix} \in \mathbb{C}^{N \times LT}. \quad (3.5)$$

If the system is overdetermined, i.e., $\mathbf{A}_k$ has more rows than columns ($M \geq LT$, which means more microphones than LT), and $\mathbf{A}_k$ has full row rank, then the space spanned by the columns of $\mathbf{X}_k$ is also spanned by the columns of $\hat{\mathbf{S}}_k$. $\hat{\mathbf{S}}_k$ in (3.4) can also be written as [186]:

$$\hat{\mathbf{S}}_k = \mathbf{U}_k \mathbf{G}_k, \quad (3.6)$$

where $\mathbf{U}_k \in \mathbb{C}^{N \times LT}$, the columns of which consist of the first LT dominant column spaces of $\mathbf{X}_k$ that can be obtained by performing a singular value decomposition on $\mathbf{X}_k$, and $\mathbf{G}_k$ is a coefficient matrix $\in \mathbb{C}^{LT \times LT}$, which acts as another mixing matrix between $\hat{\mathbf{S}}_k$ and $\mathbf{U}_k$.

The essence of the subspace methods is to omit the redundant terms, e.g., terms in the columns of the original sources $\hat{\mathbf{S}}_k$ in the LFRTF domain. Hence, each original source appears only once in the columns of the source matrix in the new model. Then, conventional independent component analysis (ICA) approaches can be employed to solve the remaining problem. Now, we briefly discuss the operations in the subspace methods.

The original sources $\hat{\mathbf{S}}_k$ in (3.5) has a block-Toeplitz structure and it can be decomposed as [167]:

$$\hat{\mathbf{S}}_k = \sum_{n=0}^{N+T-1} \mathbf{E}_n \otimes \mathbf{s}_{n,k}^H, \quad (3.7)$$

where $\otimes$ denotes a Kronecker product and $\mathbf{E}_n \in \mathbb{R}^{N \times T}$ represents a Toeplitz basis matrix, where the entry in the i^{th} row and in the j^{th} column is defined as

$$\mathbf{E}_{n_{ij}} = \begin{cases} 1, & \text{if } i = j + n + 1 - T, i > 0, j > 0 \\ 0, & \text{otherwise,} \end{cases} \quad (3.8)$$

where $n \in [0, N + T - 1]$. Combining (3.6) and (3.7) yields

$$\mathbf{U}_k \mathbf{G}_k = \sum_{n=0}^{N+T-1} \mathbf{E}_n \otimes \mathbf{s}_{n,k}^H. \quad (3.9)$$

If we partition $\mathbf{G}_k$ into blocks

$$\mathbf{G}_k = \begin{bmatrix} \mathbf{G}_{0,k} & \mathbf{G}_{1,k} & \cdots & \mathbf{G}_{T-1,k} \end{bmatrix}, \quad (3.10)$$

where $\mathbf{G}_{l,k} \in \mathbb{C}^{L \times L}$ and stack the blocks together, then (3.9) can be rewritten as:

$$\begin{bmatrix} \mathbf{I}_{T \times T} \otimes \mathbf{U}_k & -\mathbf{E}_{\text{tot}} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{0,k} \\ \vdots \\ \mathbf{G}_{T-1,k} \\ \mathbf{s}_{0,k}^H \\ \vdots \\ \mathbf{s}_{N+T-1,k}^H \end{bmatrix} = \mathbf{0}, \quad (3.11)$$

where $\mathbf{E}_{\text{tot}} = \begin{bmatrix} \text{Vec}(\mathbf{E}_0) & \text{Vec}(\mathbf{E}_1) & \cdots & \text{Vec}(\mathbf{E}_{N+T-1}) \end{bmatrix} \in \mathbb{R}^{NT \times (N+T-1)}$ and $\text{Vec}(\cdot)$ represents a vectorize operation.

The original sources and the coefficient matrix $\mathbf{G}_k$ in (3.11) can be estimated by solving the homogeneous equation. However, the solution obtained will be a mixture of the original sources, see (3.27) in Sec. 3.4.2. To find the

original sources, conventional ICA approaches can be used. The pseudocode of the subspace method is provided in Algorithm 2. More details about the approach can be found in [167, 186].

Algorithm 2 Literature approach [186]

Input: $\mathbf{x}_{p,k}$

Output: $\mathbf{y}_{p,k}$

- 1: Set T and N .
 - 2: Compute $\mathbf{E}_{\text{tot}}$ using (3.8).
 - 3: **for all** k **do**
 - 4: $\mathbf{X}_k \leftarrow \begin{bmatrix} \mathbf{x}_{T-1,k} & \mathbf{x}_{T,k} & \cdots & \mathbf{x}_{T+N-2,k} \end{bmatrix}^H$.
 - 5: $\mathbf{U}_k \leftarrow$ eigenspace of $\mathbf{X}_k$ that can be obtained by the first LT eigenvectors while performing the SVD on $\mathbf{X}_k$.
 - 6: Evaluate $\begin{bmatrix} \mathbf{I}_{T \times T} \otimes \mathbf{U}_k & -\mathbf{E}_{\text{total}} \end{bmatrix}$.
 - 7: Solve the homogeneous equation (3.11) to estimate $\mathbf{S}_k^*$, where $\mathbf{S}_k^* = \begin{bmatrix} \mathbf{s}_{0,k} & \cdots & \mathbf{s}_{N+T-1,k} \end{bmatrix}^H$.
 - 8: Perform ICA on $\mathbf{S}_k^*$ and denote the solution as $\mathbf{S}_{k\text{ICA}}$.
 - 9: Solve the scaling and permutation issues on $\mathbf{S}_{k\text{ICA}}$ and denote the solution as $\mathbf{S}_k$.
 - 10: Obtain $\mathbf{y}_{p,k}$ from the p^{th} column in $\mathbf{S}_k^H$.
 - 11: **end for**
-

The subspace approach [167] has a major disadvantage. That is, the estimated $\mathbf{G}_k$ does not have a direct relationship with $\mathbf{X}_k$, which will be problematic to a real-time system. For instance, the solution that is computed using the samples of the current block, will not be applicable to the samples of the next block without further processing. Let $\mathbf{X}_k^{N_1}$ and $\mathbf{X}_k^{N_2}$ represent the observations of the first N_1 TF samples and the observations between $N_1 + 1$ and N_2 , respectively, in the frequency bin index k . By using $\mathbf{X}_k^{N_1}$, we are able to compute $\mathbf{U}_k^{N_1}$ and $\mathbf{G}_k^{N_1}$, where $\mathbf{U}_k^{N_1}$ represents the dominant column spaces of $\mathbf{X}_k^{N_1}$ while $\mathbf{G}_k^{N_1}$ represents the coefficient matrix that

is used to construct the original sources $\hat{\mathbf{S}}_k^{N_1}$:

$$\hat{\mathbf{S}}_k^{N_1} = \mathbf{U}_k^{N_1} \mathbf{G}_k^{N_1}. \quad (3.12)$$

However, $\mathbf{G}_k^{N_1}$ is not applicable to $\mathbf{X}_k^{N_2}$ to estimate the original source at N_2 , that is

$$\hat{\mathbf{S}}_k^{N_2} \neq \mathbf{U}_k^{N_2} \mathbf{G}_k^{N_1}. \quad (3.13)$$

We have to repeat the entire estimation process on $\mathbf{X}_k^{N_2}$ to estimate $\hat{\mathbf{S}}_k^{N_2}$, which is redundant. Although it is natural to estimate a demixing matrix $\tilde{\mathbf{W}}_k$ from $\mathbf{X}_k^{N_1}$ and $\hat{\mathbf{S}}_k^{N_1}$, such that

$$\hat{\mathbf{S}}_k^{N_1} = \mathbf{X}_k^{N_1} \tilde{\mathbf{W}}_k, \quad (3.14)$$

this requires a matrix inversion operation that can lead to problems, e.g., an ill-posed problem. Furthermore, discrepancies can be present between the true original sources and the estimated original sources $\hat{\mathbf{S}}_k^{N_1}$ and this makes the estimate of the demixing matrix $\tilde{\mathbf{W}}_k$ inaccurate.

3.4 Proposed Subspace Method

In this section, we reformulate the problem in Section 3.3 and discuss our proposed approach. The proposed approach allows us to directly estimate the demixing operator, so that the demixing operator can be applied directly on the upcoming observation samples. Thus, it is well suited to a real-time system.

3.4.1 Problem Formulation

To estimate the original sources at different time instant, we are interested in finding a demixing matrix $\mathbf{W}_k \in \mathbb{C}^{M \times LT}$, such that

$$\mathbf{A}_k^H \mathbf{W}_k = \mathbf{I}. \quad (3.15)$$

By multiplying (3.4) with $\mathbf{W}_k$, we obtain

$$\mathbf{X}_k \mathbf{W}_k = \hat{\mathbf{S}}_k, \quad (3.16)$$

where $\hat{\mathbf{S}}_k$ contains the original sources at different time instant. That is, each original source is presented in multiple columns of $\hat{\mathbf{S}}_k$ but with different delays.

The essence of the proposed method is similar to the approaches [167, 186] that are discussed in Sec. 3.3. By using equations from (3.7) to (3.11), we are able to decompose $\hat{\mathbf{S}}_k$ and obtain a matrix of original sources, where each original source is represented by one column only. The difference is that the demixing matrix can be directly estimated using our model. Thus, in a time-invariant environment the estimated parameters are applicable for later operations.

We partition $\mathbf{W}_k$ into blocks:

$$\mathbf{W}_k = \begin{bmatrix} \mathbf{W}_{0,k} & \mathbf{W}_{1,k} & \cdots & \mathbf{W}_{T-1,k} \end{bmatrix}, \quad (3.17)$$

where $\mathbf{W}_{l,k} \in \mathbb{C}^{M \times L}$. By stacking the blocks together and utilizing the block-Toeplitz structure of $\hat{\mathbf{S}}_k$, (3.16) can be written as:

$$\hat{\mathbf{X}}_k \hat{\mathbf{W}}_k = \mathbf{E}_{\text{tot}} \mathbf{S}_k, \quad (3.18)$$

where

$$\hat{\mathbf{X}}_k = \begin{bmatrix} \mathbf{I}_{T \times T} \otimes \mathbf{X}_k \end{bmatrix} \in \mathbb{C}^{NT \times MT}, \quad (3.19)$$

$$\hat{\mathbf{W}}_k = \begin{bmatrix} \mathbf{W}_{0,k}^H & \cdots & \mathbf{W}_{T-1,k}^H \end{bmatrix}^H \in \mathbb{C}^{MT \times L}, \quad (3.20)$$

$$\mathbf{S}_k = \begin{bmatrix} \mathbf{s}_{0,k} & \cdots & \mathbf{s}_{N+T-1,k} \end{bmatrix}^H \in \mathbb{C}^{(N+T-1) \times L}. \quad (3.21)$$

To get insights into the arrangement of the matrices from (3.16) to (3.18), we provide an illustration of the operations in Fig. 3.1.

Figure 3.1: An illustrative example of the operations from (3.16) to (3.18).

3.4.2 Principle of the Proposed Method

The goal is to reformulate the convolutive BSS problem into an instantaneous problem in the LFRTF domain. As the original sources in the LFRTF domain have a block-Toeplitz structure, by using a different arrangement of equations a Toeplitz basis matrix can be employed to decompose convolutive mixtures into instantaneous mixtures. In this subsection, we discuss the principle of our proposed method.

Let $\hat{\mathbf{W}}_k^*$ and $\mathbf{S}_k^*$ be the solutions of (3.18). This indicates that

$$\mathbf{S}_k^* = \mathbf{E}_{\text{tot}}^\dagger \hat{\mathbf{X}}_k \hat{\mathbf{W}}_k^*, \quad (3.22)$$

where $\dagger$ represents a Moore-Penrose pseudoinverse. We note that

$$\frac{1}{T} \mathbf{E}_{\text{tot}}^T \mathbf{E}_{\text{tot}} \approx \mathbf{I}, \quad (3.23)$$

thus,

$$\mathbf{E}_{\text{tot}}^\dagger \approx \frac{1}{T} \mathbf{E}_{\text{tot}}^T \quad (3.24)$$

Then, we substitute $\hat{\mathbf{W}}_k^*$ together with (3.22) and (3.24) into (3.18) to eliminate $\mathbf{S}_k$. This yields:

$$\hat{\mathbf{X}}_k \hat{\mathbf{W}}_k^* = \frac{1}{T} \mathbf{E}_{\text{tot}} \mathbf{E}_{\text{tot}}^T \hat{\mathbf{X}}_k \hat{\mathbf{W}}_k^*. \quad (3.25)$$

We can rearrange (3.25) and form:

$$\left(\hat{\mathbf{X}}_k - \frac{1}{T} \mathbf{E}_{\text{tot}} \mathbf{E}_{\text{tot}}^T \hat{\mathbf{X}}_k \right) \hat{\mathbf{W}}_k^* = \mathbf{0}. \quad (3.26)$$

This shows that the solutions of $\hat{\mathbf{W}}_k$ lie in the null space of $\left(\hat{\mathbf{X}}_k - \frac{1}{T} \mathbf{E}_{\text{tot}} \mathbf{E}_{\text{tot}}^T \hat{\mathbf{X}}_k \right)$. The null space can be found by performing either a singular value decomposition (SVD) or a QR decomposition. After finding $\hat{\mathbf{W}}_k$, the original source $\mathbf{S}_k$ can be recovered using (3.22).

It is relevant to note that the dimensionality of the null space is $(M - LT)T + L$ as the rank of $\hat{\mathbf{X}}_k$ is LT^2 . Therefore, the dimensionality of $\hat{\mathbf{W}}_k^*$ is $MT \times (MT - LT^2 + L)$ instead of $MT \times L$. Moreover, every $\hat{\mathbf{W}}_k^* \mathbf{Q}_k^\dagger$ will be a solution of (3.26), where $\mathbf{Q}_k \in \mathbb{C}^{L \times (MT - LT^2 + L)}$ represents any matrix. This can be easily proven by substituting $\hat{\mathbf{W}}_k^* \mathbf{Q}_k^\dagger$ into (3.26):

$$\left(\hat{\mathbf{X}}_k - \frac{1}{T} \mathbf{E}_{\text{tot}} \mathbf{E}_{\text{tot}}^T \hat{\mathbf{X}}_k \right) \hat{\mathbf{W}}_k^* \mathbf{Q}_k^\dagger = \mathbf{0} \mathbf{Q}_k^\dagger = \mathbf{0}. \quad (3.27)$$

Hence, the estimated demixing matrix $\hat{\mathbf{W}}_k^*$ is not the “true” demixing matrix $\hat{\mathbf{W}}_k$, i.e., true original sources cannot be obtained by using $\hat{\mathbf{W}}_k^*$. However, they are related by, such that

$$\hat{\mathbf{W}}_k \mathbf{Q}_k = \hat{\mathbf{W}}_k^*. \quad (3.28)$$

Substitute (3.28) into (3.22), we obtain

$$\mathbf{S}_k^* = \frac{1}{T} \mathbf{E}_{\text{tot}}^T \hat{\mathbf{X}}_k \hat{\mathbf{W}}_k \mathbf{Q}_k = \mathbf{S}_k \mathbf{Q}_k, \quad (3.29)$$

which means that $\mathbf{S}_k^*$ is an instantaneous mixture of the original sources $\mathbf{S}_k$. Therefore, we can use any traditional instantaneous BSS algorithm, such as the conventional independent component analysis (ICA) approach to estimate $\mathbf{S}_k$ in (3.29).

3.4.3 Summary of the Proposed Algorithm

The proposed approach allows us to estimate the demixing matrix directly in the LFRTF domain. The idea is similar to [167, 186] but the principle is different. It differs in that the estimated parameters can be applied to the next incoming blocks of data samples provided that the environment is time-invariant or changing gradually. This saves computational cost and it is particularly useful for real-time applications.

The scaling and permutation ambiguities are also present in our proposed approach as the operation is performed separately in each individual frequency bin. The minimal distortion principle approach [110] can be used to address the scaling ambiguity while the permutation issue can be solved by using [9, 80, 157]. The algorithm of the proposed approach is summarized in Algorithm 3.

Algorithm 3 Proposed subspace method

Input: $\mathbf{x}_{p,k}$

Output: $\mathbf{y}_{p,k}, \tilde{\mathbf{W}}_k$

- 1: Set T and N .
 - 2: Compute $\mathbf{E}_{\text{tot}}$ using (3.8).
 - 3: Compute $\mathbf{E}_{\text{tot}}^\dagger$ by performing pseudoinverse on $\mathbf{E}_{\text{tot}}$ or use $\frac{1}{T}\mathbf{E}_{\text{tot}}^T$ instead.
 - 4: **for all** k **do**
 - 5: $\mathbf{X}_k \leftarrow \begin{bmatrix} \mathbf{x}_{T-1,k} & \mathbf{x}_{T,k} & \cdots & \mathbf{x}_{T+N-2,k} \end{bmatrix}^H$.
 - 6: $\hat{\mathbf{X}}_k \leftarrow \mathbf{I}_{T \times T} \otimes \mathbf{X}_k$.
 - 7: $\hat{\mathbf{W}}_k^* \leftarrow$ null space of $\left(\hat{\mathbf{X}}_k - \mathbf{E}_{\text{tot}} \mathbf{E}_{\text{tot}}^\dagger \hat{\mathbf{X}}_k \right)$ by performing the SVD.
 - 8: Compute $\hat{\mathbf{S}}_k^*$ using (3.22).
 - 9: Find $\mathbf{Q}_k^\dagger$ by performing ICA on $\hat{\mathbf{S}}_k^*$.
 - 10: $\hat{\mathbf{W}}_k \leftarrow \hat{\mathbf{W}}_k^* \mathbf{Q}_k^\dagger$.
 - 11: $\tilde{\mathbf{W}}_k \leftarrow \mathbf{W}_{0,k}$ $\triangleright$ the first $M \times L$ submatrix of $\hat{\mathbf{W}}_k$.
 - 12: Solve the scaling and permutation issues on $\tilde{\mathbf{W}}_k$.
 - 13: $\mathbf{y}_{p,k} \leftarrow \tilde{\mathbf{W}}_k^H \mathbf{x}_{p,k}$.
 - 14: **end for**
-

3.5 Experimental Results

This subsection presents the experimental results of the simulated data. We first discuss the experimental setup and procedure. Then, we compare the separation performance of our proposed method with the other literature approaches [32,33,167,172]. All the computations were conducted offline and computed by using Matlab R2014b on a PC having an Intel(R) Core(TM) i7-4790 CPU@3.60 GHz processor with 8GB random-access memory.

3.5.1 Experimental Setup and Procedure

As data we used the 10 seconds speech source signals of the Stereo Audio Source Separation Evaluation Campaign (SASSEC). The sampling rate was 16 kHz. In the experiment, 20 microphones were used. The positions of the sources were fixed while the microphones were randomly placed in a rectangular room with a size of $6\text{ m} \times 8\text{ m} \times 3\text{ m}$, at a fixed height of 1.5 m. For each experiment, three random speech signals were selected as sources and the positions of the microphones were randomly changed. An example of the arrangements is illustrated in Fig. 3.2, where the red crosses indicate the microphones while the blue dots represent the sources. We used the room impulse response (RIR) simulator [54], based on the image source method [5], to generate the 20×3 RIR with 4096 taps. Each set of experiment was conducted three times with different reverberation times, which were 0.15 s, 0.3 s and 0.6 s. The reverberation time is denoted as β in the simulator. The simulated RIR were then used to convolve with the speech signals in the time domain to obtain the observations in a noise-free environment.

A square root of Hann window with 2048 taps (128 ms) and 512 (32 ms) taps was used in the high-frequency-resolution time-frequency (HFRTF) domain and the low-frequency-resolution time-frequency (LFRTF) domain, respectively. The windows were 50% overlapped. For the time-domain method [172], the length of the demixing filter was set to 512 taps (32 ms). All the demixing parameters were estimated by the complex fast independent

Figure 3.2: The positions of the microphones and the sources

component analysis (ICA) approach [20] using the first 120 time-frequency (TF) frames of the observations. In each frequency bin, the permutation ambiguity was resolved using oracle information, which is based on the correlation coefficients between the amplitude envelope (absolute value) of the estimated sources and the original sources across the 120 TF frames. After correcting the permutation orders, we minimized the scaling effect using the minimal distortion principle method [110]. Then, the estimated demixing parameters were applied to the entire mixtures to obtain the separated signals.

We examined six different strategies and compared the separation performance. The algorithmic latency was 32 ms for all methods. The cross-band [33], shortening [32] and time-domain [172] approaches utilize the HFRTF domain to perform the estimation and compute the representation of the demixing operators in the low-latency domain. The cross-band [33] and the

shortening [32] methods performed the separation in the LFRTF domain while the time-domain strategy [172] performed the separation in the time domain. For [33] and [32], we considered only two cross-band filters during the computation of the demixing operators as the contributions of others were relatively weak. The subspace [167] and the proposed methods performed both the estimation and the separation in the LFRTF domain based on the description in Sec. 3.3 and Sec. 3.4, respectively. The stand-alone ICA approach [20] also operated in the LFRTF domain. It directly estimated the demixing parameters based on the instantaneous BSS model although the mixing process was convolutive.

In the first set of the experiments, for each technique, the demixing parameters were estimated using a fixed number of the observation samples in the TF domain, which was 120 TF samples. So, the segment length of the observation was 7.68 s for the approaches that performed estimation in the HFRTF domain (e.g., cross-band, shortening, time-domain) and 1.92 s for the methods that performed estimation in the LFRTF domain (e.g., stand-alone, subspace, proposed). In the second set of the experiments, we used an equivalent segment length of the observations in time, which was 1.92 s only, to estimate the demixing parameters. This implied that the approaches that performed estimation in the HFRTF domain used only 30 TF samples while the methods that performed estimation in the LFRTF domain used 120 TF samples. Then, we evaluated and compared the separation performance. We also included the separation performance while using the different segment length of the observations.

The source separation performance was evaluated by using the signal-to-interference ratio (SIR) and the signal-to-distortion ratio between the original source and the separated signal. The SIR and SDR values were computed using the BSS_EVAL toolbox [189]. Each set of experiments was performed 60 times by using different combination of original sources and different arrangement of the microphones. The mean and standard deviation of the values are presented in the results.

3.5.2 Experimental Results

Fig. 3.3 and Fig. 3.4 illustrate the source separation performance (SIR and SDR, respectively) by using various techniques in the first set of experiments. Fig. 3.5 and Fig. 3.6 illustrate the source separation performance (SIR and SDR, respectively) in the second set of experiments. The first set of experiments computed the demixing parameters based on the first 120 TF observation samples while the second set of experiments estimated the demixing parameters based on the first 1.92 s observation samples (i.e., 30 TF samples for the HFRTF-domain methods while 120 TF samples for the LFRTF-domain methods). The blue, red and yellow circles indicate the mean SIR values of the first, second and third sources, respectively. The error bars represent the standard deviation based on the 60 trials.

The results show that the performance of the proposed method surpasses all the other methods in terms of the SIR metric and it achieves satisfactory performance. For the SDR metric, the performance of the proposed method is lower than the shortening method. This can be due to crossband filtering. In the problem formulation of the proposed method (3.1), we consider only band-to-band filters, hence the contributions of the neighbour frequency bins are neglected. Thus, distortion can be introduced. However, the performance of the proposed method is consistent over different room reverberation times.

We compare the computational complexity of the various BSS strategies. Table. 3.1 presents the average SIR performance of the three sources and the average computational time of each BSS method in the first set of experiments with different room reverberation time. The computational time of the stand-alone ICA approach is the lowest but the performance is the worst. The time-domain technique has low computational complexity as well but the separation performance is not as good as our proposed method and the shortening strategy. Our proposed method is third in terms of computational complexity and achieves the highest SIR performance.

We also include a comparison between the separation performance of

Figure 3.3: The separation performance (SIR) of various techniques, where the demixing operators were computed using the first 120 TF observation samples.

the competing BSS methods, which is indicated by SIR and SDR values in Fig. 3.7 and Fig. 3.8, respectively, while using different observation length to compute the demixing parameters, where the reverberation time was 0.15 s. Since the ST window length of the HFRTF-domain methods was four times longer than the one of the LFRFTF-domain approaches, the number of TF samples in the HFRTF-domain was four times less than the one in the LFRFTF-domain for an equivalent segment length of the time-domain observations. The results show that the proposed method outperformed the

Figure 3.4: The separation performance (SDR) of various techniques, where the demixing operators were computed using the first 120 TF observation samples.

other approaches on average for each experiment.

3.5.3 Discussion

The proposed method outperformed the existing subspace approaches [167, 186] in every set of experiments. This is because the principle of the proposed method is to find the demixing matrix. We include the demixing matrix in the contrast function while performing estimation, as shown in (3.26).

In contrast, the approaches [167, 186] aim to estimate the original sources

Table 3.1: The average SIR performance and computational time of the first set of the experiments by using various techniques.

Methods (s)	Reverb time , β	Average SIR (dB)	Average time (s)
Stand-alone	0.15	2.53	1.79
Cross-band		9.42	42.09
Shortening		17.45	12.37
Time-domain		7.19	5.42
Sorensen		1.38	12.88
Proposed		18.68	11.17
Stand-alone	0.30	2.20	2.10
Cross-band		3.94	42.83
Shortening		11.25	13.10
Time-domain		6.32	5.53
Sorensen		1.80	12.87
Proposed		18.79	11.04
Stand-alone	0.60	2.11	2.23
Cross-band		3.16	42.96
Shortening		8.15	13.54
Time-domain		7.20	5.95
Sorensen		1.74	12.79
Proposed		17.02	11.00

that are observed during the analysis period. Thus, the demixing matrix for entire signals has to be estimated explicitly. The estimation can be affected by the noise signals, which are present in the observation signals in the LFRTF domain. The noise signals can be due to the cross-band filtering effect, see Fig. 4.3 in Sec. 4.3.3.

In addition, the proposed method allows BSS to operate with a shorter of time-segment observation samples than other approaches. Since the esti-

mation is performed in the LFRTF domain, a sufficient number of samples is easier to obtain than is the case for approaches that perform the estimation in the HFRTF domain.

We briefly discuss the computational complexity of the various BSS strategies. In the discussion, we do not include the computational complexity of the ICA method as it is used in every BSS strategy. We consider only the additional processing steps. Let N_w and N_b denotes the length of the long window and short window, respectively. Furthermore, let $K = 2\Delta k + 1$ and $P = N_w/N_b$. For the cross-band method, it requires an additional $MLN_wK(N_b/2+1)$ matrix multiplications and $L(N_b/2+1)$ evaluations of the singular value decomposition (SVD) to obtain the demixing matrix, where the computational cost of each SVD operation is $M(L-1)^2K^2P^2$. For the shortening method, it requires an additional $MLN_bK(N_b/2+1)$ matrix multiplications and $L(N_b/2+1)$ evaluations of the singular value decomposition (SVD) to obtain the demixing matrix, where the computational cost is $M(L-1)^2K^2$.

Recall that N indicates the number of samples that are used in the estimation procedure and T indicates the length of the mixing filters in the TF domain. The additional processing step for the subspace method is to perform a SVD on a matrix with a size of NT by $LT^2 + N + T - 1$, where the computational cost is $NT(LT^2 + N + T - 1)^2$. The proposed method also requires an additional SVD evaluation, but the matrix size is NT by MT , where the computational cost is NM^2T^3 .

For the time-domain method, the computational complexity of the extra processing steps is much lower than that of the others and can be ignored. However, the demixing process is performed in the time-domain, which involve ML times of convolutions. The computational complexity of each convolution is approximately $N_x \log N_x$, where N_x denotes the total number of samples of the observation signal in the time domain. For all the other methods, the demixing process is performed in the TF domain. The total number of TF blocks is $2N_x/N_b - 1$, as the window was 50% overlap. For

Table 3.2: Summary of the computational cost for each BSS method.

Method	Computational cost of overall procedure
Standalone	$4N_x \log N_b + MLN_x$
Cross-band	$\frac{N_b}{2} (MLN_w K + ML^3 K^2 P^2) + 4N_x \log N_b + MLN_x$
Shortening	$\frac{N_b}{2} (MLN_b K + ML^3 K^2) + 4N_x \log N_b + MLN_x$
Time-domain	$MLN_x \log N_x$
Subspace	$NT(LT^2 + N + T - 1)^2 + 4N_x \log N_b + MLN_x$
Proposed	$NM^2T^3 + 4N_x \log N_b + MLN_x$

each block, a forward fast-Fourier transform operation is required to convert the observation data from the time domain into the frequency domain. Next, the transformed data is multiplied by the demixing matrix. Then, a backward FFT operation is used to transform the frequency-domain processed data into the time domain. The computational complexity of all these operations is approximately $4N_x \log N_b + MLN_x$.

The computational cost of each BSS strategy is summarized in Table. 3.2. The stand-alone method has the lowest computational cost while the time-domain approach has the second lowest computational cost as they do not require extra processing step to obtain the demixing matrix. By substituting the parameters of the experimental setup, the computational cost for the shortening, subspace and proposed methods is about 10^7 while the order of the computational cost of the cross-band method is about 10^8 . Hence, the theoretical results are consistent with the actual computational time, which is presented in 3.1.

3.6 Conclusion

In this chapter, we proposed a method that allows to facilitate the deconvolution of a convolutive mixture to a new instantaneous mixture. The proposed method is able to directly estimate the demixing parameters from the new

instantaneous mixture in the LFRTF domain. In a time-invariant system, the estimated demixing parameters are shareable between TF blocks. That is, the estimated demixing parameters can be applied instantly to the next incoming TF observation sample to obtain the separated signals. Moreover, shorter segments of the time-domain observation signals are only required for the estimation process. By using a two-threaded system, the user will experience less waiting time compared to the HFRTF-domain approaches before the separated signals are outputted.

The drawback of the proposed method is that the separation performance can be affected by choosing a wrong parameter to indicate the length of the mixing filters in the LFRTF domain. Future work may consider a minimization approach with constraints that specify the length of the mixing filters.

Figure 3.5: The separation performance (SIR) of various techniques, where the demixing operators were computed using the first 1.92 s observation samples.

Figure 3.6: The separation performance of various techniques, where the demixing operators were computed using the first 1.92 s observation samples.

Figure 3.7: The separation performance (SIR) of various techniques, where the demixing operators were computed using the observations with different length and the room reverberation time was 0.15 s.

Figure 3.8: The separation performance (SDR) of various techniques, where the demixing operators were computed using the observations with different length and the room reverberation time was 0.15 s.

Chapter 4

Low-Latency Approach for Convolutional Blind Source Separation

We present a low latency approach for blind source separation (BSS). BSS algorithms generally require a long window to estimate the demixing parameters. In traditional approaches, the long analysis window leads to a long algorithmic delay. Hence, traditional BSS approaches cannot be used in real-time systems. In contrast, our approach reduces the algorithmic delay independently of the window length used for estimation, while retaining separation performance. The new method exploits that the information about the sources provided by additional microphones can be traded against algorithmic delay. The method can be integrated with existing BSS algorithms and can be implemented in the time-domain or in the time-frequency domain. Our experimental results confirm the effectiveness of our approach.

4.1 Introduction

As mentioned in Sec. 2.6, the algorithmic latency of a real-time system is mostly due to the length of the window that is used to perform the short-time Fourier transform (STFT). Most of the existing time-frequency (TF)-domain blind source separation (BSS) methods rely on a long window length to obtain good separation performance, which is not applicable in real-time implementations.

To address the problem caused by the algorithmic delay, we proposed an approach [33], based on a cross-band filtering technique [14], that enables the estimation step to be performed in a high-frequency-resolution time-frequency (HFRTF) domain while the separation step is performed in a low-frequency-resolution time-frequency (LFRTF) domain. On the one hand, this method allows us to use a long window for the estimation, so that the instantaneous TF-domain BSS model is valid. On the other hand, we can perform the separation by using a short window to reduce the algorithm latency. We further enhanced the method by improving the computational efficiency using a non-iterative impulse response shortening strategy [32], where the details of the method are discussed in Chapter 6. It enables us to shorten the estimated mixing filters and facilitate the computation of the demixing operators in the LFRTF domain.

To reduce the system latency to the desired delay, the methods [32, 33] require a sufficient number of microphones to perform the estimation. The relation between the LFRTF-domain blocks and the full HFRTF-domain representation is described by a system of equations. To obtain a short delay some of these equations must be removed. To make the problem solvable, the number of equations must be larger or equal to the number of variables. The additional equations can be obtained by increasing the number of microphones as each microphone contributes unique constraints on the separated signals.

A time-domain implementation was also developed in [172] recently. In

the time domain, the algorithmic latency is caused by the so-called non-causal part of the filters, where the non-causal part refers to the components before the main peak of the filters. This method assumes that the mixing filters are usually sparse in a non-reverberant environment, so the contribution from the non-causal part of the demixing filters is relatively weak. Thus, they truncate the non-causal part and perform the separation by convolving the observations with the truncated demixing filters in the time domain.

As noted, the time-domain implementation [172] relies on the mixing filters to be sparse, which is less useful in a moderate or severe reverberant environment. In contrast, our method [32] first shortens the mixing filters in the time domain. This is done by selecting the scaling factors in the frequency domain to obtain the smoothest spectrum possible. Our method [32] exploits that the spectrum of a signal in the frequency domain becomes smoother when we zero-pad a signal in the time domain. After shortening the mixing filters, we truncate them in a way that the maximal information is retained. For example, we keep the important parts of the mixing filters, such as the main pulse and its adjacent responses, and discharge the parts that are far from the main pulse as their contributions are less significant. Then, we design the demixing operators based on the truncated mixing filters.

The main contribution of this chapter is to exploit that each microphone adds to the number of equations specifying a sequence of source signal vectors. As a result, blocks of shorter duration can be used for the source signal vector. Seen from another perspective, the added microphones allow us to omit non-causal filter components. Hence, the associated algorithmic delay can be eliminated. We find that the latency can be reduced to any desired delay by increasing the number of microphones. Our experiments show that excellent performance can be obtained with a small increase in microphones. While the approach builds on the principles of [33] the algorithms are different and performance is improved significantly. The number of microphones over the number required is drastically reduced.

This chapter first introduces a model based on frame theory, [188], to

accurately describe the signal representations in the TF domain. Next, we propose our approach to find low-latency demixing operators based on the BSS parameters that are estimated conventionally in the HFRTF domain. We present two implementations of the method: the first implementation operates in the time domain and the second implementation operates in a LFRTF domain. Then, we present the experimental results of the proposed method and comparisons with the other literature approaches.

4.2 Time-Frequency Transform

In this section, we first introduce an operator-based model to describe the short-time Fourier transform (STFT) based on frame theory. This facilitates the further derivation and analysis as linear operators can be represented by matrices.

4.2.1 Analysis and Synthesis Operator of the STFT

We consider all signals to be periodic with a length of N_x samples. Let us denote the length of the window by N_w , where $N_w = 2^p, p \in \mathbb{Z}^+$ and the block shift by N_l , where $N_l = N_w/2^q, q \in \mathbb{Z}$. Thus, N_w is always divisible by N_l . The synthesis and analysis operator for the STFT can be represented as $\Phi : \mathbb{C}^{QN_w} \rightarrow \mathbb{R}^{N_x}$ and $\tilde{\Phi} : \mathbb{R}^{N_x} \rightarrow \mathbb{C}^{QN_w}$, respectively, where Q represents the number of blocks in the TF domain and is an integer, i.e., $Q = \frac{N_x}{N_l} + \frac{N_w}{N_l} - 1$. Note that the length of the signals can always be zero-padded so that N_x is an integer multiple of N_l .

To describe the structure of the synthesis operator we first define $\mathbf{D}_{N_w} \in \mathbb{R}^{N_w \times N_w}$ as a diagonal matrix, where the diagonal entries consist of a window function. For instance, $\mathbf{D}_{N_w} = \text{diag}\{d(0), \dots, d(N_w - 1)\}$, where $\text{diag}\{\cdot\}$ denotes a diagonal operator. We define furthermore $\overline{\mathbf{D}}_{nN_l, N_w}$ as a sub-matrix of $\mathbf{D}_{N_w}$ from column one to column nN_l and $\underline{\mathbf{D}}_{nN_l, N_w}$ as a sub-matrix of $\mathbf{D}_{N_w}$ from column $N_w - nN_l + 1$ to the last column. We can now describe the

structure of the synthesis operator $\Phi \in \mathbb{C}^{N_x \times QN_w}$

$$\Phi = \begin{bmatrix} \varphi_{-\frac{N_w}{N_l}+1} & \cdots & \varphi_{Q+1-2\frac{N_w}{N_l}} & \varphi_{-\frac{N_w}{N_l}+1} & \cdots & \varphi_{-1} \end{bmatrix} \quad (4.1)$$

$$\varphi_p = \begin{bmatrix} \mathbf{0}_{N_w \times pN_l} & \mathcal{F}_{N_w} \mathbf{D}_{N_w} & \mathbf{0}_{N_w \times (N_x - N_w - pN_l)} \end{bmatrix}^*, \quad (4.2)$$

$$\varphi_{-p} = \begin{bmatrix} \mathcal{F}_{N_w} \mathbf{D}_{(\frac{N_w}{N_l}-p)N_l, N_w} & \mathbf{0}_{N_w, N_x - N_w} & \mathcal{F}_{N_w} \overline{\mathbf{D}}_{pN_l, N_w} \end{bmatrix}^*, \quad (4.3)$$

where $\cdot^*$ indicates the adjoint operator and $\mathcal{F}_{N_w}$ denotes an $N_w \times N_w$ discrete Fourier transform matrix. φ_p in (4.2) denotes the STFT synthesis operator at the p^{th} block in the TF domain. φ_{-p} in (4.3) accounts for the periodic structure of the data. We can use zero-padding to reduce artifacts resulting from the periodicity assumption. The synthesis operator is equivalent to a weighted overlap-add technique.

It is desirable that the columns of Φ form a tight frame, so that the canonical dual frame is equivalent to the Hermitian transpose of itself [188], e.g., $\tilde{\Phi} = \Phi^*$. This implies that the same window function must be used for both the synthesis and analysis windows. To form a tight frame, the window function $d(n)$ must satisfy the condition [14]:

$$\sum_{p=-\infty}^{\infty} d(n - pN_l)^2 = \frac{c}{N_w}, \quad \forall n, \quad (4.4)$$

where c is an arbitrary constant. For instance, a square root of Hann window [159] can be used for this purpose.

The advantage of using a tight frame is that perfect reconstruction is straightforward. For a tight frame we have

$$\Phi \Phi^* = c \mathbf{I}_{N_x}, \quad (4.5)$$

where $\mathbf{I}_{N_x}$ indicates an $N_x \times N_x$ identity matrix and c is the frame bound of Φ , which can be computed using (4.4). Note that the operators Φ and Φ^* preserve distances except for the constant factor c .

4.2.2 Time-Domain Signal Model

For a discrete linear time-invariant (LTI) system, a noiseless observation signal in the time-domain can be modelled as:

$$\bar{\mathbf{x}} = \mathbf{h} * \mathbf{s} \in \mathbb{R}^{(N_x+N_h-1) \times 1}, \quad (4.6)$$

where $*$ denotes a convolution operator, $\mathbf{h} \in \mathbb{R}^{N_h \times 1}$ indicates a mixing filter or a room impulse response, which is modelled by a finite impulse response (FIR) filter with a length of N_h , $\mathbf{s} \in \mathbb{R}^{N_x \times 1}$ indicates an original source. Consider that $\mathbf{s}$ is a periodic signal with a period of N_x , then (4.6) can be approximated as:

$$\mathbf{x} = \mathbf{h} \circledast \mathbf{s} \in \mathbb{R}^{N_x \times 1}, \quad (4.7)$$

where $\circledast$ denotes a circular convolution operator. As we have assumed that $\mathbf{s}$ is periodic, (4.7) is a truncated version of (4.6) and they are equivalent, i.e., $\mathbf{x}_i = \bar{\mathbf{x}}_i$, where $i \in [0, N_x - 1]$. This expression can also be written as:

$$\mathbf{x} = \mathbf{H}\mathbf{s}, \quad (4.8)$$

where $\mathbf{H} \in \mathbb{R}^{N_x \times N_x}$ represents a circulant matrix of $\bar{\mathbf{h}} = \begin{bmatrix} \mathbf{h}^T & \mathbf{0}_{1 \times (N_x - N_h)} \end{bmatrix}^T \in \mathbb{R}^{N_x \times 1}$.

Let us assume that the window function is properly scaled, so that the columns of Φ form a 1-tight frame, e.g., $c = 1$ in (4.5). Thus, the signal can then be perfectly recovered after the transform.

4.2.3 TF-Domain Signal Representations

The representation of (4.8) in the TF domain can be written as:

$$\Phi^* \mathbf{x} = \Phi^* \mathbf{H}\mathbf{s}. \quad (4.9)$$

Note that in this formulation the observation signal $\Phi^* \mathbf{x}$ in the TF domain is a vector $\mathbb{C}^{Q N_w \times 1}$. For instance, $(\Phi^* \mathbf{x})_{p N_w + k} = (\varphi_p^* \mathbf{x})_k$ indicates the observation signal at TF block index p and frequency bin index k . Note that

φ_p describes the time-frequency transform of the signal at the p^{th} block, i.e., the signal samples between index pN_l and $pN_l + N_w - 1$, which is defined in (4.2).

Using the property of the perfect reconstruction in (4.5), we can rewrite (4.9) as

$$\Phi^* \mathbf{x} = (\Phi^* \mathbf{H} \Phi) \Phi^* \mathbf{s}, \quad (4.10)$$

where $\Phi^* \mathbf{s}$ represents the original signal in the TF domain while $(\Phi^* \mathbf{H} \Phi)$ indicates the representation of $\mathbf{H}$ in the TF domain. Using this model, the signal can be described accurately in the TF domain. Moreover, it is applicable to any window length.

The matrix $(\Phi^* \mathbf{H} \Phi)$ has a block-Toeplitz structure as $\mathbf{H}$ is a circulant matrix. We can partition $\Phi^* \mathbf{x}$ in (4.10) into blocks. Note that the subscript of the STFT operator φ identifies a sub-matrix, as shown in (4.2) and (4.3). We can write (4.10) as:

$$\varphi_p^* \mathbf{x} = (\varphi_p^* \mathbf{H} \varphi_{p'}) \Phi^* \mathbf{s}, \quad (4.11)$$

where p indicates the block index in the TF domain, $(\varphi_p^* \mathbf{H} \varphi_{p'})_{k,k'}$ indicates the impulse response between $\mathbf{x}$ at TF block index p and frequency bin index k and $\mathbf{s}$ at TF block index p' and frequency bin index k' .

The TF-domain signal model in (4.11) describes the time-domain convolutive model (4.7) accurately and independently of the block length used. For our purposes it is useful to represent the signal in the low-frequency-resolution time-frequency (LFRTF) domain, where the window length is less than the room impulse response, i.e, $N_w < N_h$.

4.2.4 Instantaneous Approximation

The system identification problem in the time domain is often transformed into the frequency domain. This is because the convolution mixture can be approximated by a series of instantaneous mixtures in the frequency domain when the window length is much greater than the length of the room impulse

response (RIR), i.e., $N_w \gg N_h$, and this facilitates the computation and estimation processes [146]. In this subsection, we show that this is a special case of the model described in (4.11).

When $N_w \gg N_h$, this results in a high-frequency-resolution time-frequency (HFRTF) domain. Let us denote by $\tilde{\mathbf{H}} \in \mathbb{R}^{N_w \times N_w}$ and $\tilde{\mathbf{H}}_c \in \mathbb{R}^{N_w \times N_w}$ the convolution matrix and the circulant matrix of $\tilde{\mathbf{h}} = \begin{bmatrix} \mathbf{h} & \mathbf{0}_{1 \times (N_w - N_h)} \end{bmatrix}^T \in \mathbb{R}^{N_w \times 1}$, respectively. Note that $\tilde{\mathbf{H}}$ is also a sub-matrix of $\mathbf{H}$ if $N_x \gg N_w$, i.e.,

$$\tilde{\mathbf{H}} = \begin{bmatrix} \mathbf{H}_{i,i} & \cdots & \mathbf{H}_{i,i+N_w-1} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{i+N_w-1,i} & \cdots & \mathbf{H}_{i+N_w-1,i+N_w-1} \end{bmatrix}, \quad (4.12)$$

where $i \in [0, N_x - N_w]$. If $N_w \gg N_h$, then

$$\|\tilde{\mathbf{H}}\mathbf{z}\|_2 \approx \|\tilde{\mathbf{H}}_c\mathbf{z}\|_2, \quad (4.13)$$

where $\|\cdot\|_2$ indicates the $\mathcal{L}_2$ norm operator and $\mathbf{z} \in \mathbb{R}^{N_w \times 1}$ is a vector of signal. In addition, (4.13) becomes more accurate when N_w is getting larger.

As the rows of the discrete Fourier transform matrix $\mathcal{F}_{N_w}^*$ form the eigenvectors of the circular convolution operator $\tilde{\mathbf{H}}_c$ [188], we can decompose $\tilde{\mathbf{H}}_c$ into:

$$\tilde{\mathbf{H}}_c = \mathcal{F}_{N_w}^* \mathbf{\Lambda} \mathcal{F}_{N_w}, \quad (4.14)$$

where $\mathbf{\Lambda}$ is a diagonal matrix, e.g.,

$$\mathbf{\Lambda} = \text{diag} \left\{ \mathcal{F}_{N_w} \tilde{\mathbf{h}} \right\}. \quad (4.15)$$

By assuming $N_w \gg N_h$, and using (4.13) and (4.14) and focusing only on the parts of $\mathbf{H}$ that are fully affected by φ_p^* , we can approximate (4.11) as:

$$\varphi_p^* \mathbf{x} \approx \varphi_p^* \begin{bmatrix} \vdots \\ \cdots & \mathcal{F}_{N_w}^* \mathbf{\Lambda} \mathcal{F}_{N_w} & \cdots \\ \vdots \end{bmatrix} \Phi \Phi^* \mathbf{s}, \quad (4.16)$$

$$= \begin{bmatrix} \vdots \\ \cdots & \mathcal{F}_{N_w} \mathbf{D}_{N_w} \mathcal{F}_{N_w}^* \mathbf{\Lambda} \mathcal{F}_{N_w} & \cdots \\ \vdots \end{bmatrix} \mathbf{s}, \quad (4.17)$$

To simplify the expression in (4.17), [13] assume that the window $d(n)$ is long and smooth relative to the RIR $\mathbf{h}$, thus $d(n)$ is approximately constant over $\mathbf{h}$, i.e., $d(n-m)\mathbf{h}_m \approx d(n)\mathbf{h}_m$. Then, (4.17) can be approximated as:

$$\varphi_p^* \mathbf{x} \approx \begin{bmatrix} \vdots \\ \cdots \mathcal{F}_{N_w} \mathcal{F}_{N_w}^* \mathbf{\Lambda} \mathcal{F}_{N_w} \mathbf{D}_{N_w} \cdots \\ \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ \cdots \mathbf{I}_{N_w} \cdots \\ \vdots \end{bmatrix} \mathbf{s}, \quad (4.18)$$

$$= \mathbf{\Lambda} \varphi_p^* \mathbf{s}. \quad (4.19)$$

The model described in (4.19) is the so-called the multiplicative transfer function (MTF) approximation model. It has been widely used to solve system identification problems. These include, but are not limited to, blind source separation, beamforming, acoustic echo cancellation and speech dereverberation.

In contrast to the MTF model, the frame theory model (4.10) describes the TF representations of the signals accurately regardless of the window length. This is later being used to transform the mixing matrix from the high-frequency-resolution time-frequency (HFRTF) domain into the low-frequency-resolution time-frequency (LFRTF) domain. The details are discussed in Sec. 4.3.3.

4.3 Proposed Approach

The MTF approximation model (4.19) requires a long window to be sufficiently accurate, and this means the delay requirements for real-time systems cannot be met by straightforward BSS implementations. The delay requirements are high because the implementation is generally on a block-by-block basis and the window length must be at least twice the length of the room impulse response (RIR), i.e., $N_w \geq 2N_h$ for reasonable performance [145].

To minimize delay of BSS, a feedback architecture with infinite impulse response (IIR) filters [91, 144, 180] may seem most natural. In this case, the

separation can be performed with only one-tap delay. However, a stable IIR filter exists only if the mixing system is minimum phase [91, 92, 142]. The real RIR is usually non-minimum phase [91], and hence the implementation BSS with IIR filters is impractical.

In this section, we first briefly discuss the blind source separation (BSS) problem. Then we propose two low-latency approaches to solve the BSS problem that are based on the same fundamental principle. The first approach operates in the time domain while the second method operates in the low-frequency-resolution time-frequency (LFRTF) domain, where the window length is shorter than the RIR, i.e., $N_w \leq N_h$. Both methods design the demixing operators based on the RIR that is estimated in the HFRTF domain. This is motivated by the fact that many state-of-the-art approaches have been proposed to address the BSS problem in the HFRTF domain. For instance, the well-known fast independent component analysis (ICA) approach [69, 71] can be applied individually in every frequency bin in the HFRTF domain to estimate the mixing matrix.

We now provide the essence of the algorithms. Consider a sequence of known microphone vectors (dimensionality equal to the number of microphones) and a longer sequence of unknown source vectors (dimensionality equal to the number of sources) that determine the sequence of microphone vectors. The system identification performed by BSS provides us with a linear relationship between the source and microphone sequences. This relation corresponds to a set of linear equations. For a causal system the number of microphone vectors must be smaller than the number of source vectors. The main contribution of this work is to exploit that the resulting system of equations is fully specified if the number of microphones is sufficiently high. We find that the algorithmic delay can be reduced to any desired rate by increasing the number of microphones.

4.3.1 Blind Source Separation Model

Let denote by $\mathbf{x}^m$ the m^{th} observation signal and by $\mathbf{s}^l$ the l^{th} original source. Recall that in the HFRTF domain a LTI BSS problem with M microphones and L original sources can be written as:

$$\mathbf{x}_{p,k} = \mathbf{A}_k \mathbf{s}_{p,k}, \quad (4.20)$$

where

$$\mathbf{x}_{p,k} = \left[(\varphi_p^* \mathbf{x}^1)_k \quad \cdots \quad (\varphi_p^* \mathbf{x}^M)_k \right]^T, \quad (4.21)$$

$$\mathbf{s}_{p,k} = \left[(\varphi_p^* \mathbf{s}^1)_k \quad \cdots \quad (\varphi_p^* \mathbf{s}^L)_k \right]^T, \quad (4.22)$$

$$\mathbf{A}_k = \begin{bmatrix} \Lambda_{k,k}^{11} & \cdots & \Lambda_{k,k}^{1L} \\ \vdots & \ddots & \vdots \\ \Lambda_{k,k}^{M1} & \cdots & \Lambda_{k,k}^{ML} \end{bmatrix}, \quad (4.23)$$

and $\Lambda_{k,k}^{ml}$ denotes the multiplicative transfer function (MTF) between the m^{th} observation and the l^{th} original source in frequency bin index k .

To estimate the original sources, we want to find a demixing matrix $\mathbf{W}_k$ in every frequency bin, such that

$$\mathbf{y}_{p,k} = \mathbf{W}_k^H \mathbf{x}_{p,k}, \quad (4.24)$$

where $\mathbf{y}_{p,k}$ denotes the estimated original sources at TF index p and frequency bin k and $\cdot^H$ denotes a Hermitian transpose. The observation signals can then be reconstructed as:

$$\mathbf{x}_{p,k} = \hat{\mathbf{A}}_k \mathbf{y}_{p,k}, \quad (4.25)$$

$$\hat{\mathbf{A}}_k = \begin{bmatrix} \hat{\Lambda}_{k,k}^{11} & \cdots & \hat{\Lambda}_{k,k}^{1L} \\ \vdots & \ddots & \vdots \\ \hat{\Lambda}_{k,k}^{M1} & \cdots & \hat{\Lambda}_{k,k}^{ML} \end{bmatrix}, \quad (4.26)$$

where $\hat{\mathbf{A}}_k$ is the estimated mixing matrix, e.g., $\hat{\mathbf{A}}_k = \mathbf{W}_k^{H\dagger}$, and $\cdot^\dagger$ denotes a pseudo-inverse.

Most state-of-the-art approaches solve the instantaneous BSS problems in the HFRTF domain and achieve satisfactory separation performance. Most of these approaches are based on one of the two main separation principles. The first [20, 28, 36, 69, 71] computes the demixing matrix by the assumption that the separated sources are uncorrelated or independent. The second [1, 2, 76, 97, 148, 154, 198] estimates the mixing matrix in every frequency bin by presuming only one source is active in a certain period. Then, the demixing matrix is designed based on the estimated mixing matrix.

4.3.2 Time-domain Approach

In this subsection, we provide a time-domain approach to perform BSS in a low-latency manner. The main idea is that first the estimation is performed using previous inputs with existing BSS methods in the HFRTF domain. Then, we focus on designing the demixing operator in the time domain to perform the actual separation operation for the future inputs in a low-latency manner based on the estimated HFRTF-domain model.

As noted, we assume that all the RIR $\hat{\mathbf{\Lambda}}_{k,k}^{ml}, m \in [1, M], l \in [1, L], k \in [0, N_w - 1]$ in (4.26) have already been estimated in the HFRTF domain. Then, the time-domain observation signal can be reconstructed as

$$\mathbf{x}^m = \mathbf{\Phi}\mathbf{\Phi}^* \mathbf{x}^m \approx \sum_{l=1}^L \mathbf{\Phi}\ddot{\mathbf{\Lambda}}^{ml}\mathbf{\Phi}^* \mathbf{y}^l, \quad (4.27)$$

where $\ddot{\mathbf{\Lambda}}^{ml} = \hat{\mathbf{\Lambda}}^{ml} \otimes \mathbf{I}_{Q \times Q}$, $\otimes$ defines a Kronecker product and $Q = \frac{N_x}{N_l} + \frac{N_w}{N_l} - 1$. Note that N_x, N_w, N_l indicate the length of the signal, the window length and the block shift, respectively.

The matrix $\mathbf{\Phi}\ddot{\mathbf{\Lambda}}^{ml}\mathbf{\Phi}^* = \hat{\mathbf{H}}^{ml}$ in (4.27) has a block-circulant structure with a block size of $N_l \times (2N_w - N_l)$ as the RIR is stationary in a LTI system. Note that if $N_l = 1$, then $\mathbf{\Phi}\ddot{\mathbf{\Lambda}}^{ml}\mathbf{\Phi}^*$ will be a circulant matrix. An example is illustrated in Fig. 4.1 by using an exponential decay function with additive white Gaussian noise to simulate the time-domain RIR with the following parameters: $N_x = 256, N_h = 32, N_w = 64$ and $N_l = 16$. We can see that a

Figure 4.1: A matrix representation of a synthetic room impulse response $\Phi \ddot{\Lambda}^{ml} \Phi^*$ in the time domain using the model in (4.27) with $m = 1$, $l = 1$, $N_x = 256$, $N_h = 32$, $N_w = 64$ and $N_l = 16$. The y-axis and x-axis represent the rows and the columns of the matrix $\Phi \ddot{\Lambda}^{ml} \Phi^*$, respectively. The white coloured parts indicate negative infinity values.

block of the observation contains the contributions from multiple blocks of the original source.

Our goal is to solve (4.27) for a set of rows of the source signals $\mathbf{y}^l$ that corresponds to a local time interval, and do so simultaneously for all sources and microphones. At first sight this seems impossible as the sequence of source signal vectors is longer than the corresponding sequence of microphone vectors. However, by using a sufficient number of microphones we obtain a sufficient number of equations to specify the source signals.

In the following we first define a suitable set of linear equations that specify the desired segments of $\mathbf{y}^l$. Let us denote by N_f the length of the acquired signals (or the data acquisition-block length in a practical system)

and by p the time-block index. For instance, at time-block index p , the observation signals between pN_f and $\bar{n}_p$ are captured, where $\bar{n}_p = (p+1)N_f$ labels the endpoint of the signal acquisition corresponds to time-block index p . Furthermore, let $\underline{n}_p = \bar{n}_p - N_w$ be the time index of the earliest observed microphone data that are affected by the source signal at time-block index p , $\underline{N}_p = \bar{n}_p - 2N_w + N_f$ be the time index of the earliest source signal data that is included in the microphone data at time-block index $\underline{n}_p$ and $\bar{N}_p = \bar{n}_p + N_w - N_f$ be the time index of the first not-yet observed microphone data. Note that the signals from time index $\underline{n}_p$ to $\bar{n}_p - 1$ are associated with the causal parts of the filters while the signals from time index $\bar{n}_p$ to $\bar{N}_p - 1$ are associated with the non-causal parts of the filters, correspond to time-block index p . Fig. 4.2 illustrates the relationship between $\underline{n}_p$, $\bar{n}_p$, $\underline{N}_p$ and $\bar{N}_p$, where $p = 7$, $N_f = 16$ and $N_w = 64$.

The source signal $\mathbf{y}^l$ associated with the causal parts can be specified by the previous observation inputs, e.g., $\begin{bmatrix} x_{N_p}^m & \cdots & x_{\bar{n}_p-1}^m \end{bmatrix}^T$, $m \in [1, M]$. To handle the signals associated with the non-causal parts of the filters, instead of acquiring the future observation inputs as is done in [14], we add more microphones, so that there is a sufficient number of equations in the system.

The window length N_w is based on the response length and used in the HFRTF domain analysis. However, the data acquisition-block length N_f is associated with the low-delay separation operation only. Recall that M indicates the number of microphones and L indicates the number of original sources. We partition the observation signals into multiple blocks and rewrite (4.27) on a block-by-block basis by stacking all the M observation signals in a matrix form:

$$\mathbf{X}_p = \mathcal{A}_p \mathbf{Y}_p, \quad (4.28)$$

where

$$\mathbf{X}_p = \begin{bmatrix} \mathbf{x}_p^{1T} & \cdots & \mathbf{x}_p^{MT} \end{bmatrix}^T \in \mathbb{R}^{MN_w \times 1}, \quad (4.29)$$

$$\mathbf{x}_p^m = \begin{bmatrix} x_{\underline{n}_p}^m & \cdots & x_{\bar{n}_p-1}^m \end{bmatrix}^T \in \mathbb{R}^{N_w \times 1}, \quad (4.30)$$

$$\mathbf{Y}_p = \begin{bmatrix} \mathbf{y}_p^{1T} & \cdots & \mathbf{y}_p^{LT} \end{bmatrix}^T \in \mathbb{R}^{L(3N_w-2N_f) \times 1}, \quad (4.31)$$

$$\mathbf{y}_p^l = \begin{bmatrix} y_{\text{mod}\{\underline{N}_p, N_x\}}^l & \cdots & y_{\text{mod}\{\bar{N}_p-1, N_x\}}^l \end{bmatrix}^T \in \mathbb{R}^{(3N_w-2N_f) \times 1}, \quad (4.32)$$

$$\mathcal{A}_p = \begin{bmatrix} \mathcal{A}_p^{11} & \cdots & \mathcal{A}_p^{1L} \\ \vdots & \ddots & \vdots \\ \mathcal{A}_p^{M1} & \cdots & \mathcal{A}_p^{ML} \end{bmatrix} \in \mathbb{R}^{MN_w \times L(3N_w-2N_f)}, \quad (4.33)$$

$$\mathcal{A}_p^{ml} = \begin{bmatrix} \hat{\mathbf{H}}_{\underline{n}_p, \text{mod}\{\underline{N}_p, N_w\}}^{ml} & \cdots & \hat{\mathbf{H}}_{\bar{n}_p, \text{mod}\{\bar{N}_p, N_w\}}^{ml} \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{H}}_{\bar{n}_p-1, \text{mod}\{\underline{N}_p, N_w\}}^{ml} & \cdots & \hat{\mathbf{H}}_{\bar{n}_p-1, \text{mod}\{\bar{N}_p, N_w\}}^{ml} \end{bmatrix} \in \mathbb{R}^{N_w \times (3N_w-2N_f)}, \quad (4.34)$$

where p indicates the time-block index and $\text{mod}\{\cdot\}$ indicates a modulus operator.

The idea of (4.28) is to formulate a standard BSS problem by combining all the observation data on a block-by-block basis. In (4.30), $\mathbf{x}_p^m$ denotes the m^{th} observation signal between time index $\underline{n}_p$ and $\bar{n}_p - 1$. $\mathbf{X}_p$ in (4.29) denotes an observation signal vector that concatenates all the M observation signals that are defined in (4.30). In (4.32), $\mathbf{y}_p^l$ denotes the l^{th} separated signal between time index $\underline{N}_p$ and $\bar{N}_p - 1$. $\mathbf{Y}_p$ in (4.31) denotes a vector of separated sources corresponding to the observations in $\mathbf{X}_p$, which concatenates all the L original sources. The matrix $\mathcal{A}_p$ in (4.33) represents an overall mixing matrix that contains all the room impulse responses $\mathcal{A}_p^{ml}$ between the observation signals $\mathbf{x}_p^m$ and the original sources $\mathbf{y}_p^l$, where $m = 1, \dots, M$ and $l = 1, \dots, L$. An example of the room impulse response $\mathcal{A}_p^{ml}$ is illustrated in Fig. 4.2.

In practice, the data acquisition-block length N_f is desired to be small, so that the system can operate with minimal latency. Note that $\mathcal{A}_p$ does

Figure 4.2: The relationship between $\underline{n}_p$, $\bar{n}_p$, $\underline{N}_p$ and $\bar{N}_p$ as shown in Fig. 4.1, where $p = 7$, $N_f = 16$ and $N_w = 64$.

not vary with p due to the block-circulant structure. Hence, (4.28) can be simplified to:

$$\mathbf{X}_p = \mathcal{A}\mathbf{Y}_p, \quad (4.35)$$

where $\mathcal{A} = \mathcal{A}_n, n \in \mathbb{Z}$.

To perform source separation in a low-latency manner (i.e., with a short data acquisition-block length), we want to find a demixing operator $\mathcal{W}$ such that

$$\tilde{\mathbf{Y}}_p = \mathcal{W}\mathbf{X}_p = \mathcal{W}\mathcal{A}\mathbf{Y}_p, \quad (4.36)$$

where $\tilde{\mathbf{Y}}_p$ denotes the recovered sources. A straightforward solution for $\mathcal{W}$ is the left pseudo-inverse of $\mathcal{A}$, i.e.,

$$\mathcal{W} = (\mathcal{A}^T \mathcal{A})^{-1} \mathcal{A}^T, \quad (4.37)$$

which is also a least squares solution for (4.36).

The mixing operator in (4.28) can be ill-conditioned, that is, $\mathcal{A}$ is close to a singular matrix. In this case, a small error induced in the HFRTF domain solution may lead to a large error to the estimated sources $\tilde{\mathbf{Y}}_p$ in the low-latency domain [85]. Therefore, instead of using the pseudo-inverse solution, we provide an alternative method to compute the demixing operator $\mathcal{W}$.

Extracting a source from the observations is equivalent to removing its interference signals. Thus, to extract the l^{th} estimated source, we want to eliminate the contributions from the other sources. This can be done by finding a demixing vector $\mathbf{w}^l$, such that, $(\mathbf{w}^l)^T \mathcal{A}^{\setminus l} = \mathbf{0}$, where $\mathcal{A}^{\setminus l}$ is spanned by all the columns associated with the other sources in $\mathcal{A}$, i.e.,

$$\mathcal{A}^{\setminus l} = \begin{bmatrix} \mathcal{A}_p^{11} & \dots & \mathcal{A}_p^{1(l-1)} & \mathcal{A}_p^{1(l+1)} & \dots & \mathcal{A}_p^{1L} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{A}_p^{M1} & \dots & \mathcal{A}_p^{M(l-1)} & \mathcal{A}_p^{M(l+1)} & \dots & \mathcal{A}_p^{ML} \end{bmatrix} \in \mathbb{R}^{MN_w \times (L-1)(3N_w - 2N_f)}. \quad (4.38)$$

To retain the maximum energy of the source of interest $\mathbf{y}_p^l$, we select the feasible demixing vector $\mathbf{w}^l$ that is nearest in direction to the direction of the source of interest $\mathcal{A}$, where

$$\mathcal{A}^l = \left[(\mathcal{A}_p^{1l})^T \quad \dots \quad (\mathcal{A}_p^{Ml})^T \right]^T. \quad (4.39)$$

Thus, the demixing vector $\mathbf{w}^l$ can be found as

$$\mathbf{w}^l = \arg \max_{\mathbf{w}^l} \left\{ \mathbf{w}^{lT} \mathcal{A}^l \mathcal{A}^{lT} \mathbf{w}^l \mid \mathbf{w}^l \in \mathcal{N}(\mathcal{A}^{\setminus l}) \right\}, \quad (4.40)$$

where $\mathcal{N}(Z)$ denotes the nullspace of Z . Note that the distortion of the recovered signal using the demixing vectors in (4.40) can be larger than the one in the HFRTF domain due to extra processing steps. The distortion affects the audio quality but not the source separation.

We now provide the singular value decomposition (SVD) method to solve (4.40) without considering the constraint. First, we rewrite the maximization problem as

$$\mathbf{w}^l = \arg \max_{\mathbf{w}^l} \left\{ \mathbf{w}^{lT} \mathbf{U}^l \mathbf{U}^{lT} \mathcal{A}^l \mathcal{A}^{lT} \mathbf{U}^l \mathbf{U}^{lT} \mathbf{w}^l \right\}, \quad (4.41)$$

where $\mathbf{U}^{lT}$ maps towards the null-space of $\mathcal{A}^{\setminus l}$. $\mathbf{U}^l$ can be computed by performing SVD on $\mathcal{A}^{\setminus l}$, where the columns of $\mathbf{U}^l$ consist of the eigenvectors that associated with zero eigenvalues. Since $\mathbf{U}^{lT}\mathbf{U}^l = \mathbf{I}$, then $\mathbf{U}^l\mathbf{U}^{lT}$ defines a projection. Hence the solution $\mathbf{w}^l$ will be projected to the null space of $\mathcal{A}^{\setminus l}$. This shows that the constraint in (4.40) is automatically taken care in (4.41).

To solve (4.41), we can perform a SVD on $\mathbf{U}^{lT}\mathcal{A}^l\mathcal{A}^{lT}\mathbf{U}^l$. Then, we select the eigenvector with the largest eigenvalue, which we denote as $\mathbf{w}_{U^l}^l$. Finally we find $\mathbf{w}^l$ as

$$\mathbf{w}^l = \mathbf{U}^l \mathbf{w}_{U^l}^l. \quad (4.42)$$

The solution in (4.42) shows that $\mathbf{w}^l$ lives in the null space of $(\mathcal{A}^{\setminus l})^T$. Hence, the row rank of $\mathcal{A}^{\setminus l}$ must be greater than its column rank, i.e., $MN_w > (L-1)(3N_w - 2N_f)$ (assuming all the rows are independent to each other). Otherwise, $\mathbf{U}^l$ will not exist. In this case, we can choose $\mathbf{U}^l$ to be the eigenvector of $\mathcal{A}^{\setminus l}\mathcal{A}^{\setminus lT}$ that is associated with the smallest eigenvalue.

As with the time-domain method, the proposed LFRTF-domain method allows us to trade between the number of microphones M and the system latency for a given source separation performance. The system latency is due to the non-causal parts of the filters and the length of the non-causal parts is affected by the data acquisition-block length N_f . To compensate for the reduction in the latency, which corresponds to a reduction in N_f , we can increase the number of microphones M to compute the demixing operator $\mathbf{w}^l$ as long as the following condition is satisfied

$$\frac{N_f}{N_w} > \frac{3}{2} - \frac{M}{2(L-1)}, \quad (4.43)$$

where $\frac{N_f}{N_w}$ represents the factor of the delay in the HFRTF-domain method and L is the number of the original sources. Ideally, $\frac{N_f}{N_w} = \frac{1}{2^n}$, where $n \in \mathbb{Z}$. For example, to separate three original source signals we need only six microphones, then the algorithmic delay can be reduced to any desired length.

4.3.3 Low-Frequency-Resolution Time-Frequency (LFRTF)-Domain Approach

In this subsection, we provide an alternate low-latency approach to perform BSS in the low-frequency-resolution time-frequency (LFRTF) domain. To decrease the system latency, the window length must be reduced. In practice, the window length must be shorter than the room impulse response (RIR) and this results in a LFRTF domain. Hence, the multiplicative transfer function (MTF) approximation in (4.19) is no longer valid in the LFRTF domain. Similarly to the proposed time-domain approach, the LFRTF domain method first computes the HFRTF domain mixing matrix based on the previous inputs. Next, we use (4.11) to compute the LFRTF-domain representation of the HFRTF-domain mixing matrix. Then, we design the LFRTF-domain demixing operator to perform the source separation for future inputs. This method has the advantage that the separation process can be performed in every frequency bin in the LFRTF domain. Instead of computing the full-band time-domain demixing operator (in practice, this requires large computational effort), we compute the narrow-band demixing matrix in every frequency bin. Hence, the computational effort is distributed among every frequency bin.

Let us denote by N_b the length of the LFRTF window (whereas N_w is the HFRTF window) and by N_a the shift length of the LFRTF. Furthermore, let $Q_s = \frac{N_x}{N_a} + \frac{N_b}{N_a} - 1$, where N_x is an integer multiple of N_a . We define $\Psi : \mathbb{C}^{Q_s N_b} \rightarrow \mathbb{R}^{N_x}$ as a synthesis operator that transforms data from the LFRTF domain into the time domain while $\Psi^* : \mathbb{R}^{N_x} \rightarrow \mathbb{C}^{Q_s N_b}$ as an analysis operator that transforms data from the time domain into the LFRTF domain.

We convert (4.27) into the LFRTF domain and apply the property of perfect recovery in (4.5). This yields:

$$\Psi^* \mathbf{x}^m = \sum_{l=1}^L \mathbf{G}^{ml} \Psi^* \mathbf{y}^l, \quad (4.44)$$

where

$$\mathbf{G}^{ml} = \Psi^* \hat{\mathbf{H}}^{ml} \Psi, \quad (4.45)$$

$$\hat{\mathbf{H}}^{ml} = \Phi \ddot{\mathbf{A}}^{ml} \Phi^*. \quad (4.46)$$

$\mathbf{G}^{ml}$ is a matrix representation of the time-domain operator $\hat{\mathbf{H}}^{ml}$ in the LFRTF domain. $\mathbf{G}^{ml}$ has a block-circulant structure, which is inherited from $\hat{\mathbf{H}}^{ml}$. An example of the TF representations of Fig. 4.1 is illustrated in Fig. 4.3 with two extra parameters, which are $N_b = 16$ and $N_a = 8$.

Figure 4.3: A matrix representation of a synthetic room impulse response $\Psi^* \hat{\mathbf{H}}^{ml} \Psi$ in the LFRTF domain using the model in (4.44) with $m = 1$, $l = 1$, $N_x = 256$, $N_h = 32$, $N_w = 64$, $N_l = 16$, $N_b = 16$ and $N_a = 8$. The y-axis and x-axis represent the rows and the columns of the matrix $\Psi^* \hat{\mathbf{H}}^{ml} \Psi$, respectively. The white coloured parts indicate negative infinity values.

The block-circulant structure of $\mathbf{G}^{ml}$ facilitates the estimation of $\Psi^* \mathbf{s}$ as we can assume that the mixing operators in the LFRTF domain are static in every TF block. So, the observation at TF-index $(p, \cdot)$ can be expressed

as:

$$\psi_p^* \mathbf{x}^m = \sum_{l=1}^L \psi_p^* \hat{\mathbf{H}}^{ml} \Psi \Psi^* \mathbf{y}^l, \quad (4.47)$$

It is worth noting that $(\psi_p^* \mathbf{H} \psi_{p'})_{k,k'}$ indicates the impulse response between the l^{th} original source at TF-index (p', k') and the m^{th} observation signal at TF-index (p, k) . As illustrated in Fig. 4.3, the contribution of $(\psi_p^* \hat{\mathbf{H}}^{ml} \tilde{\psi}_{p'})_{k,k'}$ becomes smaller when $|k - k'|$ gets bigger as it is referred as a band-to-band filter when $k = k'$ but a cross-band filter when $k \neq k'$ [14]. Hence, $\psi_p^* \hat{\mathbf{H}}^{ml} \tilde{\psi}_p$ is a sparser matrix compared to $\Phi \ddot{\mathbf{A}}^{ml} \Phi^*$.

From Fig. 4.3, we know that the observation at TF block p and frequency bin index k , $(\psi^* \mathbf{x}^m)_k$, contains the contributions of the original sources from neighbouring blocks in time (we distinguish causal and non-causal coefficients) and neighboring frequency bins. We also find that the number of blocks associated with the causal coefficients $\underline{P}$ equal in number to the number of blocks associated with the non-causal coefficients $\bar{P}$, i.e., $\underline{P} = \bar{P} = P - 1$, where $P = \frac{N_w}{N_a}$ is the number of LFRFT window shifts per HFRTF window and N_w is divisible by N_a . Let us denote by Δk the number of neighbour frequency bins and $K = 2\Delta k + 1$. By iterating through all the frequency bins and considering only Δk significant neighbour frequency bins, we can stack all the M observations in (4.47) and write:

$$\mathbf{X}_{p,k} = \Gamma_k \mathbf{Y}_{p,k}, \quad (4.48)$$

where

$$\mathbf{X}_{p,k} = \begin{bmatrix} \tilde{\mathbf{X}}_{p,k}^1 & \dots & \tilde{\mathbf{X}}_{p,k}^M \end{bmatrix}^T \in \mathbb{C}^{MPK \times 1}, \quad (4.49)$$

$$\tilde{\mathbf{X}}_{p,k}^m = \begin{bmatrix} \mathbf{X}_{p-\underline{P},k}^m & \dots & \mathbf{X}_{p,k}^m \end{bmatrix}^T \in \mathbb{C}^{PK \times 1}, \quad (4.50)$$

$$\mathbf{X}_{p',k}^m = \begin{bmatrix} (\psi_{p'}^* \mathbf{x}^m)_{k-\Delta k} & \dots & (\psi_{p'}^* \mathbf{x}^m)_{k+\Delta k} \end{bmatrix}^T \in \mathbb{C}^{K \times 1}, \quad (4.51)$$

$$(4.52)$$

$$\mathbf{Y}_{p,k} = \begin{bmatrix} \tilde{\mathbf{Y}}_{p,k}^{1\top} & \cdots & \tilde{\mathbf{Y}}_{p,k}^{L\top} \end{bmatrix}^\top \in \mathbb{C}^{L(3P-2)K \times 1}, \quad (4.53)$$

$$\tilde{\mathbf{Y}}_{p,k}^l = \begin{bmatrix} \mathbf{Y}_{p-2\underline{P},k}^{l\top} & \cdots & \mathbf{Y}_{p+\overline{P},k}^{l\top} \end{bmatrix}^\top \in \mathbb{C}^{(3P-2)K \times 1}, \quad (4.54)$$

$$\mathbf{Y}_{p',k}^l = \begin{bmatrix} (\psi_{p'}^* \mathbf{y}^l)_{k-2\Delta k} & \cdots & (\psi_{p'}^* \mathbf{y}^l)_{k+2\Delta k} \end{bmatrix}^\top \in \mathbb{C}^{K \times 1}, \quad (4.55)$$

$$\mathbf{\Gamma}_k = \begin{bmatrix} \tilde{\mathbf{G}}_k^{11} & \cdots & \tilde{\mathbf{G}}_k^{1L} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{G}}_k^{M1} & \cdots & \tilde{\mathbf{G}}_k^{ML} \end{bmatrix} \in \mathbb{C}^{MPK \times L(3P-2)K}, \quad (4.56)$$

$$\tilde{\mathbf{G}}_k^{ml} = \begin{bmatrix} \tilde{\mathbf{G}}_{p-\underline{P},p-2\underline{P},k}^{ml} & \cdots & \tilde{\mathbf{G}}_{p-\underline{P},p+\overline{P},k}^{ml} \\ \vdots & \ddots & \vdots \\ \tilde{\mathbf{G}}_{p,p-2\underline{P},k}^{ml} & \cdots & \tilde{\mathbf{G}}_{p,p+\overline{P},k}^{ml} \end{bmatrix} \in \mathbb{C}^{PK \times (3P-2)K}, \quad (4.57)$$

$$\tilde{\mathbf{G}}_{q,q',k}^{ml} = \begin{bmatrix} \hat{\mathbf{G}}_{q,q',k-\Delta k,k-\Delta k}^{ml} & \cdots & \hat{\mathbf{G}}_{q,q',k-\Delta k,k+\Delta k}^{ml} \\ \vdots & \ddots & \vdots \\ \hat{\mathbf{G}}_{q,q',k+\Delta k,k-\Delta k}^{ml} & \cdots & \hat{\mathbf{G}}_{q,q',k+\Delta k,k+\Delta k}^{ml} \end{bmatrix} \in \mathbb{C}^{K \times K}, \quad (4.58)$$

$$\hat{\mathbf{G}}_{q,q',k,k'}^{ml} = \left(\psi_q^* \hat{\mathbf{H}}^{ml} \tilde{\psi}_{q'} \right)_{k,k'}. \quad (4.59)$$

The formulation in (4.48) is equivalent to a standard BSS problem in the TF domain by combining all the observation data. In (4.51), $\mathbf{X}_{p',k}^m$ denotes the m^{th} observation signal between frequencies $k - \Delta k$ and $k + \Delta k$ at time block p' in the LFRTF domain while $\tilde{\mathbf{X}}_{p,k}^m$ in (4.50) indicates a vector of $\mathbf{X}_{p',k}^m$ between time blocks $p - \underline{P}$ and p in the LFRTF domain. $\mathbf{X}_{p,k}$ in (4.49) denotes an observation signal vector that concatenates all the M observation signals that are defined in (4.50) and (4.51) in the LFRTF domain. In (4.55), $\mathbf{Y}_{p',k}^l$ denotes the l^{th} original source between frequencies $k - \Delta k$ and $k + \Delta k$ at time block p' in the LFRTF domain while $\tilde{\mathbf{Y}}_{p,k}^l$ in (4.50) indicates a vector of $\mathbf{Y}_{p',k}^l$ between time blocks $p - \underline{P}$ and p in the LFRTF domain. $\mathbf{Y}_{p,k}$ in (4.49) denotes a source signal vector that concatenates all the L original sources that are defined in (4.54) and (4.55) in the LFRTF domain. In (4.56), $\mathbf{\Gamma}_k$ represents an overall mixing matrix in the LFRTF domain that contains all the transfer functions between the observation signals $\tilde{\mathbf{X}}_{p,k}^m$ and the originals sources $\tilde{\mathbf{Y}}_{p,k}^l$, where $m = 1, \dots, M$ and $l = 1, \dots, L$.

Similar to the time-domain approach, the l^{th} estimated original source $(\psi_p^* \tilde{\mathbf{y}}^l)_k$ in the LFRTF domain can be extracted by finding a demixing vector $\mathbf{w}_k^l$ to cancel the contributions of the columns in $\mathbf{\Gamma}_k$ that are associated with the other interference signals, i.e.,

$$\mathbf{\Gamma}_k^{\setminus l} = \begin{bmatrix} \tilde{\mathbf{G}}_k^{11} & \dots & \tilde{\mathbf{G}}_k^{1(l-1)} & \tilde{\mathbf{G}}_k^{1(l+1)} & \dots & \tilde{\mathbf{G}}_k^{1L} \\ \vdots & \ddots & \vdots & & & \\ \tilde{\mathbf{G}}_k^{M1} & \dots & \tilde{\mathbf{G}}_k^{M(l-1)} & \tilde{\mathbf{G}}_k^{M(l+1)} & \dots & \tilde{\mathbf{G}}_k^{ML} \end{bmatrix} \in \mathbb{C}^{MPK \times (L-1)(3P-2)K}. \quad (4.60)$$

We also want the demixing vector $\mathbf{w}_k^l$ to points in the direction of the vector that spans the subspace of the source of interest $\mathbf{\Gamma}_k^l$, so as to retain most of the energy of the source of interest $\mathbf{Y}_{p,k}^l$, where

$$\mathbf{\Gamma}_k^l = \begin{bmatrix} \tilde{G}_k^{1l \text{ T}} & \dots & \tilde{G}_k^{Ml \text{ T}} \end{bmatrix}^{\text{T}}. \quad (4.61)$$

By using the same approach as the time-domain approach, we define $\mathbf{U}_k^l$ be the null space of $\mathbf{\Gamma}_k^{\setminus l \text{ H}}$. Then, the demixing vector $\mathbf{w}_k^l$ can be found by maximizing the following cost function:

$$\mathbf{w}_k^l = \arg \max_{\mathbf{w}_k^l} \left\{ \mathbf{w}_k^{l \text{ H}} \mathbf{U}_k^l \mathbf{U}_k^{l \text{ H}} \mathbf{\Gamma}_k^l \mathbf{\Gamma}_k^{l \text{ H}} \mathbf{U}_k^l \mathbf{U}_k^{l \text{ H}} \mathbf{w}_k^l \right\}, \quad (4.62)$$

where $\mathbf{U}_k^{l \text{ H}}$ is a unitary matrix that maps to the null space of the subspace of the interference signals, $\mathcal{N}(\mathbf{\Gamma}_k^{\setminus l \text{ H}})$. Equation (4.62) is equivalent to finding the eigenvector that is associated with the largest eigenvalue of $\mathbf{U}_k^{l \text{ H}} \mathbf{\Gamma}_k^l \mathbf{\Gamma}_k^{l \text{ H}} \mathbf{U}_k^l$, which we denote by $\mathbf{w}_{U_k^l}^l$. The final solution $\mathbf{w}_k^l$ is then

$$\mathbf{w}_k^l = \mathbf{U}_k^l \mathbf{w}_{U_k^l}^l. \quad (4.63)$$

Note that distortion can also be introduced due to extra processing steps. The distortion only affects the audio quality but not the source separation.

The solution (4.63) shows that $\mathbf{w}_k^l$ lives in the null space of $\mathbf{\Gamma}_k^{\setminus l \text{ H}}$. The null space $\mathbf{U}^l$ will only exist when $MPK \geq (L-1)(3P-2)K$. Otherwise,

we can choose $\mathbf{U}_k^l$ to be the eigenvector of $\mathbf{\Gamma}^{\setminus l} \mathbf{\Gamma}^{\setminus l \text{H}}$ that associated with the smallest eigenvalue.

By using the proposed LFRTF-domain method, we can trade the number of the microphone M with the system latency to retain the separation performance. The system latency can be reduced by keeping the length of the window N_b as short as possible, but this increases P as $P = \frac{N_w}{N_a}$ and $\frac{N_b}{N_a} = 2^t$, where $t \in \mathbb{Z}$. Hence, the number of microphones has to be increased as follows:

$$\frac{N_b}{N_w} = \frac{N_b}{N_a} \frac{1}{P}, \quad (4.64)$$

$$\frac{1}{P} > \frac{3}{2} - \frac{M}{2(L-1)}, \quad (4.65)$$

where $\frac{N_b}{N_w}$ represents the factor of delay. Ideally, the number of LFRTF window shifts per HFRTF window is $P = 2^r$, where $r \in \mathbb{Z}$. For instance, to separate three original source signals, the delay can be reduced to any desired rate by using only six microphones. Note that the number of neighbour frequency bins does not affect the number of microphones required in (4.65).

4.4 Results

In this section, we first discuss the experimental setup and procedure. Then we provide the experimental results. Lastly, we provide a brief discussion on the computational complexity of the proposed approaches.

4.4.1 Experimental Setup and Procedure

As data we used the 10 seconds speech source signals of the Stereo Audio Source Separation Evaluation Campaign (SASSEC) [190]. The sampling rate was 8 kHz. In the experiment, eight microphones were used. The positions of the sources were fixed while the microphones were randomly placed in a rectangular room with a size of 4 m $\times$ 3 m $\times$ 3 m, at a fixed height of 1.5 m. For each experiment, three random speech signals were selected as sources.

We used the room impulse response (RIR) simulator [54], based on the image source method [5], to generate the 8×3 RIRs with 4096 taps. Each set of experiment was conducted three times with different reverberation times, which were 0.1 s, 0.2 s and 0.3 s. The simulated RIRs were then convolved with the speech signals to obtain the observations.

In all experiments, a square root of Hann window was used. In the high-frequency-resolution time-frequency (HFRTF) domain, the window length was 512 taps (64 ms) and the windows were 75% overlapped, i.e., $N_w = 512$, $N_l = 128$ in (4.1). The HFRTF-domain demixing matrix in every frequency bin was estimated by the complex fast independent component analysis (ICA) approach [20] using the first 120 time-frequency (TF) frames of the observations. In each frequency bin, the permutation ambiguity was resolved using oracle information, which is based on the correlation coefficients between the amplitude envelope (absolute value) of the estimated sources and the original sources across the 120 TF frames. After correcting the permutation orders, we minimized the scaling effect using the minimal distortion principle method [110]. Then, the HFRTF-domain mixing matrix in every frequency bin was computed by taking the pseudo-inverse of the HFRTF-domain demixing matrix.

We compared the performance between various methods. These approaches include our two proposed methods, the impulse response truncation method [172] (Sunohara) and the original independent component analysis (ICA) [69] that was operated in both the HFRTF domain and the low-frequency-resolution time-frequency (LFRTF) domain. Note that the scaling and permutation ambiguities were not present in our proposed methods and Sunohara's method. These issues were resolved in the HFRTF domain. The separation performance was evaluated using the signal-to-interference ratio (SIR) and the signal-to-distortion ratio (SDR) between the original source and the separated signal, which are commonly used in BSS applications [?, 107]. The SIR value indicates the ratio between the signal energy of the desired source and the signal energy of the interference signals. The SDR

value indicates the ratio between the signal energy of the desired source and the energy of the error signal, which is due to interfering sources and artifacts. The SIR and SDR values were computed using the BSS_EVAL toolbox [189] and higher values indicate better separation performance. Each set of experiment was repeated 30 trials and the averaged values were presented in the results.

For our time-domain method (Proposed-TD), the block length is 128 taps (16 ms), e.g., $N_f = 128$. The demixing parameters in the time domain was computed based on (4.41) by using the mixing matrix that was previously estimated in the HFRTF domain. The system algorithmic latency caused by the Proposed-TD method was 16 ms.

For our LFRTF-domain method (Proposed-LFR), a 128-taps (16 ms) square-root of Hann window with 50% overlapped was used, e.g., $N_b = 128$ and $N_a = 64$. The LFRTF-domain demixing parameters were designed based on (4.62) using the estimated mixing matrix in the HFRTF domain by considering three neighbour frequency bins, i.e., $\Delta k = 3$. The system algorithmic latency caused by the Proposed-LFR method was also 16 ms.

For the original ICA approach, the estimated demixing matrix was directly applied to the observations without post-processing. The ICA method that operated in the HFRTF domain (ICA-HFR) used a long window ($N_w = 512$) while the ICA method that operated in the LFRTF domain (ICA-LFR) used a short window ($N_w = 128$). Thus, the system algorithmic latency caused by the ICA-HFR method and the ICA-LFR method were 64 ms and 16 ms, respectively.

For Sunohara's method [172], the time-domain demixing filters were first obtained by performing the inverse Fourier transform on the HFRTF demixing matrix. Next, the first 128 samples of the demixing filters were truncated. Then, the source separation was performed using the truncated demixing filter in the time domain. Thus, the system algorithmic latency of this method was 16 ms.

Figure 4.4: The separation performance (SIR) of various techniques with different reverberation times, where the number of microphones $M = 8$ and the algorithmic latency was 16 ms.

4.4.2 Experimental Results

Both Fig. 4.4 and Fig. 4.5 show comparisons of the blind source separation performance between different methods with different room reverberations. The blue, red and yellow circles in Fig. 4.4 indicate the mean SIR values of the first, second and third sources, respectively. In Fig. 4.5, the blue, red and yellow circles indicate the mean SDR values of the first, second and third sources, respectively. The error bars represent the standard deviation based on the 30 trials.

Figure 4.5: The separation performance (SDR) of various techniques with different reverberation times, where the number of microphones $M = 8$ and the algorithmic latency was 16 ms.

The results show that the performance of both our proposed method surpasses the other low-latency approaches, e.g., the impulse response truncating method (Sunohara) and the method that directly implements the ICA approach in the LFRF domain (ICA-LFR). It is interesting that the performance of the LFRF-domain method (Proposed-LFR) is similar with the proposed time-domain method (Proposed-TD), although only three neighbour frequency bins ($\Delta k = 3$) were considered in the experiment. This shows that the effect of the further neighbour frequency bins is not significant.

Fig. ?? illustrates the performance of the proposed methods against the

Figure 4.6: Performance of the proposed methods (SIR) against different number of microphones M at different algorithmic latency, where the reverberation time is $\beta = 0.2$ s.

number of microphones for a reverberation time of 0.2 s. The performance was evaluated using the average SIR values of the three sources. The experimental results agree with the theoretical predictions (4.43) and (4.65). That is, the performance of the source separation can be retained in the low-latency domain provided that there is a sufficient number of microphones. When the number of microphones increases, the separation performance continues to increase slowly as the solution in the HFRTF domain becomes more accurate.

Fig. 4.6 and Fig. 4.7 depict the performance of the proposed methods against the number of microphones with different algorithmic latency for a reverberation time of 0.2 s. The performance was evaluated using the average SIR values and the average SDR values of the three sources in Fig. 4.6 and Fig. 4.7, respectively. When the number of microphones increases, the separation performance continues to increase slowly as the solution in the HFRTF domain becomes more accurate. We notice that the Proposed-LFR method performs better than proposed-TD especially in terms of the SDR metric, when the condition (4.43) is fulfilled (e.g., $M \geq 6$ for three sources). Because the Proposed-TD method can be considered as a full-band approach

Figure 4.7: Performance of the proposed methods (SDR) against different number of microphones M at different algorithmic latency, where the reverberation time is $\beta = 0.2$ s.

Figure 4.8: Performance of the proposed methods (SIR in (a) and SDR in (b)) against algorithmic latency using six microphones, where the reverberation time is $\beta = 0.2$ s.

that utilizes all the frequency bins when computing the demixing parameters, it requires more processing effort and this introduces more distortion.

Fig. 4.8 and Fig. 4.9 illustrate the performance of the proposed methods against the algorithmic latency using six microphones for a reverberation time of 0.2 s and 0.5 s, respectively. For the latter scenario, the room size

(a) Average SIR against algorithmic latency. (b) Average SDR against algorithmic latency.

Figure 4.9: Performance of the proposed methods (SIR in (a) and SDR in (b)) against algorithmic latency using six microphones, where the room size is $9 \text{ m} \times 6 \text{ m} \times 2.7 \text{ m}$ and the reverberation time is $\beta = 0.5 \text{ s}$.

was $9 \text{ m} \times 6 \text{ m} \times 2.7 \text{ m}$. This is to simulate a classroom setting. Since the reverberation time was higher than previous experiments, satisfactory separation performance cannot be obtained by using a window with only 512 taps (64 ms) in the HFRTF domain. Thus, a window with 1024 taps (128 ms) was used instead. The performance was evaluated using the average SIR values and the average SDR values of the three sources. The experimental results agree with the theoretical predictions (4.43) and (4.65). That is, the performance of the source separation (SIR values) can be retained in the low-latency domain at any desired delay provided that there is a sufficient number of microphones. However, the SDR values are affected when the algorithmic latency is reduced due to processing artifacts. We also repeated the experiments by using more microphones, e.g., ($M = 7$ and $M = 8$) and obtained similar results. Hence, the performance of the proposed methods is consistent over different room settings.

We briefly discuss the computational complexity of the proposed approaches. Recall that M is the number of microphones, L is the number of sources, N_w is the window length in the HFRTF domain, N_f is the data

acquisition-block length, N_b is the window length in the LFRTF domain, N_a is the length of the shift, $K = 2\Delta k + 1$, where Δk is the number of neighbour frequency bins, and $P = \frac{N_w}{N_a}$.

Compared to the ICA-HFR method, the Proposed-TD method requires to perform ML times of matrix multiplication to obtain (4.33), L evaluations of singular value decomposition (SVD) to obtain the null spaces $\mathbf{U}^l$ and another L evaluations for the SVD to obtain the demixing vector $\mathbf{w}^l$. Among these operations, the computational cost for the SVD for finding the null spaces $\mathbf{U}^l$ is the largest, i.e., $\mathcal{O}(LMN_w(L-1)^2(3N_w-2N_f)^2)$. Thus, the additional computational cost of the Proposed-TD method is of the order of $\mathcal{O}(ML^3N_w^3)$.

Compared to the ICA-HFR method, the Proposed-LFD approach requires to perform ML matrix multiplications to obtain (4.56), $L(\frac{N_b}{2}+1)$ evaluations for the SVD to find the null spaces $\mathbf{U}_k^l$ in every frequency bin and another $L(\frac{N_b}{2}+1)$ evaluations for the SVD to obtain the demixing vector $\mathbf{w}_k^l$. The computational cost for the SVD for finding the null spaces $\mathbf{U}_k^l$ is the largest, i.e., $\mathcal{O}(L(\frac{N_b}{2}+1)MPK(L-1)^2(3P-2)^2K^2)$, among these operations. Thus, the additional computational cost of the Proposed-LFD method is of the order of $\mathcal{O}(MN_bL^3P^3K^3)$.

4.5 Conclusion

In this chapter, we proposed two related methods to reduce the algorithmic delay of blind source separation (BSS). Both methods utilize the high-frequency-resolution time-frequency domain (HFRTF) to perform the estimation. Then, the separation parameters in the low-latency domain are computed based on the estimation in the HFRTF. The methods can be integrated with existing BSS algorithms, for instance, the conventional independent component analysis method without affecting separation performance. They help to avoid using a long window for separating process, and hence, this will result in a reduction of the system latency. This means the imple-

mentation of existing BSS algorithms in real-time systems becomes possible.

Our methods exploit that adding microphones leads to additional specification of the relation between source signals and microphone observations. In theory, the system latency can be reduced to any desired value by adjusting the number of microphones.

Chapter 5

Blind Source Separation Method for Non-Stationary Scenarios Using Meta-Information

In this chapter, we present a method to deal with the blind source separation (BSS) problem when a new source is introduced. The new source can be either static or moving. A conventional generalized sidelobe canceller (GSC) is particularly useful in this scenario, but it requires a sufficient number of samples to obtain an accurate covariance matrix. This increases the estimation delay. Our proposed method helps to resolve this difficulty. It enables the estimation of an accurate covariance matrix using only a single snapshot of observation data by utilizing the entire spectrum. Hence, the estimation delay reduces, which makes it more applicable to real-world applications. Our experiment results confirm the effectiveness of the proposed approach.

5.1 Introduction

Most existing blind source separation (BSS) approaches were proposed for time-invariant scenarios, e.g., static sources. These methods include but are not limited to the independent component analysis (ICA) approach [20], the degenerate unmixing estimation technique (DUET) algorithm [198], the time-frequency ratio of mixtures (TIFROM) algorithm, etc. In a non-stationary scenario, e.g., when a new static or moving source is introduced to the system, the separation performance may degrade. To cope with non-stationary scenarios, a real-time batch-wise implementation of conventional ICA approaches was used in [120, 121]. The observation signals are disassembled into batches by assuming that each batch contains time-invariant or slightly time-varying mixtures, so that conventional ICA approaches can be adopted in each batch to perform source separation.

The drawback of the approaches [120, 121] is that the users will experience a delay for the system to update the separation parameters when the environment is changing. Since ICA approaches are statistical, the delay can be long as a sufficient number of samples is required to obtain good separation performance. If the environment keeps changing, then the overall separation performance can be poor. This is because each batch can no longer be approximated as time-invariant unless the length of each batch is short and this affects the separation performance due to an insufficient number of samples.

Recently, a few approaches [126, 128, 175] have been proposed to perform source separation on moving sources. These methods involve tracking of moving sources, which can be done either by using a video camera [126] or by estimating directions of arrival (DOA) of the moving sources. The source separation is then performed by using beamforming techniques with the reconstructed model based on the estimated source trajectories. All these approaches perform analysis from scratch without utilizing “meta-information” that can be preserved from previous estimations. For instance, when a new source appears while the desired sources remain stationary. In this circum-

stance, the mixing channels of the desired sources are still the same, hence the information about the mixing channels of the desired sources from previous analysis can be used to facilitate the later estimation. Moreover, the existing approaches require a sufficient of samples to perform analysis and this increases the estimation delay.

In this chapter, we propose an approach combined with the generalized sidelobe canceller (GSC) beamformer to perform source separation in a scenario, when a new source is introduced to the system. The new source can be either static or moving. Our proposed method is able to perform BSS within one single snapshot of observation data in a free-field scenario or in a less reverberant environment. This reduces the estimation delay as one block of data is only needed for analysis. The essence of the proposed method is that we utilize the entire spectrum to create multiple measurements, so that a new covariance matrix can be reconstructed. By using an innovative structure that is based on a GSC beamformer, the “meta-information” from previous operations can be reused. Hence, the efficiency of the BSS system can be increased.

This chapter starts with a problem formulation to describe a model of the BSS when a new source is introduced. Second, we review the conventional method to cope with the new source and discuss the limitation. Next, we provide our proposed method to improve the conventional method. Then, we compare our proposed approach with the conventional method and present the experimental results.

5.2 Problem Formulation

This section provides the problem formulation when a new incoming signal is introduced to a blind source separation (BSS) system. The new incoming signal acts as an interfering signal and the interfering source can be either static or moving.

Consider M microphones, L desired sources and one interfering source,

where $M > L$. Denote by p and k the block index and the frequency bin index, respectively, in the time-frequency (TF) domain. The observation signals in the short-time Fourier transform (STFT) domain in a noiseless scenario can be expressed as:

$$\mathbf{x}_{p,k} \approx \mathbf{A}_k \mathbf{s}_{p,k} + \mathbf{v}_{p,k} z_{p,k}, \quad (5.1)$$

where $\mathbf{x}_{p,k} \in \mathbb{C}^{M \times 1}$, $\mathbf{s}_{p,k} \in \mathbb{C}^{L \times 1}$ and $z_{p,k} \in \mathbb{C}^{1 \times 1}$ indicate the observation signals, the desired sources and the interfering source, respectively, at TF-index (p, k) while $\mathbf{A}_k \in \mathbb{C}^{M \times L}$ represents the mixing matrix at frequency bin k and $\mathbf{v}_{p,k} \in \mathbb{C}^{M \times 1}$ represents the mixing vector of the interfering source at TF-index (p, k) .

We assume that we are provided with a demixing matrix $\mathbf{W}_{p,k} \in \mathbb{C}^{L \times M}$ that has previously been estimated during the periods when the interfering source $z_{p,k}$ is absent. We also assume that there are no permutation and scaling ambiguities. Then

$$\mathbf{W}_k^H \mathbf{A}_k = \mathbf{I}, \quad (5.2)$$

where $\cdot^H$ denotes the Hermitian transpose operator. By applying $\mathbf{W}_{p,k}$ on the mixtures in (5.1), the desired sources can still be separated as:

$$\mathbf{y}_{p,k} = \mathbf{W}_k^H \mathbf{x}_{p,k} = \mathbf{s}_{p,k} + \mathbf{W}_k^H \mathbf{v}_{p,k} z_{p,k}. \quad (5.3)$$

However, the l^{th} separated source contains the l^{th} desired source and also the interfering source with a scaling factor of $\mathbf{w}_k^l \mathbf{v}_{p,k}$ as expressed in (5.3), where $\mathbf{w}_k^l \in \mathbb{C}^{1 \times M}$ is a column vector, located at the l^{th} column in $\mathbf{W}_k$.

To separate the desired source from the mixture signal $y_{p,k}$ in (5.3), the obvious solution is to perform again the BSS algorithm directly on the new mixtures $x_{p,k}$ in (5.1). However, there are two major drawbacks to this approach. First, it suffers from the estimation delay. The BSS algorithms require a sufficient number of samples to perform the separation accurately. Second, the efficiency is low as the operation is redundant. Most of the low-complexity BSS algorithms operate with static sources only.

In (5.3) we show that by using the previous estimated demixing matrix $\mathbf{W}_{p,k}$ as the meta-information, we are able to separate each individual desired source although the results contain components of the interfering signal. The only task that remains is to minimize the contribution of the interfering signal in each separated source. Methods for coping with a time-invariant channel due to a static source and a time-varying channel due to a moving source are discussed in Sec. 5.3 and in Sec. 5.4, respectively. In the derivations, we assume that the mean of the desired sources and the interfering signal are all zero. Thus, the mean can be discarded while computing the correlation coefficient and the covariance matrix.

5.3 Time-Invariant Channel

In this section, we provide a least-squares approach to estimating each desired source from the separated sources (5.3). Consider new observation signals that contain previous desired sources and a static interfering source, i.e., the interfering source is not moving. By applying “meta-information” that is available from previous operations, e.g., the previous demixing matrix $\mathbf{W}_k$, to perform BSS on the new observation signals, we are able to obtain one desired source, albeit contaminated with an interfering signal, in each separated signal. Hence, the aim of the approach is to eliminate the components of the interfering signal from each separated signal and update the demixing matrix. If multi-dimensional interfering signals, e.g., multiple estimates of the interfering signals, are considered, we show that the solution of the problem leads to a solution based on a generalized sidelobe canceller (GSC) beamformer [53].

If the interfering source is not moving, then the unknown mixing vector of the interfering source $\mathbf{v}_{p,k}$ is a constant across the TF blocks, i.e.,

$$\mathbf{v}_{p,k} = \mathbf{v}_k \quad \forall p. \quad (5.4)$$

Hence, the l^{th} separated source in (5.3) can be simplified to

$$y_{p,k}^l = s_{p,k}^l + \mathbf{w}_k^l \mathbf{H} \mathbf{v}_k z_{p,k} = s_{p,k}^l + c_k^l z_{p,k}, \quad (5.5)$$

where $c_k^l = \mathbf{w}_k^l \mathbf{H} \mathbf{v}_k$ denotes the scaling factor of the interfering signal in the l^{th} separated source. Now, we use (5.5) and define an error signal for every separated source in every frequency bin as:

$$e_{p,k}^l = y_{p,k}^l - d_k^l \hat{z}_{p,k} = s_{p,k}^l + c_k^l z_{p,k} - d_k^l \hat{z}_{p,k}, \quad (5.6)$$

where d_k^l is a weighting factor and $\hat{z}_{p,k}$ represents the estimated interfering signal, which can be expressed as:

$$\hat{z}_{p,k} = \hat{c}_k z_{p,k}, \quad (5.7)$$

where $\hat{c}_k$ denotes a scaling factor. The interfering signal can be easily estimated if we have a sufficient number of microphones, e.g., $M > L$. We can design a vector that is orthogonal to the mixing vectors of the desired sources by using “meta-information”. The vector is then used to estimate the interfering signal by suppressing the contributions of the desired sources in the new observation signals. However, the scaling of the estimated interfering signal is incorrect due to the unknown mixing vector of the interfering source $\mathbf{v}_k$, which will be discussed later in (5.14).

The goal of the approach is to minimize the error signal by finding an optimal d_k^l , so that the components of the interfering signal $z_{p,k}$ in the separated source $y_{p,k}^l$ can be suppressed, so a clean desired source $s_{p,k}^l$ can be obtained. By minimizing the mean square error of (5.6), the optimal d_k^l can be obtained as:

$$d_k^l = \frac{r_{y^l \hat{z}, k}}{r_{\hat{z} \hat{z}, k}}. \quad (5.8)$$

The correlation coefficient $r_{y_k^l \hat{z}_k}$ between $y_{p,k}^l$ and $\hat{z}_{p,k}$ is defined as:

$$r_{y^l \hat{z}, k} = \frac{1}{N_p} \sum_{p=0}^{N_p-1} y_{p,k}^l \hat{z}_{p,k}^*, \quad (5.9)$$

where $\cdot^*$ is the conjugate operator.

In the following, we prove that the weight d_k^l obtained in (5.8) is an optimum weight factor to eliminate the components of the interfering signal $z_{p,k}$ in the separated source $y_{p,k}^l$ in (5.5). Substituting (5.5) and (5.7) into (5.9) and (5.8) yields:

$$\begin{aligned} d_k^l &= \frac{\sum_{p=0}^{N_p-1} (s_{p,k}^l + c_k^l z_{p,k}) z_{p,k}^* \hat{c}_k^*}{\sum_{p=0}^{N_p-1} \hat{c}_k z_{p,k} z_{p,k}^* \hat{c}_k^*} \\ &= \frac{\sum_{p=0}^{N_p-1} s_{p,k}^l z_{p,k}^* \hat{c}_k^* + c_k^l \hat{c}_k^* z_{p,k} z_{p,k}^*}{\sum_{p=0}^{N_p-1} \hat{c}_k \hat{c}_k^* z_{p,k} z_{p,k}^*} \end{aligned} \quad (5.10)$$

Let us assume that the l^{th} desired source $s_{p,k}^l$ and the interference signal $z_{p,k}$ are uncorrelated. Then we have

$$\sum_{p=0}^{N_p-1} s_{p,k}^l z_{p,k}^* \approx 0. \quad (5.11)$$

Thus, (5.10) can be approximated as:

$$d_k^l = \frac{c_k^l \hat{c}_k^* \sum_{p=0}^{N_p-1} z_{p,k} z_{p,k}^*}{\hat{c}_k \hat{c}_k^* \sum_{p=0}^{N_p-1} z_{p,k} z_{p,k}^*} = \frac{c_k^l}{\hat{c}_k}. \quad (5.12)$$

Substitute (5.12) into (5.6) and we obtain the l^{th} desired source as:

$$e_{p,k}^l = s_{p,k}^l + c_k^l z_{p,k} - \frac{c_k^l}{\hat{c}_k} \hat{c}_k z_{p,k} = s_{p,k}^l. \quad (5.13)$$

We have seen so far the method to extract the desired source $y_{p,k}^l$ by using the previous estimated demixing vector of the desired source $\mathbf{w}_{p,k}^l$ and the estimated interfering signal $\hat{z}_{p,k}$. Now, we provide a method to estimate the interfering signal. The estimated interfering signal is a scaled version of the original interfering signal.

To estimate the interfering signal, we first want to find a vector $\mathbf{b}_k$ such that $\mathbf{b}_k^H \mathbf{A}_k = \mathbf{0}$. Then, the interference signal with a scaling factor of $\hat{c}_k$ can be obtained by

$$\hat{z}_{p,k} = \mathbf{b}_k^H \mathbf{x}_{p,k} = \mathbf{b}_k^H \mathbf{A}_k \mathbf{s}_{p,k} + \mathbf{b}_k^H v_k z_{p,k} = \mathbf{0} + \mathbf{b}_k^H v_k z_{p,k} = \hat{c}_k z_{p,k}. \quad (5.14)$$

Thus, $\mathbf{b}_k$ lies in the null space of $\mathbf{A}_k^H \in \mathbb{C}^{L \times M}$ and the null space of $\mathbf{A}_k^H$ always exists as $M > L$.

By using the previously estimated demixing vector of the desired source $\mathbf{w}_k^l$ and (5.14), we can find the new demixing vector $\mathbf{w}_{k,\text{new}}^l$ of the l^{th} desired source through (5.13). The new demixing vector $\mathbf{w}_k^l$ can be computed as:

$$s_{p,k}^l = \mathbf{w}_k^{lH} \mathbf{x}_{p,k} - d_k^l \mathbf{b}_k^H \mathbf{x}_{p,k} = \left(\mathbf{w}_k^l - \mathbf{b}_k d_k^{l*} \right)^H \mathbf{x}_{p,k}, \quad (5.15)$$

and hence,

$$\mathbf{w}_{k,\text{new}}^l = \mathbf{w}_k^l - \mathbf{b}_k d_k^{l*}. \quad (5.16)$$

The correlation coefficients $r_{y^l \hat{z},k}$ and $r_{\hat{z} \hat{z},k}$ can be written as:

$$r_{y^l \hat{z},k} = \frac{1}{N_p} \sum_{p=0}^{N-1} \mathbf{w}_k^{lH} \mathbf{x}_{p,k} \mathbf{x}_{p,k}^H \mathbf{b}_k = \mathbf{w}_k^{lH} \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{b}_k, \quad (5.17)$$

and

$$r_{\hat{z} \hat{z},k} = \frac{1}{N_p} \sum_{p=0}^{N-1} \mathbf{b}_k^H \mathbf{x}_{p,k} \mathbf{x}_{p,k}^H \mathbf{b}_k = \mathbf{b}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{b}_k, \quad (5.18)$$

respectively, where $\mathbf{R}_{\mathbf{x}\mathbf{x},k}$ denotes the covariance matrix of $\mathbf{x}_{p,k}$, e.g.,

$$\mathbf{R}_{\mathbf{x}\mathbf{x},k} = \sum_{p=0}^{N-1} \mathbf{x}_{p,k} \mathbf{x}_{p,k}^H. \quad (5.19)$$

Thus, the optimum weighting factor d_k^l can also be expressed as:

$$d_k^{l*} = \left(\mathbf{b}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{b}_k \right)^{-1} \mathbf{b}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{w}_k^l. \quad (5.20)$$

It is worth noting that the solution (5.16) is a simplified version of the generalized sidelobe canceller (GSC) beamformer [53], where $\mathbf{w}_k^l$ acts as a steering vector, $\mathbf{b}_k$ is a blocking vector instead of a blocking matrix and d_k^{l*} is the interference cancelling filter [93].

Fig. 5.1 illustrates the GSC-like structure of our problem when multiple observation signals are considered. The estimated original sources can be computed as:

$$\hat{\mathbf{s}}_{p,k}^l = \mathbf{w}_{k,\text{new}}^{lH} \mathbf{x}_{p,k}. \quad (5.21)$$

The new demixing vector of the l^{th} desired source can be updated as:

$$\mathbf{w}_{k,\text{new}}^l = \mathbf{w}_k^l - \mathbf{B}_k \mathbf{d}_k^l, \quad (5.22)$$

where $\mathbf{B}_k$ denotes as the blocking matrix, i.e., $\mathbf{B}_k \mathbf{A}_k = \mathbf{0}$ and $\mathbf{d}_k^l$ now is a vector of interference cancelling filters, which can be computed as:

$$\mathbf{d}_k^l = (\mathbf{B}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{B}_k)^{-1} \mathbf{B}_k^H \mathbf{R}_{\mathbf{x}\mathbf{x},k} \mathbf{w}_k^l. \quad (5.23)$$

Figure 5.1: The GSC-like structure of the BSS problem for extracting the l^{th} desired source.

5.4 Time-Varying Channel

In this section, we first show that the conventional GSC beamformer can also be useful when dealing with a time-varying channel due to a new moving interfering source by using a model with two interfering sources to simulate the interfering signal. Next, we discuss the problem to be solved. Then, we propose an approach to enhance the performance of the conventional GSC method. The proposed method enables the conventional GSC method to find solutions by using one single snapshot of observation data. This is particularly useful while coping with a moving interfering source as the mixing channel varies continuously. We also discuss methods that can be used to improve the robustness of the proposed method.

5.4.1 Model of a Moving Source

A moving source can be seen as multiple sources at different locations across time. We provide a model for the simplest case by considering only two locations. Let us denote by k the frequency bin index. In the STFT domain, the mixture of the desired source $\mathbf{s}_{\cdot,k}$ and the interfering signal $z_{\cdot,k}$ at the first location as well as the second location can be expressed as

$$\mathbf{x}_{p_1,k} = \mathbf{A}_k \mathbf{s}_{p_1,k} + \mathbf{v}_{p_1,k} z_{p,k}, \quad (5.24)$$

and

$$\mathbf{x}_{p_2,k} = \mathbf{A}_k \mathbf{s}_{p_2,k} + \mathbf{v}_{p_2,k} z_{p_2,k}, \quad (5.25)$$

respectively, where $\mathbf{A}_k$ is the mixing matrix of the desired sources at frequency bin index k , p_1 denotes the block index when the interfering source is at the first location and p_2 denotes the block index when the interfering source is at the second location. Alternatively, (5.24) and (5.25) can be rewritten as:

$$\mathbf{x}_{p,k} = \mathbf{A}_k \mathbf{s}_{p,k} + \mathbf{v}_{p_1,k} \tilde{z}_{1,p,k} + \mathbf{v}_{p_2,k} \tilde{z}_{2,p,k}, \quad (5.26)$$

where $p \in [p_1, p_2]$ and

$$\tilde{z}_{1,p,k} = \begin{cases} z_{p,k} & p = p_1, \\ 0 & p = p_2, \end{cases} \quad (5.27)$$

$$\tilde{z}_{2,p,k} = \begin{cases} 0 & p = p_1, \\ z_{p,k} & p = p_2. \end{cases} \quad (5.28)$$

We now show that by using a sufficient number of microphones M , the conventional GSC method can still be employed to update the demixing vector $\mathbf{w}_k^l$. Recall that L represents the number of the desired sources. By using (5.26), the output $y_{p,k}^l$ after applying the demixing vector $\mathbf{w}_k^l$, which is displayed in Fig. 5.1, can be expressed as:

$$y_{p,k}^l = s_{p,k}^l + \mathbf{w}_k^{lH} \mathbf{v}_{p_1,k} \tilde{z}_{1,p,k} + \mathbf{w}_k^{lH} \mathbf{v}_{p_2,k} \tilde{z}_{2,p,k}, \quad (5.29)$$

and the output $\tilde{\mathbf{z}}_{p,k}$ after applying the blocking matrix $\mathbf{B}_k \in \mathbb{C}^{M \times (M-L)}$ can be expressed as:

$$\tilde{\mathbf{z}}_{p,k}^l = \mathbf{B}_k^H \mathbf{v}_{p1,k} \tilde{z}_{1,p,k} + \mathbf{B}_k^H \mathbf{v}_{p2,k} \tilde{z}_{2,p,k}. \quad (5.30)$$

The optimal interference cancelling filters $\mathbf{d}_k^l \in \mathbb{C}^{(M-L) \times 1}$ can be found by minimizing

$$\sum_p \left| y_{p,k}^l - \mathbf{d}_k^{lH} \tilde{\mathbf{z}}_{p,k} \right|^2. \quad (5.31)$$

After finding the optimal $\mathbf{d}_k^l$, the cost function in (5.31) can be expressed as:

$$y_{p,k}^l - \mathbf{d}_k^{lH} \tilde{\mathbf{z}}_{p,k} = \hat{s}_{p,k}^l. \quad (5.32)$$

Substituting (5.29) and (5.30) into (5.32) yields:

$$s_{p,k}^l + \mathbf{w}_k^{lH} (\mathbf{v}_{p1,k} \tilde{z}_{1,p,k} + \mathbf{v}_{p2,k} \tilde{z}_{2,p,k}) - \mathbf{d}_k^{lH} \mathbf{B}_k^H (\mathbf{v}_{p1,k} \tilde{z}_{1,p,k} + \mathbf{v}_{p2,k} \tilde{z}_{2,p,k}) = \hat{s}_{p,k}^l. \quad (5.33)$$

Rearrange (5.33) and we obtain

$$(\mathbf{w}_k^l - \mathbf{d}_k^l \mathbf{B}_k)^H \mathbf{v}_{p1,k} \tilde{z}_{1,p,k} + (\mathbf{w}_k^l - \mathbf{d}_k^l \mathbf{B}_k)^H \mathbf{v}_{p2,k} \tilde{z}_{2,p,k} = \hat{s}_{p,k}^l - s_{p,k}^l. \quad (5.34)$$

Since $\mathbf{d}_k^l$ is the optimum solution, the right-hand-side of (5.34) vanishes. Then, we can rewrite (5.34) into:

$$\begin{bmatrix} \mathbf{v}_{p1,k}^H \mathbf{B}_k \\ \mathbf{v}_{p2,k}^H \mathbf{B}_k \end{bmatrix} \mathbf{d}_k^l = \begin{bmatrix} \mathbf{v}_{p1,k}^H \mathbf{w}_k^l \\ \mathbf{v}_{p2,k}^H \mathbf{w}_k^l \end{bmatrix}, \quad (5.35)$$

which defines a system that consists of two linear equations.

To solve (5.35), the dimensionality of the interference cancelling filter $\mathbf{d}_k^l$ must exceed two, e.g., $M-L \geq 2$. This implies that if we perform the analysis on N observation data, we need at least $N+L$ microphones to employ GSC beamformer for suppressing components of the moving interfering source. By using more microphones, the source separation performance can be improved and the algorithmic delay can also be reduced by using the proposed approach that is discussed in Chapter 4. However, to ensure the accuracy of the sample

covariance matrix that is needed in the GSC beamformer, a large number of data samples is needed. Hence, the number of microphones required can be large, which is impractical. Thus, we propose an approach to reconstruct the sample covariance matrix by using only one observation data sample. The details are discussed in the next subsection.

5.4.2 Proposed Method

Our proposed method is based on the GSC structure, but it is able to update the demixing matrix by using only one snapshot of observation. If we would use the conventional GSC approach as discussed in (5.23), we obtain a rank-1 covariance matrix while considering only one snapshot of observation, which is problematic for the later estimation step. This problem can be avoided by using the proposed approach. The essence of the proposed method is that, based on the free-field assumption, we are able to utilize other frequency bins to reconstruct the covariance matrix in the analysed frequency bin. Then, the rank of the reconstructed covariance matrix is no longer one as it now contains the information of all the desired sources and the interfering signal. The new demixing matrix can be obtained by using the GSC method with a reconstructed covariance matrix.

Consider a two-dimensional plane. In a free-field scenario, the observation signal of one signal coming from the direction θ_p^l in every TF-index (p, k) can be expressed as:

$$\mathbf{x}_{p,k}^l = \mathbf{a}(\theta_p^l)_k \mathbf{s}_{p,k}^l, \quad (5.36)$$

where $\mathbf{a}(\theta_p^l)_k$ denotes the steering vector of the direction θ_p^l . There is a unique relationship of the steering vectors between frequency bins. For instance, for a uniform circular microphone array, the m^{th} entry of $\mathbf{a}(\theta_p^l)_k$ can be computed as:

$$a(\theta_p^l)_k^m = \exp^{-j \frac{2\pi}{K} k \frac{f_s}{c} \cos(\theta_p^l - \frac{2\pi m}{M})} = \left(\exp^{-j \frac{2\pi}{K} \frac{f_s}{c} \cos(\theta_p^l - \frac{2\pi m}{M})} \right)^k, \quad (5.37)$$

where K denotes the number of frequency bins, f_s is the sampling rate, c is the speed of the sound and M is the number of microphones.

If $\mathbf{a}(\theta_p^l)_{k'}$ and $\mathbf{a}(\theta_p^l)_{\tilde{k}}$ are given, where $k' + \tilde{k} = k$, then we are able to compute the steering vector $\mathbf{a}(\theta_p^l)_k$, i.e.,

$$\mathbf{a}(\theta_p^l)_k = \mathbf{a}(\theta_p^l)_{k'} \circ \mathbf{a}(\theta_p^l)_{\tilde{k}}, \quad (5.38)$$

where $\circ$ denotes a Hadamard product. Note that (5.38) is valid for any arrangement of the microphone array in a free-field scenario. Thus, we are able to generate a new observation signal in frequency bin k by using the observation signal in other frequency bins, e.g., frequency bins k' and $\tilde{k}$, where $k' + \tilde{k} = k$. The new observation signal can be computed as:

$$\mathbf{x}_{p,k'+\tilde{k}}^l = \mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l = \mathbf{a}(\theta_p^l)_{k'} \circ \mathbf{a}(\theta_p^l)_{\tilde{k}} \circ \mathbf{s}_{p,k'}^l \circ \mathbf{s}_{p,\tilde{k}}^l. \quad (5.39)$$

The vector of the new observation signal $\mathbf{x}_{p,k'+\tilde{k}}^l$ points to the same direction of the vector of the original observation signal in frequency bin k $\mathbf{x}_{p,k}^l$ but with a different complex scaling factor, i.e., $\mathbf{s}_{p,k'}^l \circ \mathbf{s}_{p,\tilde{k}}^l$.

Due to the circular and symmetry properties of the complex number, there are always $\lfloor \frac{K}{2} \rfloor$ distinct pairs of k' and $\tilde{k}$ that fulfill the condition $k' + \tilde{k} = k$. Hence, we can always add at least $\lfloor \frac{K}{2} \rfloor$ new observation signals while performing analysis in each frequency bin.

The reconstructed covariance matrix $\mathbf{R}_{\mathbf{x}^l \mathbf{x}^l, p, k}$ at TF-index (p, k) of the observation signal that contains only one single source can be computed as:

$$\begin{aligned} \mathbf{R}_{\mathbf{x}^l \mathbf{x}^l, p, k} &= \frac{1}{\lfloor \frac{K}{2} \rfloor + 1} \sum_{k', \tilde{k} \in \mathcal{K}} \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right) \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right)^H, \\ \mathcal{K} &= \left\{ k', \tilde{k} \mid k' + \tilde{k} = k \cap k', \tilde{k} \in \left[-\frac{K}{2}, \frac{K}{2} \right] \right\}. \end{aligned} \quad (5.40)$$

This makes the estimation more robust as more samples, e.g., $\lfloor \frac{K}{2} \rfloor + 1$ samples, are considered for the minimization problem.

For multiple sources, we consider the simplest scenario, where the observation $\mathbf{x}_{p,k}$ contains only two uncorrelated sources, e.g., $s_{p,k}^1$ and $s_{p,k}^2$. Hence, the observation $\mathbf{x}_{p,k}$ can be written as:

$$\mathbf{x}_{p,k} = \mathbf{x}_{p,k}^1 + \mathbf{x}_{p,k}^2. \quad (5.41)$$

The reconstructed covariance matrix $\mathbf{R}_{\mathbf{x}\mathbf{x},p,k}$ by considering all frequency bins can be expressed as:

$$\mathbf{R}_{\mathbf{x}\mathbf{x},p,k} = \frac{1}{\lfloor \frac{K}{2} \rfloor + 1} \left(\left[\sum_{l=1}^L \sum_{k', \tilde{k} \in \mathcal{K}} \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right) \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right)^H \right] + \mathbf{n}_{p,k} \right), \quad (5.42)$$

where $\mathbf{n}_{p,k}$ denotes a noise term due to cross terms between $\mathbf{x}_{p,k'}^i$ and $\mathbf{x}_{p,\tilde{k}}^j$ for $i, j = [1, 2]$ and $i \neq j$. The performance can be affected by the noise term. Since the two sources are uncorrelated, the contribution of the noise term is smaller than the main terms $\sum_{l=1}^L \sum_{k', \tilde{k} \in \mathcal{K}} \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right) \left(\mathbf{x}_{p,k'}^l \circ \mathbf{x}_{p,\tilde{k}}^l \right)^H$. Moreover, the two sources can have different pitches. The frequency occupied by the sources will be distinct to each other. By summing over all frequencies, the main terms will be dominant over the cross terms.

There are few strategies that can be used to minimize the effect caused by the noise term. This can further improve the robustness of the system. The details of the methods will be discussed in the next subsection.

5.4.3 Robustness Improvement

In this subsection, we discuss two approaches to improve the robustness of the system. The first method is an approach to approximate the covariance matrix while the second approach is diagonal loading.

Tyler's covariance estimator [183] is a well-known approach to approximate the covariance matrix due to its simplicity. It is also robust to outliers and guaranteed to converge [199]. Tyler showed that given M -dimensional independent and identically distributed (iid) samples, e.g., $\mathbf{x}_i$, where $i = 1, \dots, N$, which is drawn from an elliptical distribution, a good covariance estimator is the solution to the fixed point equation

$$\mathbf{Q} = \frac{M}{N} \sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{Q}^{-1} \mathbf{x}_i}. \quad (5.43)$$

Tyler's covariance estimator is a minimax robust estimator of the covariance matrix. In addition, it is distribution-free in the sense that its asymptotic

variance does not depend on a parametric form of the underlying distribution [171].

Diagonal loading is a useful tool to improve the robustness of matrix inverses. It simply adds a scaled identity matrix to the covariance matrix [94]. The diagonal loading technique can be seen as a form of Tikhonov regularization [137]. It improves the system robustness by adding a penalty function to reduce the $\mathcal{L}_2$ norm of the estimated parameters, so that the performance will not be significantly affected by perturbations as discussed in Sec. 2.5.3. Furthermore, ill-posed inverse problems can be avoided as the covariance matrix will always be full rank after diagonal loading.

By using Tyler's covariance estimator and the diagonal loading technique, the l^{th} desired source can be obtained by using the GSC method as:

$$\hat{s}_{p,k}^l = \left(\mathbf{w}_k^l - \mathbf{B}_k \hat{\mathbf{d}}_{p,k}^l \right)^H \mathbf{x}_{p,k}, \quad (5.44)$$

$$\hat{\mathbf{d}}_{p,k}^l = \left(\mathbf{B}_k^H \left(\hat{\mathbf{R}}_{\mathbf{xx},k}(p) + \lambda \mathbf{I} \right) \mathbf{B}_k \right)^{-1} \mathbf{B}_k^H \left(\hat{\mathbf{R}}_{\mathbf{xx},k}(p) + \lambda \mathbf{I} \right) \mathbf{w}_k^l, \quad (5.45)$$

where $\hat{\mathbf{R}}_{\mathbf{xx},p,k}$ is Tyler's covariance estimator of $\mathbf{R}_{\mathbf{xx},p,k}$ and λ is the diagonal loading value.

5.5 Results

In this section, we first discuss the experimental setup and procedure. Then we provide the experimental results of our proposed method and comparisons with the conventional GSC approach [53] for both the scenarios when the interfering source is static and moving. The separation performance was evaluated using the signal-to-interference ratio (SIR) between the original source and the separated signal. The SIR values were computed using the BSS_EVAL toolbox [189].

5.5.1 Experimental Setup and Procedure

In all experiments, the sampling rate was set to be 8000 Hz. Two speech signals were used as desired sources and one chirp signal was used as interfering signal. The entire spectrum of the chirp signal was occupied and the length of one cycle was 1024 samples (128 ms). The length of all the signals was three seconds. The transfer function between microphone m and source l was computed using the free-field approximation:

$$a_k^{ml} = \frac{e^{-j\frac{2\pi}{K}\frac{d_{ml}}{c}f_s k}}{4\pi d_{ml}}, \quad (5.46)$$

where $K = 1024$ is the number of FFT bins, $f_s = 8000$ is the sampling rate, $c = 330$ is the speed of sound (m/s), and d_{ml} is the Euclidean distance between microphone m and source l . A square-root of Hann window with 1024 taps and 50% overlap was used to perform the short-time Fourier transform. Hence, the algorithmic delay of the system is equivalent to one window length, which was 128 ms.

In the first set of the experiments, the arrangement of the desired sources and the microphones are displayed in Fig. 5.2. The black crosses indicate the microphones, which were arranged in a circular array with an inner radius of 0.5 m at the center of the room [5 m, 5 m]. The blue dot indicates the first desired source while the red dot indicates the second desired source. For each trial, the interfering source was randomly placed in the room. The experiment was performed 30 times.

For the moving interfering source, we conducted two sets of experiments with different moving speed, i.e., at 1 m/s (a walking waiter) and at 10 m/s (a moving vehicle). The initial position and the moving direction of the interfering source were set randomly, but the interfering source was guaranteed to reach the point [10 m, 5 m] at 1.5 s. This was to ensure that the interfering source approached the microphones array as the length of the signals were three seconds. We also assume that direction of the interfering source were not be changed when it was moving. Each set of the experiments was

Figure 5.2: The positions of the microphones and the desired sources. The black crosses indicate the microphones, the blue dot indicates the first desired source while the red dot indicates the second desired source.

repeated by 30 times. Fig. 5.3 illustrates the positions of the microphones, desired sources and the moving interference source in the two sets of experiments by using the parameters that were set in the first trial. The black crosses represent positions of the microphones. The blue and red dots indicate the two stationary desired sources. The green dot indicates the initial position of the interfering source while the green line indicates its moving path.

We compared the separation performance of our proposed method with the conventional GSC method. The performance, for the case where the BSS was only performed by using the previous demixing matrix without post-processing by a GSC method, was included as well. We also considered the performance of the proposed method combined with various robust improvement strategies, such as Tyler’s covariance estimator and the diagonal loading technique.

The diagonal loading value λ was set to be $\text{std}\{\text{diag}\{\mathbf{R}\}\}$ [105], where

(a) Moving speed = 1 m/s.

(b) Moving speed = 10 m/s.

Figure 5.3: The positions of the microphones indicated by black crosses, desired sources indicated by red and blue dots as well as the moving interference signal, where the green dot indicates the initial position while the green line indicates the moving path.

$\text{std}\{\mathbf{r}\}$ indicates the standard deviation of $\mathbf{r}$ and $\text{diag}\{\cdot\}$ denotes a diagonal operator. The demixing matrix $\mathbf{W}_k$ was computed using the pseudoinverse of $\mathbf{A}_k$, i.e., $\mathbf{W}_k = \mathbf{A}_k (\mathbf{A}_k^H \mathbf{A}_k)^{-1}$.

All estimations were performed using only single snapshot of observation data. The previous demixing matrix $\mathbf{W}_k$ was obtained using (5.2) based on the oracle information of the mixing channels.

5.5.2 Experimental Results for a Static Interfering Source

Fig. 5.4 illustrates the source separation performance using various techniques for a static interfering source. The blue and red circles represent the means of the SIR values over 30 trials of the first source and the second source, respectively. The error bars indicate the standard deviation.

Based on the results, the conventional GSC method performed the worst as the sample covariance matrix was inaccurate due to an insufficient number of data samples since one observation data was only used for estimations

in the experiment. The conventional GSC method tends to suppress the contribution of the desired sources as well as the interfering signal. This is similar to a linear regression problem with one observation data that contains noise; the best-fitting line with no error can always be found because it compensates with the noise term as well as the variable. Hence the method gives poor results.

The separation performance can be significantly improved by using the proposed method. We observed that the performance of the proposed method can be barely improved by using Tyler's covariance estimator and the diagonal loading technique. The improvement was not significant for this case. This showed that the contribution of the noise term, which is caused by the cross-terms between every pair of distinct sources, is small compared to the main terms that are computed using the same source.

Figure 5.4: Separation performance of various methods when a new static interfering source is introduced to the BSS system.

5.5.3 Experimental Results for a Moving Interfering Source

Fig. 5.5 and Fig. 5.6 illustrate the source separation performance using various techniques for a slow-moving and a fast-moving interfering source, respectively. The blue and red circles represent the means of the SIR values over 30 trials of the first source and the second source, respectively. The error bars indicate the standard deviation.

Based on the results, the proposed method outperformed the conventional GSC approach and the “Unprocessed” method. The conventional GSC approach still performed the worst as the sample covariance matrix was inaccurate by considering only one observation data. The overall results for the moving interference source were consistent with the case for a static interference source. This implies that the proposed method can be employed to cope with the BSS scenario, when a new moving interfering source is introduced.

Figure 5.5: Separation performance of various methods when a new slow-moving interfering source (1 m/s) is introduced to the BSS system.

Figure 5.6: Separation performance of various methods when a new fast-moving interfering source (10 m/s) is introduced to the BSS system.

We also include the separation results obtained in the first trial of the experiments in Fig. 5.7 for the slow-moving interfering source and Fig. 5.8 for the fast-moving interfering source. The first and the second rows indicate the first and the second original sources, respectively. The third row shows the first observation signal. The fourth and the fifth rows represent the first and the second separated sources using the “Unprocessed” method, respectively. The separated sources using the conventional GSC method are illustrated in the sixth and the seventh row. The eighth and the ninth rows indicate the separated sources using the proposed method discussed in Sec. 5.4.2.

From both the figures, we can clearly see that the interfering signal, i.e., the chirp signal, is present in the separation results that were obtained by using either the “Unprocessed” method or the conventional GSC approach. For the conventional GSC approach, we observe that the low-frequency components of the chirp signal (the peaks) were not suppressed as much as the high-frequency components. This is because the mixing vectors of two sources

Figure 5.7: Partial results obtained in the first trial of experiments for a slow-moving interfering source.

in high frequencies are much different compared to the mixing vectors in low frequencies.

Figure 5.8: Partial results obtained in the first trial of experiments for a fast-moving interfering source.

5.6 Conclusion

In this chapter, we proposed a method to improve the performance of the conventional GSC beamforming technique while dealing with an interfering source. The proposed method allows the estimation to be performed by using only a single snapshot of observation data, so the estimation delay can be significantly reduced compared to the conventional GSC beamformer.

We utilize the entire spectrum to create multiple measurements in every frequency bin, hence the reconstructed covariance matrix is more accurate and robust. The proposed method can be applied for both static and moving interfering sources and it achieves good separation performance.

Since the proposed method is able to reconstruct an accurate covariance matrix with single observation data, it can be particularly useful for practical real-world BSS applications that desire low latency. By combining the proposed method with the GSC beamformer, the previous estimated demixing matrix can be used as “meta-information” to facilitate the later operation. Thus, the separation parameters can be updated frequently to ensure good source separation performance at all times.

Chapter 6

Non-Iterative Impulse Response Shortening Method

In this chapter we present a non-iterative impulse response shortening method aiming to reduce the latency of a system. Our method exploits that smoothing the frequency-domain response generally leads to a shorter time-domain response. The method is simple to implement and has a computational complexity that is significantly lower than that of competing methods. Yet it achieves good performance. It can be used for applications involving system identification such as blind source separation (BSS), cross-talk cancellation and channel equalization. Our experimental results confirm the effectiveness of the method, demonstrating the benefit of the approach in the BSS and cross-talk cancelling applications.

6.1 Introduction

In recent years, impulse response shortening methods have been studied widely for different applications. These include, but are not limited to, blind source separation (BSS) [110, 112, 113], speech dereverberation [200], channel equalization [114] and cross-talk cancellation [19, 87]. It has been used to minimize the artifacts resulting from circular convolution [112, 113] associ-

ated with the use of the fast Fourier transform and to compensate for the room reverberation with minimal latency [19,87,114]. We distinguish latency due to estimation computational time and due to algorithmic delay.

BSS is an example application where the latency problem occurs. To resolve the algorithmic latency problem, we [33] proposed a cross-band filtering approach, where the details are discussed in Chapter 4, to compute a low-frequency-resolution time-frequency (LFRTF)-domain representation of the mixing filters using the high-frequency-resolution time-frequency (HFRTF)-domain mixing matrices. It reduces the time-lag of the system once the calculation is completed but the computational effort to estimate the cross-band filters is high compared to that of the basic HFRTF-domain approach. Hence, the cross-band filtering approach is only useful for stationary scenarios, where the source locations are fixed in time and the estimates can be updated infrequently.

The latency problem also occurs in cross-talk cancellation application. Cross-talk cancellation aims to deliver multiple signals to multiple listeners independently and simultaneously. Differently to BSS, the information of the room impulse response (RIR) is usually complete during the designation of the pre-filters. The pre-filters are used to process the signals before propagating to the listeners. The time-domain approaches [19,87] have been proposed to address this problem but suffer from high computational cost when the length of the RIR increases.

Impulse response shortening can be applied to both the BSS and cross-talk cancellation applications to reduce the algorithmic latency. However, all the aforementioned approaches except [113] involve $\mathcal{L}_1$ -norm or $\mathcal{L}_\infty$ -norm minimization, which can lead to a slow convergence rate, prolonging the processing time. This again results in a system with long estimation delay and it is undesirable to real-time applications. Although [113] facilitates the finding of the optimal solution by using the least-squares method, it is difficult to select the correct parameters and involves inversions of large matrices.

In this chapter, we propose a non-iterative method to perform impulse response shortening to reduce the system latency. It reduces both the computational burden and shortens the time for signal acquisition. Our approach is based on the fact that the spectrum of a signal becomes smoother when the signal is zero-padded in the time-domain. Instead of computing the scaling factors by finding the sparsest representation of the RIR in the time domain, we search for the complex scaling factors in the frequency domain that results in the smoothest spectrum. We demonstrate the advantage of the method in the BSS and cross-talk cancelling applications.

The chapter first starts with an introduction of the system identification model. Second, we provide our proposed method for impulse response shortening. Next, we discuss the implementations of the proposed approach in BSS and cross-talk cancellation applications. Then, we present the simulation results of our proposed method.

6.2 Impulse Response Shortening

In this section, we first review the model of system identification and discuss the approach for impulse response shortening of [113]. Next, we introduce our proposed method.

6.2.1 System Identification Model

We denote by h_k^m the frequency response of the RIR between the microphone m and a source, which is labelled as s in frequency bin k . When the length of the window N_w is at least twice the length of the RIR N_h and it is generally considered that the linear system, the observation mixtures in the HFRTF domain in a noiseless scenario can be approximated as [42]:

$$\mathbf{x}_{p,k} \approx \mathbf{h}_k s_{p,k}, \quad (6.1)$$

where $\mathbf{x}_{p,k}$ and $s_{p,k}$ denote the vectors of the observations and the source, respectively, at time-block index p and frequency bin k . The vector $\mathbf{h}_k$ con-

tains the frequency responses between all the microphones and the source, e.g., $\mathbf{h}_k = [h_k^1 \ \cdots \ h_k^M]^\text{H}$, where M indicates the number of microphones.

Denote by $\mathbf{H}^m$ the estimated RIR between microphone m and the source, which is a diagonal matrix, i.e., $\mathbf{H}_{k,k}^m = \mathbf{h}_k^m$. To shorten the estimated RIR $\mathbf{H}^m$, a complex scaling factor c_k can be applied to $\mathbf{H}_{k,k}^m$ in each frequency bin to obtain a time-domain response. The complex scaling factor introduces a time-shift to each frequency component signal. When the factors are chosen suitably, the frequency component signals can be aligned such that a short time response is obtained. This causes a filtering effect to the recovered signal [123]. The filtering effect is generally not a significant issue in most applications. This is particularly true for BSS, where the scaling ambiguity is usually present in the frequency-domain approaches.

The approach of [112,113] now uses a short-time Fourier transform (STFT) and the time-domain response can be expressed as $\mathbf{V}^m \mathbf{c}$, where

$$\mathbf{V}^m = \sum_{N_l} \mathbf{E}_{N_l} \mathcal{F}^{-1} \text{diag}(\mathcal{F} \mathbf{D}_{N_l} \delta) \mathbf{H}^m, \quad (6.2)$$

and $\mathbf{c} = [c_0 \ \cdots \ c_{N_w-1}]^\text{T}$, is a vector of scaling factors. $\mathcal{F}$ is a discrete Fourier transform matrix and $\text{diag}(\cdot)$ denotes an operator that converts a vector to a diagonal matrix. $\mathbf{D}_{N_l}$, which is a diagonal matrix, defines the analysis window that is shifted to the l position. $\mathbf{E}_{N_l}$ is a shifting matrix that ensures the overlapping block are merged correctly. Note that the window length in $\mathbf{D}_{N_l}$ is shorter than N_w .

To perform impulse shortening, the authors [113] approximate the time-domain response to a pulse-like response, such as a delta function. The delta function is the shortest response as it contains only a single one and zeros otherwise. An optimal set of the complex scaling factor $\mathbf{c}$ can be obtained by minimizing

$$\|\mathbf{d} - \mathbf{V} \mathbf{c}\|_2, \quad (6.3)$$

where, $\|\cdot\|_2$ represents an $\mathcal{L}_2$ -norm operator,

$$\mathbf{V} = [\mathbf{V}^{1\text{T}} \ \cdots \ \mathbf{V}^{M\text{T}}]^\text{T}, \quad (6.4)$$

$$\mathbf{d} = \begin{bmatrix} \mathbf{d}^{1^T} & \cdots & \mathbf{d}^{M^T} \end{bmatrix}^T, \quad (6.5)$$

where $\mathbf{d}^m$ denotes the delta function that contains a single one at the peak position of $\mathbf{V}^m \mathbf{1}$. A trivial solution of (6.3) will be $\mathbf{c} = \mathbf{V}^\dagger \mathbf{d}$, where $\{\cdot\}^\dagger$ indicates a Moore-Penrose pseudoinverse.

In practice, (6.3) is not a good criterion to be minimized. It requires the estimated RIR $\hat{h}_m$ to be sparse, i.e., the amplitude of the main peak should be relatively large compared to the other peaks. This is not the case in some applications, e.g. BSS, as the estimated mixing matrices suffer from the so-called scaling ambiguity across the spectrum. More details will be discussed in Section 6.3. In addition, it needs to perform an inverse operation of large matrices, which increases the computational burden.

6.2.2 Proposed Method

In this subsection, we propose a non-iterative method to perform impulse response shortening. Similarly to [112, 113], we introduce a new set of complex scaling factors to shorten the estimated RIR.

The motivation for our method is based on the fact that zero-padding a signal in the time domain leads to a smoother frequency spectrum. We hypothesize that this implies that smoothing the frequency spectrum leads to a short response. Instead of finding the scaling factors by making the estimated RIR sparse in the time domain as done in [112, 113], we search for a new set of complex scaling factors in the frequency domain that lead to the smoothest spectrum.

A smooth spectrum can be obtained by altering the coefficients of the frequency response $\mathbf{h}_k$, such that they are maximally similar to each other in adjacent bins. These coefficients can be estimated by minimizing the following equation:

$$\min_{c_k} \|\mathbf{h}_{k-1} - c_k \mathbf{h}_k\|_2, \quad k = 1, \dots, L_w - 1. \quad (6.6)$$

The closed-form solution for (6.6) is given as:

$$c_k = \frac{\mathbf{h}_k^H \tilde{\mathbf{h}}_{k-1}}{\mathbf{h}_k^H \mathbf{h}_k}, \quad (6.7)$$

where $\{\cdot\}^H$ denotes a Hermitian transpose. Hence, the frequency response $\tilde{\mathbf{h}}_k$ that varies smoothly in frequency can be computed as:

$$\tilde{\mathbf{h}}_k = \begin{cases} \frac{\mathbf{h}_k}{\|\mathbf{h}_k\|}, & k = 0, \\ \frac{c_k \mathbf{h}_k}{\|c_k \mathbf{h}_k\|}, & k = 1, \dots, L_w - 1. \end{cases} \quad (6.8)$$

Our proposed approach differs from [112, 113] as the desired time-domain responses are not required. It does not involve matrix inversion. Unlike [19, 87], our method is non-iterative. Hence, the computational efficiency is higher, and is simpler to implement. We show that our proposed method can be implemented to design both the post-filters as shown in Section 6.3 for BSS applications and the pre-filters as presented in Section 6.4 for cross-talk cancellation implementations.

6.3 Blind Source Separation Application

This section first reviews the formulation of the BSS problem in the TF domain. Next, we apply our proposed approach for impulse response shortening. This facilitates the computation of the representations of the LFRTF-domain mixing filters. Then, we estimate the demixing operator in the LFRTF-domain based on the shortened mixing filters.

6.3.1 Problem Formulation of BSS

In this subsection, we first briefly provide the necessary background for BSS. We neglect the effect of noise in the derivation and consider the over-determined scenario, where the number of microphones M is larger than the number of original sources L , i.e. $M > L$.

The observation mixtures in the HFRTF domain can be written as:

$$\mathbf{x}_{p,k} \approx \mathbf{A}_k \mathbf{s}_{p,k}, \quad (6.9)$$

where $\mathbf{x}_{p,k}$ and $\mathbf{s}_{p,k}$ now denote the vectors of the observation mixtures and the original sources at time-block index p and frequency bin k , respectively, while $\mathbf{A}_k$ represents the mixing matrix at frequency bin k . The mixing matrices $\mathbf{A}_k$ (or, alternatively, the demixing matrices), can be estimated using the aforementioned BSS algorithms [2, 71, 84, 134, 148, 198].

To minimize the algorithmic latency, the BSS problem can be solved by using a cross-band filtering approach [33]. The approach designs the demixing operators based on the mixing filters in the LFRTF domain, which are obtained from the HFRTF-domain mixing matrices. Hence, the methods that aim to estimate $\mathbf{A}_k$ [2, 148, 198] are preferred.

The estimation delay is large as the computational effort is extremely high if the estimated mixing matrix, denoted as $\hat{\mathbf{A}}_k$, is directly used in [33]. Due to the scaling ambiguity, a random complex scaling factor is introduced into each column of the HFRTF-domain mixing matrix in every frequency bin. This causes a random time-shift in each frequency band signal and leads to a long RIR.

The computational efficiency in [33] can be improved by shortening and truncating the mixing filters beforehand. In this case, the effect of the cross-band filters will be minimal and can be neglected. This significantly improves the calculation speed. The details will be discussed in the next subsection.

6.3.2 Estimating the Demixing Operator

We follow the method described in [33] to estimate the demixing operators in the LFRTF domain, which is based on the HFRTF mixing filters. To perform separation, we need to compute the representation of the shortened mixing filters in the LFRTF domain. The shortened mixing filters can be obtained by using the proposed method in Section 6.2.2. It is done by repeating (6.8)

and (6.7) for each column of $\hat{\mathbf{A}}_k$ for every frequency bin, e.g. $\mathbf{h}_k$ is replaced by the n^{th} column of $\hat{\mathbf{A}}_k$ and $\tilde{\mathbf{h}}_k$ is replaced by the n^{th} column of $\tilde{\mathbf{A}}_k$, where $n = 1, \dots, L$.

To facilitate the computation in [14], the length of the truncated RIR is desired to be the length of the window in the LFRTF domain. This is to diminish the effect of the cross-band filters. The truncation can be done by applying a rectangular window to capture the segments containing the highest $\mathcal{L}_1$ -norm, so that the maximal information is retained.

Denote by $\tilde{a}_t^{ml}$ the processed RIR between the m^{th} microphone and the l^{th} source in the time domain, where t indicates the time index. Note that $\tilde{a}_t^{ml}$ can be obtained by performing inverse Fourier transform on $\tilde{\mathbf{A}}_k^{ml}$, where m and l denote the m^{th} row and the l^{th} column, respectively. Let $\tilde{\mathbf{a}}_t^l$ represents the RIR between all the microphones and the l^{th} source, i.e., $\tilde{\mathbf{a}}_t^l = \begin{bmatrix} \tilde{a}_t^{1l} & \dots & \tilde{a}_t^{Ml} \end{bmatrix}^T \in \mathbb{R}^{M \times 1}$. Furthermore, let $\tilde{\mathcal{A}}_l$ represents a concatenated version of $\tilde{\mathbf{a}}_t^l$ with all the time indexes, i.e., $\tilde{\mathcal{A}}_l = \begin{bmatrix} \tilde{\mathbf{a}}_0^{1l} & \dots & \tilde{\mathbf{a}}_{N_w-1}^{1l} \end{bmatrix} \in \mathbb{R}^{N_w \times M}$.

Without loss of generality, we rotate the RIR, such that the segments are located in the middle. The truncated RIR can be represented as:

$$\bar{\mathcal{A}}_l = D_l^* \tilde{\mathcal{A}}_l, \quad (6.10)$$

$$D_l^* = \arg \max_{D_l \in \mathcal{D}} \mathbf{1}^T \left| D_l \tilde{\mathcal{A}}_l \right| \mathbf{1}, \quad (6.11)$$

where $\mathcal{D}$ is a set of zero matrices with an identity matrix located at the i^{th} column index:

$$\mathcal{D} = \left\{ \begin{bmatrix} \mathbf{0}^{N_a \times i} & \mathbf{I}^{N_a \times N_a} & \mathbf{0}^{N_a \times (N_w - N_a - i)} \end{bmatrix} \mid i \in [0, N_w - N_a] \right\}, \quad (6.12)$$

where N_a is the length of the ST window in the LFRTF domain.

The truncated RIR $\bar{a}^{ml}$ can be found by computing (6.10) and (6.11) repeatedly for $l = 1, \dots, L$. Then, we calculate the LFRTF-domain representation of the band-to-band mixing filters [14] and design the demixing operators in the LFRTF domain using the method described in [33] (Section 4).

6.4 Cross-talk cancellation application

This section first briefly reviews the problem definition of the cross-talk cancellation. Then, we implement the proposed response shortening approach to design pre-filters for cross-talk cancellation.

The objective of the cross-talk cancellation is to deliver the signals from Q loudspeakers to R listeners independently and simultaneously. In general, $Q \geq R$. This can be done by designing $R \times Q$ pre-filters to compensate for the $Q \times R$ RIR between the loudspeakers and the listeners.

Let us define g^{rq} as the time-domain pre-filter that compensates for the signal, which propagates from the q^{th} loudspeaker to the r^{th} listener. Let us denote by b^{qr} the RIR between the loudspeaker q and the listener r in the time domain. A time-domain approach to finding the pre-filters can be found in [19, 87]. Both methods involve norm minimization and are implemented iteratively, which can lead to a reduced convergence rate and prolong the estimation delay.

We simplify the problem by transforming the problem into the frequency domain. Hence, the pre-filter can be estimated in every frequency bin in the HFRTF domain independently:

$$\mathbf{G}_k \mathbf{B}_k = \mathbf{I}, \quad (6.13)$$

where $\mathbf{G}_k \in \mathbb{C}^{R \times Q}$ and $\mathbf{B}_k \in \mathbb{C}^{Q \times R}$ represent the frequency responses of the pre-filters and the RIR in frequency bin k , respectively.

To shorten the pre-filters, we apply the approach proposed in Section 6.2. Although the scaling ambiguity is not present in the cross-talk-cancellation application (the RIR are generally fully known), the performance is generally not significantly affected. The shortened pre-filters can be obtained by substituting $\mathbf{h}_k = \mathbf{g}_k^r$, where $\mathbf{g}_k^r = \begin{bmatrix} g_k^{r1} & \cdots & g_k^{rQ} \end{bmatrix}$, and $\tilde{\mathbf{h}}_k = \tilde{\mathbf{g}}_k^r$ in both (6.8) and (6.7) for every frequency bin k .

6.5 Results

In this section, we discuss the experimental results for impulse response shortening for both the BSS and cross-talk cancellation applications. In both cases, we first provide the setup and then the simulation results. We note that the main focus is to reduce the system latency. Hence, we focus on the computational effort and the separation performance in the performance evaluation.

6.5.1 BSS Experimental Setup

Three 10 second speech signals sampled at 16 kHz, which were obtained from Stereo Audio Source Separation Evaluation Campaign (SASSEC) [190], were used. We altered the activity period, such that, for each source, certain periods existed where only one source was active. This is not necessary but guarantees that the mixing matrices can be correctly estimated using the sparsity-based method [2, 148, 198]. All the separations were conducted offline and computed by Matlab R2015b on a PC having an Intel(R) Core(TM) i5-5200 CPU@2.20 GHz processor with 8GB random-access memory.

In the simulation, 24 microphones were used. The observations were obtained by convolving the speech signals with the simulated room impulse responses (RIR). 24×3 RIR with 1024 taps were computed using the image-source method [5], where the microphones and sources were randomly placed in a room with a size of $3 \text{ m} \times 3 \text{ m} \times 3 \text{ m}$. The reverberation time was 0.2 s.

Square-root of Hann windows with 2048 taps (128 ms) and 512 taps (32 ms) were used in the high-frequency-resolution time-frequency (HFRTF) domain and in the low-frequency-resolution time-frequency (LFRTF) domain, respectively. The windows were 50% overlapped. The HFRTF-domain mixing matrices were estimated using both the independent component analysis (ICA) method [71] and the Modified-TIFROM approach [2, 148]. The Modified-TIFROM approach is an improved version of the Time-Frequency Ratio Of Mixtures (TIFROM) method by combining [2] and [148]. Instead

of computing variances for each row of the observation vectors, we compute variances of the angle between observation vectors in neighbouring blocks. The permutation ambiguity was resolved using oracle information, so that the separation performance was not affected by the permutation issue.

We examined five different approaches for each estimation method. The stand-alone strategy identified the demixing matrices in the LFRTF domain directly while the crossband approach [33] utilized the HFRTF domain mixing matrices to compute the LFRTF-domain mixing filters and designed the LFRTF-domain demixing operators. In addition to our proposed approach, we tested two different impulse response shortening methods for comparison. The minimal approach shortens the estimated RIR by resolving the scaling ambiguity based on the minimal distortion principle [110]. The method proposed by Mazur et. al. [113] is described in Section 6.2. It obtains an optimal set of complex scaling factors by approximating the time-domain shortened filters to a desired pulse-like response. After shortening, the estimated HFRTF-domain mixing filters were truncated as suggested in Section 6.3.2 and the demixing operators were designed based on the LFRTF-domain representation of the truncated mixing filters.

6.5.2 BSS Simulation Results

The BSS_EVAL toolbox [189] was used to compute the signal-to-interference ratio (SIR) between the separated source and the original source to indicate the source separation performance. A higher score indicates better performance. We also compare the computation time of each method. The results of all the approaches are tabulated in Table 6.1 and Table 6.2. The core BSS estimation methods were the ICA approach [71] and the Modified-TIFROM method [1, 148] in Table 6.1 and Table 6.1, respectively. SIR_l represents the SIR of the l^{th} while SIR_{avg} indicates the average value of the SIR values in each method.

The results show that the proposed approach achieves the highest average

Table 6.1: Performance comparison between various approaches low-latency for BSS using the ICA method.

Metrics	ICA				
	Stand-alone	Crossband [33]	Minimal [110]	Mazur [113]	Proposed
SIR_1 (dB)	9.70	16.47	13.62	8.24	18.32
SIR_2 (dB)	10.90	12.42	15.64	13.18	16.20
SIR_3 (dB)	10.69	19.13	13.66	15.74	15.60
SIR_{avg} (dB)	10.43	16.01	14.31	12.38	16.71
Time (s)	5.39	66.71	20.52	158.35	20.37

SIR value using both the ICA approach and the Modified-TIFROM method. In terms of the computation time for the estimation of the response, the proposed approach was slower than the stand-alone method but is significantly faster than the state-of-the-art procedures.

6.5.3 Cross-talk Cancellation Experimental Setup

In the simulation, four loudspeakers and two microphones, acting as listeners, were randomly located in a room with a size of $3\text{ m} \times 3\text{ m} \times 3\text{ m}$. The room impulse responses (RIR) with 1024 taps were generated using the image source method [5], where the reverberation time was 0.2 s. The sampling rate was 16 kHz.

We compare the proposed approach with [19]. The $\mathcal{L}_2$ -norm was chosen as a criterion to be minimized to facilitate the computation in [19].

6.5.4 Cross-talk Cancellation Simulation Results

The performance of the cross-talk cancellation is measured using a direct signal to cross-talk ratio (DSCR) [78], which is the ratio of the maximum direct response to the maximum cross-talk. High DSCR value indicates good cross-

Table 6.2: Performance comparison between various low-latency approaches for BSS using the Modified-TIFROM approach.

Metrics	Modified-TIFROM				
	Stand-alone	Crossband [33]	Minimal [110]	Mazur [113]	Proposed
SIR_1 (dB)	5.31	16.24	14.62	17.77	18.72
SIR_2 (dB)	20.14	22.31	22.87	19.25	25.45
SIR_3 (dB)	12.82	18.44	16.04	21.07	21.26
SIR_{avg} (dB)	12.76	18.99	17.84	19.36	21.81
Time (s)	5.01	63.45	18.13	152.72	18.24

talk cancellation performance. Fig. 6.1 shows the cross-talk cancellation performance of our approach. The DSCR values between the direct response and the cross-talk for the first signal and for the second signal were 17.81 dB and 17.59 dB, respectively. The computational time was 0.10 s. For the approach in [19] with the same setup, the DSCR values were 17.67 dB and 16.98 dB while the computational time was 3.26 s. Both methods obtained similar DSCR values but our proposed method was significantly faster than the method of [19].

In Fig. 6.2, we compare the performance of the two approaches against the length of the pre-filters. The results show that [19] surpasses our proposed approach when the filter length is longer. This is a consequence of the fact that the proposed method designs the pre-filters in the frequency domain, resulting in the distortion associated with circular convolution. Long filters make this distortion more severe. However, our method is more efficient than [19] independently of the length of the pre-filters.

Figure 6.1: Cross-talk cancellation performance of the proposed approach, where $Q = 4$ and the pre-filter length was 128.

Figure 6.2: Performance comparison with different pre-filter length.

6.6 Conclusion

In this chapter, we presented a non-iterative impulse response shortening method. It aims to reduce the system latency, including both the signal acquisition time and the processing time. It exploits the advantage that zero-padding a time-domain signal yields a smooth frequency spectrum. Thus, we are able to obtain a sparser representation of the room impulse response

(RIR) by computing a set of complex scaling factors in the frequency domain, which yields the smoothest spectrum. It is a non-iterative method and it does not require matrix inversion. Hence, it is computationally efficient and easy to be implemented, which provides a simple and practical solution for real-time applications involving system identification. The simulation results show that the approach can be used to design both post-filters for the blind source separation applications and pre-filters for the cross-talk cancellation applications.

Chapter 7

Conclusions and Future Work

In this chapter we draw general conclusions from the contributions of this thesis. Some future research directions that arise from our works are also outlined.

7.1 Conclusions

In this thesis, we have addressed the major methodological limitation that encountered by most conventional blind source separation (BSS) algorithms in practical real-world applications, which is the system latency problem. The system latency can be separated into an algorithmic latency and an estimation delay. We proposed several approaches to address the latency problem, so that BSS algorithms can be implemented in real-world applications with minimal latency.

We now provide conclusions from our contributions for each of the latency problems.

7.1.1 Algorithmic latency

In digital signal processing applications, an algorithmic latency of one window length is unavoidable while using time-frequency (TF) domain BSS ap-

proaches due to the overlap-add method. For conventional TF-domain BSS methods, the window length has to be at least twice the channel response, so that the instantaneous mixtures model in the TF domain is sufficiently accurate to describe the time-domain mixtures model. In practice, the channel response can be long due to reverberation, and the algorithmic latency is outside the range that is acceptable for devices.

In Chapter 3, we propose a subspace method that is able to perform BSS by using a window length that is shorter than the channel response. The mixture model is now in the low-frequency-resolution time-frequency (LFRTF) domain. In the LFRTF domain, the mixing process is still a convolutive process but the length of the mixing filters is significantly shorter than the ones in the time domain. This facilitates the estimation process as the size of the matrices that are used for analysis, is small compared to the ones in the time domain. By using the proposed method, an instantaneous mixture model in the LFRTF domain can be obtained by selecting the correct length for the mixing filters in the LFRTF domain. The demixing parameters can then be solved by using an independent component analysis (ICA) approach. Simulation results show the proposed method is able to reduce the algorithmic latency of the system and achieves satisfactory performance. The drawback of the proposed method is that the separation performance can be adversely affected when an unsuitable length is selected for the length of the LFRTF-domain mixing filters.

In Chapter 4, we propose an alternative method to address the algorithmic latency problem. The proposed method can be integrated with most existing TF-domain BSS methods. The idea is to perform the estimation using conventional TF-domain approaches in the high-frequency-resolution time-frequency (HFRTF) domain, where the window length is at least twice the length of the mixing channels. The HFRTF-domain mixtures can be described by a system of linear equations. When we decrease the window length, the number of variables increases as the length of the mixing filters becomes larger. To compensate for this problem, we add more microphones

to increase the number of equations. This allows us to cancel the non-causal components of the mixing filters when designing the separation parameters either in the time domain or in the LFRTF domain. Theoretically, the window length for the separation process can be reduced to any desired length by increasing the number of microphones. Experimental results show that our proposed method can be combined with existing TF-domain BSS methods to perform BSS with a low delay while the separation performance can still be retained. We conclude that the proposed method allows integration with existing TF-domain BSS methods to shorten the algorithmic latency and achieve good separation performance. This makes BSS possible for real-time implementations.

7.1.2 Estimation delay

The estimation delay can be caused by two factors. The first factor is due to the methodological limitation of the methods. For instance, a sample covariance matrix is required in most of the beamforming methods. Hence, a large number of samples is required to make the sample covariance matrix useful. This increases the estimation delay as the analysis can only be performed after acquiring enough data samples. The second factor is due to the computational complexity of the methods. For instance, iterative algorithms with a large amount of data can be time-consuming.

In Chapter 5, we proposed a method to address the BSS scenario when a new, relatively loud, source is introduced into the system. The new source can be either static or moving. We show that “meta-information”, such as separation parameters that are estimated from previous operations, in combination with a conventional generalized sidelobe canceller (GSC) method can be employed to extract the desired sources. However, a covariance matrix is needed for the GSC method and it requires a sufficient number of samples to make the sample covariance matrix accurate. This prolongs the estimation delay. By using the proposed method together with the GSC

method, the estimation can be done by using a single snapshot of observation data. Hence, the estimation delay can be drastically reduced. We utilize information from the entire spectrum and create more measurements in the analysed frequency bin. Thus, for every block of data, a robust and accurate sample covariance matrix can be obtained. Experimental results show that, when a new source appears, the proposed method can perform BSS using a single block of observation data with minimal effort by utilizing the “meta-information”. The drawback of the proposed approach is that the derivations are based on a free-field assumption, e.g., in a non-reverberant environment. The performance can be adversely affected by reverberation. We conclude that it is now possible to update the separation parameters of a BSS system to account for the appearance of a new interfering signal with minimal estimation effort in a non-reverberant scenario. This is essential for real-time BSS applications to ensure good separation quality at all times.

In Chapter 6, we proposed a method that enables the long mixing filters to be described by shorter representations. This is beneficial to applications that involve system identification, such as BSS and cross-talk cancellation, as the computational cost can be reduced while dealing with shorter impulse responses. Due to the scaling ambiguity in the TF-domain BSS approaches, we are able to apply complex scaling factors to frequency-domain responses. We can make the frequency-domain responses smoother, so that in the time domain the responses become sparser. The method is non-iterative, hence it is computationally efficient. Simulation results confirm the effectiveness of the method and show that it can be implemented as a post-processing technique for BSS applications and as a pre-processing technique for cross-talk cancellation applications.

7.2 Future Research Directions

A number of future research directions form a natural continuation of the work described in this thesis. The first would be to implement the proposed

approaches in practical devices, such as hearing instruments, to perform BSS in real time. Users can concentrate on the talkers of interest. The user experience can be improved due to the low latency of the system.

The second is to apply the proposed approaches to other applications that involve system identification, such as beamforming, source localization, acoustic echo cancellation, system equalization, etc. The proposed methods focus on the system modelling, and they are not restricted to only BSS applications.

Another interesting research direction can be exploring the implementation of the system model described in Chapter 4 on the BSS methods that fit in neither Category 1 nor Category 2. For example, nonnegative matrix factorization (NMF) methods and machine learning approaches [61] that are usually employed for monaural source separation problems, e.g., a system contains only one observation. By considering the system model of Chapter 4, these approaches might be implemented with a low delay.

In recent years, virtual-reality (VR) and augmented-reality (AR) applications have become more and more popular. The immersive experience in these devices can be enhanced by including spatial audio [37]. BSS can be beneficial to these applications by decomposing mixtures into multiple independent sources and filtering unwanted signals. The desired sources are then composed to form a new meaningful mixture. Spatial cues can also be added to indicate the location of the desired sources to enhance the user experience.

Bibliography

- [1] ABRARD, F., AND DEVILLE, Y. Blind separation of dependent sources using the time-frequency ratio of mixtures approach. In *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on* (2003), vol. 2, IEEE, pp. 81–84.
- [2] ABRARD, F., AND DEVILLE, Y. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing* 85, 7 (2005), 1389–1403.
- [3] ADEL, H., SOUAD, M., ALAQEELI, A., AND HAMID, A. Beamforming techniques for multichannel audio signal separation. *arXiv preprint arXiv:1212.6080* (2012).
- [4] AGNEW, J., AND THORNTON, J. M. Just noticeable and objectionable group delays in digital hearing aids. *Journal of the American Academy of Audiology* 11, 6 (2000), 330–336.
- [5] ALLEN, J. B., AND BERKLEY, D. A. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65, 4 (1979), 943–950.
- [6] ALLEN, J. B., AND RABINER, L. R. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE* 65, 11 (1977), 1558–1564.

- [7] AMARI, S.-I. Natural gradient learning for over-and under-complete bases in ICA. *Neural Computation* 11, 8 (1999), 1875–1883.
- [8] AMARI, S.-I., DOUGLAS, S. C., CICHOCKI, A., AND YANG, H. H. Multichannel blind deconvolution and equalization using the natural gradient. In *Signal Processing Advances in Wireless Communications, First IEEE Signal Processing Workshop on* (1997), IEEE, pp. 101–104.
- [9] ANEMÜLLER, J., AND KOLLMEIER, B. Amplitude modulation decorrelation for convolutive blind source separation. In *Proc. ICA* (2000), pp. 215–220.
- [10] ARAKI, S., MUKAI, R., MAKINO, S., NISHIKAWA, T., AND SARUWATARI, H. The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Transactions on Speech and Audio Processing* 11, 2 (2003), 109–116.
- [11] ARBERET, S., GRIBONVAL, R., AND BIMBOT, F. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *Signal Processing, IEEE Transactions on* 58, 1 (2010), 121–133.
- [12] ASANO, F., IKEDA, S., OGAWA, M., ASOH, H., AND KITAWAKI, N. Combined approach of array processing and independent component analysis for blind separation of acoustic signals. *Speech and Audio Processing, IEEE Transactions on* 11, 3 (2003), 204–215.
- [13] AVARGEL, Y., AND COHEN, I. On multiplicative transfer function approximation in the short-time Fourier transform domain. *IEEE Signal Processing Letters* 14, 5 (2007), 337–340.
- [14] AVARGEL, Y., AND COHEN, I. System identification in the short-time Fourier transform domain with crossband filtering. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 4 (2007), 1305–1319.

- [15] BAUMANN, W., KOHLER, B.-U., KOLOSSA, D., AND ORGLMEISTER, R. Real time separation of convolutive mixtures. *TW Lee, TP Jung, S. Makeig and TJ Sejnowski eds* (2001), 65–69.
- [16] BELL, A. J., AND SEJNOWSKI, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation* 7, 6 (1995), 1129–1159.
- [17] BENAROYA, L., McDONAGH, L., BIMBOT, F., AND GRIBONVAL, R. Non negative sparse representation for wiener based source separation with a single sensor. In *IEEE international conference on acoustics, speech and signal processing (ICASSP 2003)* (2003), vol. 6, IEEE, pp. VI–613.
- [18] BENVENISTE, A., AND GOURSAT, M. Blind equalizers. *IEEE Transactions on communications* 32, 8 (1984), 871–883.
- [19] BETLEHEM, T., TEAL, P. D., AND HIOKA, Y. Efficient crosstalk canceler design with impulse response shortening filters. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (March 2012), pp. 393–396.
- [20] BINGHAM, E., AND HYVÄRINEN, A. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International journal of neural systems* 10, 01 (2000), 1–8.
- [21] BLIN, A., ARAKI, S., AND MAKINO, S. Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 88, 7 (2005), 1693–1700.
- [22] BLUMENSON, L. A derivation of n-dimensional spherical coordinates. *American Mathematical Monthly* (1960), 63–66.

- [23] BOOTHROYD, A. Room acoustics and speech perception. In *Seminars in Hearing* (2004), vol. 25, Thieme Medical Publishers Inc, pp. 155–166.
- [24] BÖTTCHER, A., AND SILBERMANN, B. *Introduction to large truncated Toeplitz matrices*. Springer Science & Business Media, 2012.
- [25] BOYD, S., AND VANDENBERGHE, L. *Convex optimization*. Cambridge university press, 2004.
- [26] BRUNET, J.-P., TAMAYO, P., GOLUB, T. R., AND MESIROV, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 101, 12 (2004), 4164–4169.
- [27] CAPON, J. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE* 57, 8 (1969), 1408–1418.
- [28] CARDOSO, J.-F., AND SOULOUMIAC, A. Blind beamforming for non-Gaussian signals. In *IEE Proceedings F-Radar and Signal Processing* (1993), vol. 140, IET, pp. 362–370.
- [29] CHAMPENEY, D. C. *A handbook of Fourier theorems*. Cambridge University Press, 1987.
- [30] CHARIS, G., AND SHOWME, N. Beamforming in wireless communication standards: A survey. *Indian Journal of Science and Technology* 10, 5 (2017).
- [31] CHEVREUIL, A., AND LOUBATON, P. Blind signal separation for digital communication data. In *Academic Press Library in Signal Processing*, vol. 2. Elsevier, 2014, pp. 135–186.
- [32] CHUA, J., AND KLEIJN, W. B. Non-iterative impulse response shortening method for system latency reduction. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (2017), IEEE, pp. 581–585.

- [33] CHUA, J., WANG, G., AND KLEIJN, W. B. Convolutional blind source separation with low latency. In *Acoustic Signal Enhancement (IWAENC), 2016 IEEE International Workshop on* (2016), IEEE, pp. 1–5.
- [34] COMON, P. Independent component analysis. *Higher-Order Statistics* (1992), 29–38.
- [35] COMON, P. Independent component analysis, a new concept? *Signal processing* 36, 3 (1994), 287–314.
- [36] COMON, P., AND JUTTEN, C. *Handbook of Blind Source Separation: Independent component analysis and applications*. Amsterdam: Elsevier, 2010.
- [37] CORREA, J. F. From sound art to virtual/augmented/mixed reality: the expansion of immersion and agency through 3D and interactive sound in digital media.
- [38] DE MOURA, N. N., SIMAS FILHO, E. F., AND DE SEIXAS, J. M. Narrow-band short-time frequency-domain blind signal separation of passive sonar signals. In *International Conference on Independent Component Analysis and Signal Separation* (2009), Springer, pp. 686–693.
- [39] DELCROIX, M., KINOSHITA, K., NAKATANI, T., ARAKI, S., OGAWA, A., HORI, T., WATANABE, S., FUJIMOTO, M., YOSHIOKA, T., OBA, T., ET AL. Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)* (2011), pp. 12–17.

- [40] DiBIASE, J. H. *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University Providence, RI, 2000.
- [41] DO, H., SILVERMAN, H. F., AND YU, Y. A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (2007), vol. 1, IEEE, pp. I–121.
- [42] DOCLO, S., AND MOONEN, M. Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement. In *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC '01)* (September 2001), Darmstadt, Germany.
- [43] ENZNER, G., SCHMID, D., AND HAEB-UMBACH, R. On acoustic channel identification in multi-microphone systems via adaptive blind signal enhancement techniques. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European* (2013), IEEE, pp. 1–5.
- [44] FAN, W., ZHANG, X., AND JIANG, B. A new passive sonar bearing estimation algorithm combined with blind source separation. In *Computational Science and Optimization (CSO), 2010 Third International Joint Conference on* (2010), vol. 1, IEEE, pp. 15–18.
- [45] FÉVOTTE, C., BERTIN, N., AND DURRIEU, J.-L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural computation* 21, 3 (2009), 793–830.
- [46] FISCHER, J., AND DOOLAN, C. Beamforming in a reverberant environment using numerical and experimental steering vector formulations. *Mechanical Systems and Signal Processing* 91 (2017), 10–22.

- [47] FITZGERALD, D., CRANITCH, M., AND COYLE, E. Non-negative tensor factorisation for sound source separation. *Irish Signals and Systems Conference* (2005).
- [48] FREDRIKSSON, H. On the short-time Fourier transform and Gabor frames generated by B-splines. Master's thesis, Linnaeus University, Sweden, 2012.
- [49] FROST, O. L. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE* 60, 8 (1972), 926–935.
- [50] GODARD, D. N. Self-recovering equalization and carrier tracking in two-dimensional data communication systems. *Communications, IEEE Transactions on* 28, 11 (1980), 1867–1875.
- [51] GOODE, B. Voice over internet protocol (VOIP). *Proceedings of the IEEE* 90, 9 (2002), 1495–1517.
- [52] GOROKHOV, A., AND LOUBATON, P. Subspace-based techniques for blind separation of convolutive mixtures with temporally correlated sources. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 44, 9 (1997), 813–820.
- [53] GRIFFITHS, L., AND JIM, C. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on antennas and propagation* 30, 1 (1982), 27–34.
- [54] HABETS, E. A. P. Room impulse response (RIR) generator, 2006.
- [55] HELEN, M., AND VIRTANEN, T. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Signal Processing Conference, 2005 13th European* (2005), IEEE, pp. 1–4.
- [56] HERBIG, R., AND CHALUPPER, J. Acceptable processing delay in digital hearing aids. *Hearing Review* 17, 1 (2010), 28–31.

- [57] HIDRI, A., MEDDEB, S., AND AMIRI, H. About multichannel speech signal extraction and separation techniques. *Journal of Signal and Information Processing* 3 (2012), 238–247.
- [58] HODGSON, M., AND NOSAL, E.-M. Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms. *The Journal of the Acoustical Society of America* 111, 2 (2002), 931–939.
- [59] HOYER, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research* 5, Nov (2004), 1457–1469.
- [60] HU, T.-Y., WANG, R.-Q., AND WHITE, R. Beamforming in seismic data processing. *Chinese Journal of Geophysics* 43, 1 (2000), 89–100.
- [61] HUANG, P.-S., KIM, M., HASEGAWA-JOHNSON, M., AND SMARAGDIS, P. Deep learning for monaural speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (2014), IEEE, pp. 1562–1566.
- [62] HUANG, Y., AND BENESTY, J. Adaptive blind channel identification: multi-channel least mean square and Newton algorithms. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* (2002), vol. 2, IEEE, pp. II–1637.
- [63] HUANG, Y., BENESTY, J., AND CHEN, J. Identification of acoustic MIMO systems: Challenges and opportunities. *Signal Processing* 86, 6 (2006), 1278–1295.
- [64] HUANG, Y., BENESTY, J., AND ELKO, G. W. Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on* (1999), vol. 2, IEEE, pp. 937–940.

- [65] HUANG, Y. A., BENESTY, J., AND CHEN, J. A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment. *Speech and Audio Processing, IEEE Transactions on* 13, 5 (2005), 882–895.
- [66] HUBER, P. J., ET AL. Robust estimation of a location parameter. *The annals of mathematical statistics* 35, 1 (1964), 73–101.
- [67] HYVÄRINEN, A. *Independent component analysis by minimization of mutual information*. Helsinki University of Technology, 1997.
- [68] HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on* 10, 3 (1999), 626–634.
- [69] HYVÄRINEN, A., KARHUNEN, J., AND OJA, E. *Independent component analysis*, vol. 46. New York: John Wiley & Sons, 2004.
- [70] HYVÄRINEN, A., AND OJA, E. A fast fixed-point algorithm for independent component analysis. *Neural computation* 9, 7 (1997), 1483–1492.
- [71] HYVÄRINEN, A., AND OJA, E. Independent component analysis: algorithms and applications. *Neural networks* 13, 4 (2000), 411–430.
- [72] ICART, S., AND GAUTIER, R. Blind separation of convolutive mixtures using second and fourth order moments. In *1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1996), IEEE, pp. 3018–3021.
- [73] IKRAM, M. Z., AND MORGAN, D. R. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* (2002), vol. 1, IEEE, pp. 881–884.

- [74] JAMES, C. J., AND HESSE, C. W. Independent component analysis for biomedical signals. *Physiological measurement* 26, 1 (2004), R15.
- [75] JIA, S., AND QIAN, Y. Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing* 47, 1 (2009), 161–173.
- [76] JOURJINE, A., RICKARD, S., AND YILMAZ, O. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Acoustics, Speech and Signal Processing (ICASSP), 2000 IEEE International Conference on* (2000), vol. 5, IEEE, pp. 2985–2988.
- [77] JUNG, T.-P., MAKEIG, S., HUMPHRIES, C., LEE, T.-W., MCKEOWN, M. J., IRAGUI, V., AND SEJNOWSKI, T. J. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37, 2 (2000), 163–178.
- [78] JUNGSMANN, J. O., MAZUR, R., KALLINGER, M., MEI, T., AND MERTINS, A. Combined acoustic MIMO channel crosstalk cancellation and room impulse response reshaping. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 6 (2012), 1829–1842.
- [79] JUTTEN, C. *Calcul neuromimétique et traitement du signal: analyse en composantes indépendantes*. PhD thesis, Grenoble INPG, 1987.
- [80] KAMATA, K., HU, X., AND KOBATAKE, H. A new approach to the permutation problem in frequency domain blind source separation. In *Independent Component Analysis and Blind Signal Separation*. Berlin Heidelberg: Springer, 2004, pp. 849–856.
- [81] KESSY, A., LEWIN, A., AND STRIMMER, K. Optimal whitening and decorrelation. *The American Statistician* (2018), 1–6.

- [82] KIM, H., AND PARK, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23, 12 (2007), 1495–1502.
- [83] KIM, T. Real-time independent vector analysis for convolutive blind source separation. *IEEE Transactions on Circuits and Systems I: Regular Papers* 57, 7 (2010), 1431–1438.
- [84] KIM, T., ELTOFT, T., AND LEE, T.-W. Independent vector analysis: An extension of ICA to multivariate components. In *Independent Component Analysis and Blind Signal Separation*. Berlin Heidelberg: Springer, 2006, pp. 165–172.
- [85] KLINGER, A. Approximate pseudoinverse solutions to ill-conditioned linear systems. *Journal of Optimization Theory and Applications* 2, 2 (1968), 117–124.
- [86] KNEIPFER, R. R. Sonar beamforming - An overview of its history and status. Tech. rep., NAVAL UNDERSEA WARFARE CENTER NEWPORT DIV NEW LONDON CT NEW LONDON DETACHMENT, 1992.
- [87] KRISHNAN, L., TEAL, P. D., AND BETLEHEM, T. A robust sparse approach to acoustic impulse response shaping. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2015), pp. 738–742.
- [88] KUNDUR, D., AND HATZINAKOS, D. Blind image deconvolution. *IEEE signal processing magazine* 13, 3 (1996), 43–64.
- [89] KURITA, S., SARUWATARI, H., KAJITA, S., TAKEDA, K., AND ITAKURA, F. Evaluation of blind signal separation method using directivity pattern under reverberant conditions. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on* (2000), vol. 5, IEEE, pp. 3140–3143.

- [90] LEE, D. D., AND SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788.
- [91] LEE, T.-W., BELL, A. J., AND LAMBERT, R. H. Blind separation of delayed and convolved sources. In *Advances in neural information processing systems* (1997), pp. 758–764.
- [92] LEE, T.-W., BELL, A. J., AND ORGLMEISTER, R. Blind source separation of real world signals. In *Neural Networks, 1997., International Conference on* (1997), vol. 4, IEEE, pp. 2129–2134.
- [93] LEE, Y., AND WU, W.-R. A robust adaptive generalized sidelobe canceller with decision feedback. *IEEE Transactions on Antennas and Propagation* 53, 11 (2005), 3822–3832.
- [94] LI, J., STOICA, P., AND WANG, Z. On robust Capon beamforming and diagonal loading. *IEEE transactions on signal processing* 51, 7 (2003), 1702–1715.
- [95] LI, S. Z., HOU, X. W., ZHANG, H. J., AND CHENG, Q. S. Learning spatially localized, parts-based representation. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (2001), vol. 1, IEEE, pp. I–I.
- [96] LI, X., GIRIN, L., AND HORAUD, R. Audio source separation based on convolutive transfer function and frequency-domain lasso optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017).
- [97] LI, Y., AMARI, S.-I., CICHOCKI, A., HO, D. W., AND XIE, S. Underdetermined blind source separation based on sparse representation. *Signal Processing, IEEE Transactions on* 54, 2 (2006), 423–437.

- [98] LIN, X., GAUBITCH, N. D., NAYLOR, P., ET AL. Two-stage blind identification of SIMO systems with common zeros. In *Signal Processing Conference, 2006 14th European* (2006), IEEE, pp. 1–5.
- [99] LINDEMANN, E. Dynamic intensity beamforming system for noise reduction in a binaural hearing aid, Apr. 23 1996. US Patent 5,511,128.
- [100] LITVA, J., AND LO, T. K. *Digital beamforming in wireless communications*. Artech House, Inc., 1996.
- [101] LIU, H., AND XU, G. Closed-form blind symbol estimation in digital communications. *IEEE Transactions on Signal Processing* 43, 11 (1995), 2714–2723.
- [102] LIU, H., AND XU, G. Multiuser blind channel estimation and spatial channel pre-equalization. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (1995), vol. 3, IEEE, pp. 1756–1759.
- [103] LIU, K.-H., AND DRAGOSSET, W. H. Blind-source separation of seismic signals based on information maximization. *Geophysics* 78, 4 (2013), V119–V130.
- [104] LU, J.-Y., ZOU, H., AND GREENLEAF, J. F. Biomedical ultrasound beam forming. *Ultrasound in Medicine and Biology* 20, 5 (1994), 403–428.
- [105] MA, N., AND GOH, J. T. Efficient method to determine diagonal loading value. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* (2003), vol. 5, IEEE, pp. V–341.
- [106] MAKINO, S., LEE, T.-W., AND SAWADA, H. *Blind speech separation*, vol. 615. Springer, 2007.

- [107] MAKINO, S., SAWADA, H., MUKAI, R., AND ARAKI, S. Blind source separation of convolutive mixtures of speech in frequency domain. *IE-ICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 88, 7 (2005), 1640–1655.
- [108] MANSOUR, A., JUTTEN, C., AND LOUBATON, P. Subspace method for blind separation of sources in convolutive mixture. In *European Signal Processing Conference, 1996. EUSIPCO 1996. 8th* (1996), IEEE, pp. 1–4.
- [109] MARONNA, R. A. Robust M-estimators of multivariate location and scatter. *The annals of statistics* (1976), 51–67.
- [110] MATSUOKA, K. Minimal distortion principle for blind source separation. In *SICE 2002. Proceedings of the 41st SICE Annual Conference* (2002), vol. 4, IEEE, pp. 2138–2143.
- [111] MAULER, D., AND MARTIN, R. A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement. In *Signal Processing Conference, 2007 15th European* (2007), IEEE, pp. 222–226.
- [112] MAZUR, R., AND MERTINS, A. A method for filter shaping in convolutive blind source separation. In *International Conference on Independent Component Analysis and Signal Separation* (Berlin, Heidelberg, 2009), Springer Berlin Heidelberg, pp. 282–289.
- [113] MAZUR, R., AND MERTINS, A. Using the scaling ambiguity for filter shortening in convolutive blind source separation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (April 2009), pp. 1709–1712.
- [114] MERTINS, A., MEI, T., AND KALLINGER, M. Room impulse response shortening/reshaping with infinity-and-norm optimization.

- IEEE Transactions on Audio, Speech, and Language Processing* 18, 2 (2010), 249–259.
- [115] MESIN, L., HOLOHAR, A., AND MERLETTI, R. Blind source separation: Application to biomedical signals. *Advanced Methods of Biomedical Signal Processing* (2011), 379–409.
- [116] MIRZAL, A. NMF versus ICA for blind source separation. *Advances in Data Analysis and Classification* (2014), 1–24.
- [117] MISKIN, J., AND MACKAY, D. J. Ensemble learning for blind image separation and deconvolution. In *Advances in independent component analysis*. Springer, 2000, pp. 123–141.
- [118] MITIANOUDIS, N., AND DAVIES, M. Permutation alignment for frequency domain ICA using subspace beamforming methods. In *Independent Component Analysis and Blind Signal Separation*. Springer, 2004, pp. 669–676.
- [119] MIYOSHI, M., AND KANEDA, Y. Inverse filtering of room acoustics. *IEEE Transactions on acoustics, speech, and signal processing* 36, 2 (1988), 145–152.
- [120] MUKAI, R., SAWADA, H., ARAKI, S., AND MAKINO, S. Real-time blind source separation for moving speakers using blockwise ICA and residual crosstalk subtraction. In *Proc. ICA* (2003), Citeseer, pp. 975–980.
- [121] MUKAI, R., SAWADA, H., ARAKI, S., AND MAKINO, S. Robust real-time blind source separation for moving speakers in a room. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* (2003), vol. 5, IEEE, pp. V–469.

- [122] MUKAI, R., SAWADA, H., ARAKI, S., AND MAKINO, S. Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 87, 8 (2004), 1941–1948.
- [123] MUKAI, R., SAWADA, H., ARAKI, S., AND MAKINO, S. Frequency-domain blind source separation of many speech signals using near-field and far-field models. *EURASIP Journal on Applied Signal Processing* 2006 (2006), 200–200.
- [124] MURATA, N., IKEDA, S., AND ZIEHE, A. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing* 41, 1 (2001), 1–24.
- [125] NAIK, G. R., AND WANG, W. *Blind Source Separation: Advances in Theory, Algorithms and Applications*. Springer, 2014.
- [126] NAQVI, S. M., ZHANG, Y., AND CHAMBERS, J. A. Multimodal blind source separation for moving sources. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (2009), IEEE, pp. 125–128.
- [127] NAVIDI, W. C. *Statistics for engineers and scientists*. McGraw-Hill Higher Education, 2008.
- [128] NIKUNEN, J., DIMENT, A., AND VIRTANEN, T. Separation of moving sound sources using multichannel NMF and acoustic tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 2 (2018), 281–295.
- [129] NISHIKAWA, T., SARUWATARI, H., AND SHIKANO, K. Comparison of time-domain ICA, frequency-domain ICA and multistage ICA for blind source separation. In *Signal Processing Conference, 2002 11th European* (2002), IEEE, pp. 1–4.

- [130] ONO, N. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on* (2011), IEEE, pp. 189–192.
- [131] ONO, N. Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific* (2012), IEEE, pp. 1–4.
- [132] ONO, N. Fast stereo independent vector analysis and its implementation on mobile phone. In *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on* (2012), VDE, pp. 1–4.
- [133] OPPENHEIM, A. V., AND SCHAFER, R. *Digital Signal Processing*. MIT video course. Prentice-Hall, 1975.
- [134] OZEROV, A., AND FÉVOTTE, C. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 3 (2010), 550–563.
- [135] PAATERO, P. Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems* 37, 1 (1997), 23–35.
- [136] PAATERO, P., AND TAPPER, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 2 (1994), 111–126.
- [137] PAJOVIC, M., PREISIG, J. C., AND BAGGEROER, A. B. Analysis of optimal diagonal loading for MPDR-based spatial power estimators in the snapshot deficient regime. *IEEE Journal of Oceanic Engineering*, 99 (2018), 1–15.

- [138] PARRA, L., AND SAJDA, P. Blind source separation via generalized eigenvalue decomposition. *The Journal of Machine Learning Research* 4 (2003), 1261–1269.
- [139] PARRA, L., AND SPENCE, C. Convolutional blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing* 8, 3 (2000), 320–327.
- [140] PARRY, R. M., AND ESSA, I. Estimating the spatial position of spectral components in audio. In *Independent Component Analysis and Blind Signal Separation*. Berlin Heidelberg: Springer, 2006, pp. 666–673.
- [141] PAUCA, V. P., PIPER, J., AND PLEMMONS, R. J. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications* 416, 1 (2006), 29–47.
- [142] PEDERSEN, M. S., LARSEN, J., KJEMS, U., AND PARRA, L. C. A survey of convolutional blind source separation methods. *Multichannel Speech Processing Handbook* (2007), 1065–1084.
- [143] PHAM, D.-T., SERVIERE, C., AND BOUMARAF, H. Blind separation of convolutional audio mixtures using nonstationarity. In *Proc. ICA* (2003), pp. 981–986.
- [144] PLATT, J. C., AND FAGGIN, F. Networks for the separation of sources that are superimposed and delayed. In *Advances in neural information processing systems* (1992), pp. 730–737.
- [145] PORTNOFF, M. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 3 (Jun 1976), 243–248.

- [146] PORTNOFF, M. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 1 (1980), 55–69.
- [147] PRASAD, R., SARUWATARI, H., AND SHIKANO, K. Problems in blind separation of convolutive speech mixtures by negentropy maximization. In *Acoustic Signal Enhancement (IWAENC), 2003 IEEE International Workshop on* (2003), IEEE, pp. 287–290.
- [148] REJU, V. G., KOH, S. N., AND SOON, I. Y. Underdetermined convolutive blind source separation via time-frequency masking. *IEEE Transactions on Audio, Speech & Language Processing* 18, 1 (2010), 101–116.
- [149] RIVET, B. Blind non-stationnary sources separation by sparsity in a linear instantaneous mixture. In *Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 314–321.
- [150] RIVET, B., DUARTE, L. T., AND JUTTEN, C. Blind extraction of intermittent sources. In *Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 402–409.
- [151] RIVET, B., AND JUTTEN, C. When silence is gold. In *Signal Processing Conference, 2011 19th European* (2011), IEEE, pp. 1070–1074.
- [152] ROBLEDO-ARNUNCIO, E., AND JUANG, B.-H. Issues in frequency domain blind source separation-a critical revisit. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on* (2005), vol. 5, IEEE, pp. v–281.
- [153] SATO, Y. A method of self-recovering equalization for multilevel amplitude-modulation systems. *Communications, IEEE Transactions on* 23, 6 (1975), 679–682.

- [154] SAWADA, H., ARAKI, S., AND MAKINO, S. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *Audio, Speech, and Language Processing, IEEE Transactions on* 19, 3 (2011), 516–527.
- [155] SAWADA, H., ARAKI, S., MUKAI, R., AND MAKINO, S. Blind extraction of dominant target sources using ICA and time-frequency masking. *Audio, Speech, and Language Processing, IEEE Transactions on* 14, 6 (2006), 2165–2173.
- [156] SAWADA, H., KAMEOKA, H., ARAKI, S., AND UEDA, N. Multi-channel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 5 (2013), 971–982.
- [157] SAWADA, H., MUKAI, R., ARAKI, S., AND MAKINO, S. A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing* 12, 5 (2004), 530–538.
- [158] SCHMIDT, R. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation* 34, 3 (1986), 276–280.
- [159] SELESNICK, I. W. Short-time Fourier transform and its inverse. *Signal* 10, 1 (2009), 2.
- [160] SHAHNAZ, F., BERRY, M. W., PAUCA, V. P., AND PLEMMONS, R. J. Document clustering using nonnegative matrix factorization. *Information Processing & Management* 42, 2 (2006), 373–386.
- [161] SHALVI, O., AND WEINSTEIN, E. New criteria for blind deconvolution of nonminimum phase systems (channels). *Information Theory, IEEE Transactions on* 36, 2 (1990), 312–321.

- [162] SHERRY, T. Real time blind source separation in reverberant environments. Master's thesis, Victoria University of Wellington, New Zealand, 2014.
- [163] SIMON, C., LOUBATON, P., VIGNAT, C., JUTTEN, C., AND D'URSO, G. Separation of a class of convolutive mixtures: a contrast function approach. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on* (1999), vol. 3, IEEE, pp. 1429–1432.
- [164] SMARAGDIS, P., AND BROWN, J. C. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on*. (2003), IEEE, pp. 177–180.
- [165] SMARAGDIS, P., ET AL. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on audio speech and language processing* 15, 1 (2007), 1.
- [166] SMARAGDIS, P., RAJ, B., AND SHASHANKA, M. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation* (2007), Springer, pp. 414–421.
- [167] SORENSEN, M., EEGHEM, F. V., AND LATHAUWER, L. D. Blind multichannel deconvolution and convolutive extensions of canonical polyadic and block term decompositions. *IEEE Transactions on Signal Processing* 65, 15 (Aug 2017), 4132–4145.
- [168] SUDHAKAR, P., AND GRIBONVAL, R. A sparsity-based method to solve permutation indeterminacy in frequency-domain convolutive blind source separation. In *Independent Component Analysis and Signal Separation*. Springer, 2009, pp. 338–345.

- [169] SUN, T.-Y., LIU, C.-C., TSAI, S.-J., AND HSIEH, S.-T. Blind source separation with dynamic source number using adaptive neural algorithm. *Expert systems with applications* 36, 5 (2009), 8855–8861.
- [170] SUN, X., AND DOUGLAS, S. C. Adaptive paraunitary filter banks for contrast-based multichannel blind deconvolution. In *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2001), IEEE, pp. 2753–2756.
- [171] SUN, Y., BABU, P., AND PALOMAR, D. P. Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing* 64, 14 (2016), 3576–3590.
- [172] SUNOHARA, M., AND HARUTA, C. Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 216–220.
- [173] TAKAHATA, A. K., NADALIN, E. Z., FERRARI, R., DUARTE, L. T., SUYAMA, R., LOPES, R. R., ROMANO, J. M., AND TYGEL, M. Un-supervised processing of geophysical signals: A review of some key aspects of blind deconvolution and blind source separation. *IEEE Signal Processing Magazine* 29, 4 (2012), 27–35.
- [174] TALMON, R., COHEN, I., AND GANNOT, S. Relative transfer function identification using convolutive transfer function approximation. *IEEE Transactions on audio, speech, and language processing* 17, 4 (2009), 546–555.
- [175] TASESKA, M., AND HABETS, E. A. Blind source separation of moving sources using sparsity-based source detection and tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 3 (2018), 657–670.

- [176] THIEMANN, J., AND VINCENT, E. An experimental comparison of source separation and beamforming techniques for microphone array signal enhancement. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on* (2013), IEEE, pp. 1–5.
- [177] THOMAS, J., DEVILLE, Y., AND HOSSEINI, S. Time-domain fast fixed-point algorithms for convolutive ICA. *IEEE Signal Processing Letters* 13, 4 (2006), 228–231.
- [178] TJOA, S. K., AND LIU, K. R. Multiplicative update rules for non-negative matrix factorization with co-occurrence constraints. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (2010), IEEE, pp. 449–452.
- [179] TONG, L., XU, G., AND KAILATH, T. Blind identification and equalization based on second-order statistics: A time domain approach. *Information Theory, IEEE Transactions on* 40, 2 (1994), 340–349.
- [180] TORKKOLA, K. Blind separation of convolved sources based on information maximization. In *Neural Networks for Signal Processing [1996] VI. Proceedings of the 1996 IEEE Signal Processing Society Workshop* (1996), IEEE, pp. 423–432.
- [181] TORKKOLA, K. Blind separation for audio signals—are we there yet? In *First International Workshop on Independent component analysis and blind source separation* (1999), pp. 239–244.
- [182] TREICHLER, J., AND AGEE, B. G. A new approach to multipath correction of constant modulus signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 31, 2 (1983), 459–472.
- [183] TYLER, D. E. A distribution-free M-estimator of multivariate scatter. *The Annals of Statistics* (1987), 234–251.

- [184] VAN DER VEEN, A. J., TALWAR, S., AND PAULRAJ, A. Blind estimation of multiple digital signals transmitted over fir channels. *IEEE Signal Processing Letters* 2, 5 (1995), 99–102.
- [185] VAN DER VEEN, A.-J., TALWAR, S., AND PAULRAJ, A. A subspace approach to blind space-time signal processing for wireless communication systems. *IEEE Transactions on Signal Processing* 45, 1 (1997), 173–190.
- [186] VAN EEGHEM, F., AND DE LATHAUWER, L. Second-order tensor-based convolutive ICA: Deconvolution versus tensorization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 2252–2256.
- [187] VAN VEEN, B. D., AND BUCKLEY, K. M. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine* 5, 2 (1988), 4–24.
- [188] VETTERLI, M., KOVAČEVIĆ, J., AND GOYAL, V. K. *Foundations of signal processing*. Cambridge University Press, 2014.
- [189] VINCENT, E., GRIBONVAL, R., AND FÉVOTTE, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 4 (2006), 1462–1469.
- [190] VINCENT, E., SAWADA, H., BOFILL, P., MAKINO, S., AND ROSCA, J. P. First stereo audio source separation evaluation campaign: data, algorithms and results. In *Independent Component Analysis and Signal Separation*. Berlin Heidelberg: Springer, 2007, pp. 552–559.
- [191] VIRTANEN, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing* 15, 3 (2007), 1066–1074.

- [192] VOROBYOV, S. A. Principles of minimum variance robust adaptive beamforming design. *Signal Processing* 93, 12 (2013), 3264–3277.
- [193] WOOD, N., AND COWAN, N. The cocktail party phenomenon revisited: how frequent are attention shifts to one’s name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21, 1 (1995), 255.
- [194] WU, H.-C., AND PRINCIPE, J. C. Simultaneous diagonalization in the frequency domain (SDIF) for source separation. In *Proc. ICA* (1999), vol. 99, Citeseer, pp. 245–250.
- [195] XU, W., LIU, X., AND GONG, Y. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM, pp. 267–273.
- [196] YAZAWA, S., HAMANAKA, M., AND UTSURO, T. Novel approach to separation of musical signal sources by NMF. In *Signal Processing (ICSP), 2014 12th International Conference on* (2014), IEEE, pp. 610–615.
- [197] YELLIN, D., AND WEINSTEIN, E. Multichannel signal separation: Methods and analysis. *IEEE Transactions on signal processing* 44, 1 (1996), 106–118.
- [198] YILMAZ, O., AND RICKARD, S. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing* 52, 7 (2004), 1830–1847.
- [199] ZHANG, T., AND WIESEL, A. Automatic diagonal loading for Tyler’s robust covariance estimator. In *Statistical Signal Processing Workshop (SSP), 2016 IEEE* (2016), IEEE, pp. 1–5.

- [200] ZHANG, W., KHONG, A. W. H., AND NAYLOR, P. A. Acoustic system equalization using channel shortening techniques for speech dereverberation. In *2009 17th European Signal Processing Conference* (Aug 2009), pp. 1427–1431.