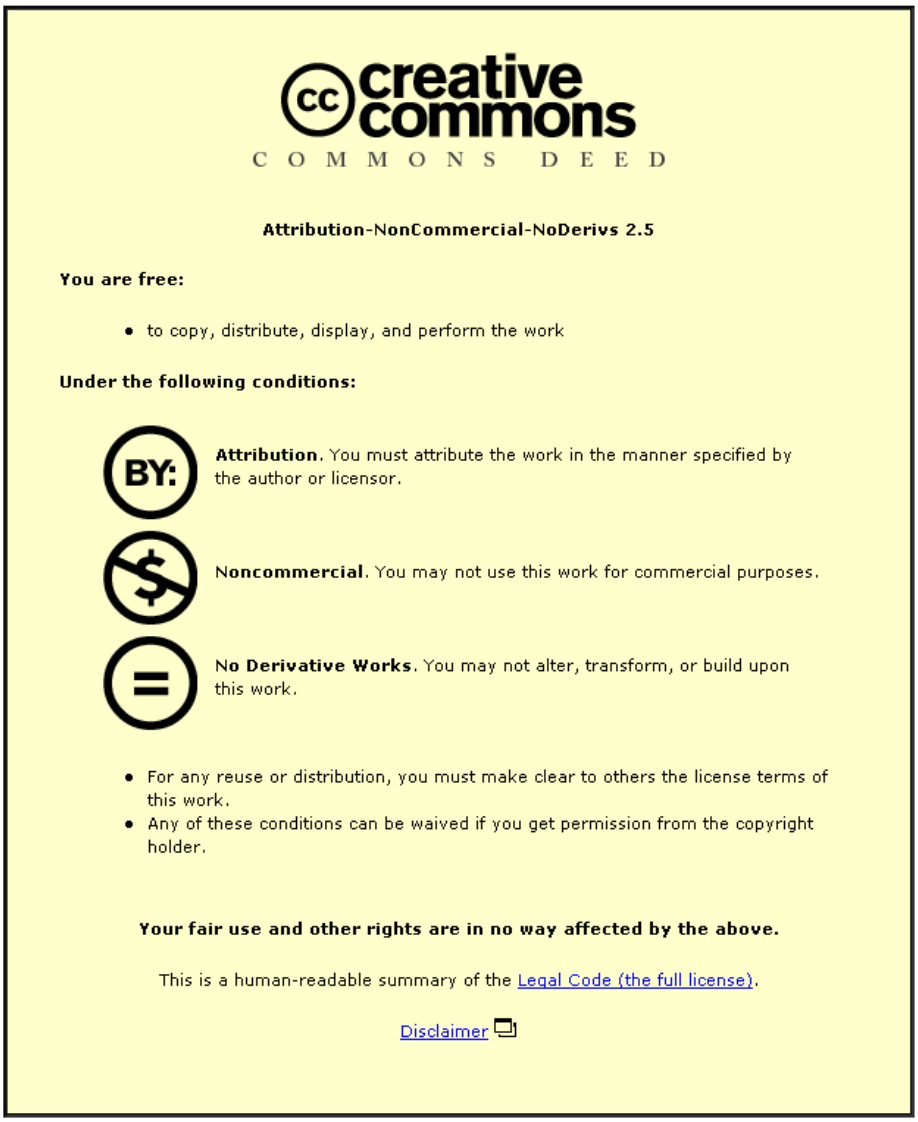
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Loughborough University Institutional Repository

This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (https://dspace.lboro.ac.uk/) under the following Creative Commons Licence conditions.



creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

Informed algorithms for sound source separation in enclosed reverberant environments

Muhammad Salman Khan

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy (PhD)

Advanced Signal Processing Group
School of Electronic, Electrical and Systems Engineering
Loughborough University
Leicestershire, England, UK, LE11 3TU

© by Muhammad Salman Khan, 2013

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (candidate)

This thesis is dedicated to my parents and family

Abstract

While humans can separate a sound of interest amidst a cacophony of contending sounds in an echoic environment, machine-based methods lag behind in solving this task. This thesis thus aims at improving performance of audio separation algorithms when they are "informed" i.e. have access to source location information. These locations are assumed to be known a priori in this work, for example by video processing.

Initially, a multi-microphone array based method combined with binary time-frequency masking is proposed. A robust least squares frequency invariant data independent beamformer designed with the location information is utilized to estimate the sources. To further enhance the estimated sources, binary time-frequency masking based post-processing is used but cepstral domain smoothing is required to mitigate musical noise.

To tackle the under-determined case and further improve separation performance at higher reverberation times, a two-microphone based method which is inspired by human auditory processing and generates soft time-frequency masks is described. In this approach interaural level difference, interaural phase difference and mixing vectors are probabilistically modeled in the time-frequency domain and the model parameters are learned through the expectation-maximization (EM) algorithm. A direction vector is estimated for each source, using the location information, which is used as the mean parameter of the mixing vector model. Soft time-frequency masks are used to reconstruct the sources. A spatial covariance model is then in-

tegrated into the probabilistic model framework that encodes the spatial characteristics of the enclosure and further improves the separation performance in challenging scenarios i.e. when sources are in close proximity and when the level of reverberation is high.

Finally, new dereverberation based pre-processing is proposed based on the cascade of three dereverberation stages where each enhances the two-microphone reverberant mixture. The dereverberation stages are based on amplitude spectral subtraction, where the late reverberation is estimated and suppressed. The combination of such dereverberation based pre-processing and use of soft mask separation yields the best separation performance. All methods are evaluated with real and synthetic mixtures formed for example from speech signals from the TIMIT database and measured room impulse responses.

Contents

Statement of Originality

The contributions of this thesis are primarily on the system integration aspects of sound source separation in enclosed reverberant environments by exploiting the source locations, potentially estimated through video processing. The contributions are supported by the following international journal and conference papers:

Book Chapter

1. **M. S. Khan**, S. M. Naqvi, and J. A. Chambers, "Model-Based Source Separation and Dereverberation Exploiting Location Information in Reverberant Environments", *Advances in Modern Blind Source Separation Techniques: Theory and Applications, (G. Naik and W. Wang, eds.) Springer, 2013.* (submitted)

Journal

2. **M. S. Khan**, S. M. Naqvi, A.-ur-Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms", *IEEE Transactions on Audio, Speech and Language Processing, vol. 21, issue 9, pp. 1900-1912, 2013.*

3. **M. S. Khan**, M. Yu, L. Wang, and J. A. Chambers, "An unsupervised acoustic fall detection system using blind source separation for sound interference suppression", *IET Signal Processing, 2013.* (provisionally accepted, revising)

4. S. M. Naqvi, W. Wang, **M. S. Khan**, M. Barnard, and J. A. Chambers, "Multimodal (audio-visual) source separation exploiting multi-speaker tracking, robust beamforming, and time-frequency masking, *IET Signal Processing, Special Issue on Multi-Sensor Signal Processing for Defence: Detection, Localisation and Classification, vol. 6, no. 5, pp. 466-477, 2012.*

Conference

5. **M. S. Khan**, S. M. Naqvi, and J. A. Chambers, "Two-stage audio-visual speech dereverberation and separation based on models of the interaural spatial cues and spatial covariance", *in Proc. IEEE DSP, Santorini, Greece, 2013.*

6. **M. S. Khan**, S. M. Naqvi, and J. A. Chambers, "Speech separation with dereverberation-based pre-processing incorporating visual cues", *in Proc. 2nd International workshop on machine listening in multi-source environments (CHIME), Vancouver, Canada, 2013.*

7. **M. S. Khan**, S. M. Naqvi, and J. A. Chambers, "A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation", *in Proc. IEEE ICASSP, Vancouver, Canada, 2013.*

8. **M. S. Khan**, A.-Rehman, S. M. Naqvi, and J. A. Chambers, "Convolutive speech separation by combining probabilistic models employing the interaural spatial cues and properties of the room assisted by vision", *in Proc. 9th IMA Mathematics in Signal Processing, Birmingham, UK, 2012.*

9. S. M. Naqvi, **M. S. Khan**, Q. Liu, W. Wang, and J. A. Chambers, "Multimodal blind source separation with a circular microphone ar-

ray and robust beamforming", *in Proc. EUSIPCO, Barcelona, Spain, 2011.*

Acknowledgements

All praise is due to Allah, the One, the most Beneficent and Merciful, the only Creator, the Controller, and the Sustainer. I wish I could be a truly thankful servant of His for the infinite bounties and blessings He has showered upon me. He indeed is the all-seeing, the all-knowing, the all-doer, and has no partner.

I express my deepest indebtedness to my parents, who raised me, cared for me, and did every effort to provide me with the best. To my father, a civil engineer, who has inspired me so much. His training has been instrumental in dealing with every stage of my life. To my mother, who provided the love and the spiritual support. I strongly believe her prayers were the main contributors in all my achievements.

I would like to show my deepest gratitude to my supervisor Prof. Jonathon Chambers. His commitment and support was unparalleled. He introduced me to such an interesting and exciting research area that kept me so indulged that I remembered to do my car's MOT six months after it had expired! In cricketing terms, being a fan and player of cricket, he could be referred to as an all-rounder! His technical expertise, supervisory experience, exceptional leadership and communication skills find no match. His feedback on my papers and reports was always prompt, which encouraged me a lot and allowed me to do more work in the time that was saved. He always returned emails swiftly, even in out-of-office hours, on weekends, and even while on a holiday! I thank him for the mentoring, guidance and support including the contribution towards my maintenance costs.

I would also like to thank all the staff and colleagues within the Advanced Signal Processing Group, for their guidance and assistance throughout the duration of my PhD, specifically at the beginning of my PhD studies. They provided me with a friendly environment to work in and were always happy to answer questions.

I want to specially thank Michael Mandel, Ngoc Duong, Philip Loizou, Deliang Wang, and Tariqullah Jan for sharing implementations of their works or making them publicly available online, and helping with any related issues.

I of course owe so much to my wife who provided a strong support throughout my PhD journey. She was patient and understanding, that allowed me to work hard and get the work done, many times on weekends and in holidays. She patiently remained away from her parents and family for me, and gave me all the support I needed. She deserves an honorary PhD award! I am also very grateful to her parents for their understanding, prayers, and support.

To my sons, all three under five, who had to wait for long hours before I would return home from the university. Upon seeing them all my tiredness would go. They would be so happy when I would take them out to play in the park after many days.

I would like to sincerely thank University of Engineering and Technology, Peshawar and the Higher Education Commission (HEC) of Pakistan for sponsoring my studies and providing me with this valuable opportunity which many do not have access to.

Finally, I thank my grandparents for their love and prayers, my brothers and sisters, and all other family members. Thanks to all my teachers and mentors in school, college, university and the institutions I attended, without whom this achievement would not have been possible. To all my friends and to everyone who were a means of support one way or another.

List of Acronyms

AIR	Aachen Impulse Response
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
BM	Binary Mask
CASA	Computational Auditory Scene Analysis
CSS	Convolutive Source Separation
DOA	Direction Of Arrival
DRR	Direct-to-Reverberation Ratio
DUET	Degenerate Unmixing Estimation Technique
EM	Expectation-Maximization
E-step	Expectation Step
FFT	Fast Fourier Transform
FT	Fourier Transform
HOS	Higher-Order Statistics
IBM	Ideal Binary Mask
ICA	Independent Component Analysis

ILD Interaural Level Difference

IPD Interaural Phase Difference

IID Interaural Intensity Difference

ISTFT Inverse Short-Time Fourier Transform

ITD Interaural Time Difference

IVA Independent Vector Analysis

KL Kullback-Leibler

MOS Mean Opinion Score

M-step Maximization Step

PEL Percentage of Energy Loss

PESQ Perceptual Evaluation of Speech Quality

PHAT Phase Transform

PI Performance Index

PNR Percentage of Noise Residue

RIR Room Impulse Response

RLSFIDI Robust Least Squares Frequency Invariant Data Independent

RT60 Reverberation Time for 60 dB decay

SAR Signal-to-Artifact Ratio

SDR Signal-to-Distortion Ratio

SINR Signal-to-Interference-Noise Ratio

SIMO Single-Input Multiple-Output

SIR Signal-to-Interference Ratio

SNR Signal-to-Noise Ratio

segSNR Segmental Signal-to-Noise Ratio

SOI Source Of Interest

SOS Second-Order Statistics

STFT Short-Time Fourier Transform

TDOA Time Difference Of Arrival

TF Time-Frequency

WDO Windowed-Disjoint Orthogonality

WNG White Noise Gain

List of Symbols

Some frequently used notations are as follows:

$\sum$	Summation
I	Number of sources
T	FFT length
$\|.\|$	Absolute value
ln	Natural logarithm
exp	Exponential
P	Length of mixing filter
Q	Length of unmixing filter
$\mathbf{H}$	Mixing matrix
$\mathbf{w}$	Estimated unmixing vector
$\mathbf{W}$	Estimated unmixing matrix
ω	Frequency index
t	Time index
$\mathbf{G}$	Overall system matrix
$\max(\cdot)$	Maximum value

$\min(\cdot)$	Minimum value
$\|\|.\|\|_2$	Euclidean norm
$(.)^T$	Transpose operator
$(.)^H$	Hermitian transpose operator
α	Interaural level difference (ILD)
ϕ	Interaural phase difference (IPD)
τ	Delay
c	Speed of sound in air at room temperature
ξ	Mean of the IPD model
σ^2	Variance of the IPD model
μ	Mean of the ILD model
η^2	Variance of the ILD model
$\mathbf{d}$	Direction vector
ς^2	Variance of the mixing vector model
θ	Elevation angle
ϕ	Azimuth angle
$\mathcal{L}$	Log likelihood function
$\mathbf{R}$	Covariance matrix
v	Scalar variance

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation

Almost everyday we encounter numerous instances where we need to focus on one sound of interest in the presence of many distracting sounds. It could be

- the parents' call to the children among other sound sources such as a television, a pet, or a vacuum cleaner within a home, or
- it may be a meeting room or an office setting where multiple speakers are simultaneously active and there is a need to follow one speaker, or
- it could be listening to a certain talker while multiple talkers are also active along with other background noise, as in a cocktail party situation [5], illustrated in Fig. 1.1.

Humans with normal hearing abilities, if required to undertake the aforementioned tasks will perform reasonably well. This remarkable performance of the human hearing system in conducting such complicated tasks is due to the complex auditory processing that is yet to be fully understood. Humans exploit multiple cues or features and there are numerous processes and complex mechanisms that make the difficult task of isolating a single

Figure 1.1. The Cocktail Party, 1965. Alex Katz.

sound among other competing sounds in realistic reverberant environments possible.

As technology progresses, more and more research is being done for the development of advanced machines that could benefit mankind in one way or another. Among many others, one need for these machines is to acquire human-like hearing capabilities (machine audition), or specifically, separate sounds from their reverberant mixtures as this would enable multiple application areas. To name a few, consider

- the performance of automatic speech recognition (ASR) systems (in smart phones, and computers) in realistic environments, with competing sources, reverberation and background noise. The performance degradation of such ASR systems could be considerably reduced by incorporating a pre-processing stage for reverberant speech separation.

- People with hearing difficulties require more sophisticated devices, such as, hearing aids or cochlear implants to better deal with everyday challenging acoustic scenarios. This will help tens of millions of people around the world and they can also carry out tasks that other humans with normal-hearing do.

- Within a meeting or teleconference room with typically multiple speakers and high levels of reverberation, robust source separation systems are required to enable convenient hands-free operation, and automatic speech transcription.

- In robotics applications, for instance, the robot needs to understand the directives in realistic environments in order to fulfil different tasks [6].

- In surveillance or forensic applications, where either there are recordings with mixtures of sounds or it is a real-time data feed; the source of interest could be extracted from the acoustic mixture.

The above-mentioned examples are just a few among many more where a sound of interest needs to be separated or extracted from a reverberant mixture of multiple sounds. Hence, there is sufficient motivation to develop efficient algorithms for machine-based reverberant sound source separation.

The current source separation algorithms can solve limited (with constraints on source statistics, the number of sources and microphones, or the amount of reverberation) versions of the source separation problem. Some methods, i.e. beamforming, typically require a large number of observations (microphones) to enhance a source coming from a certain direction and reject interferers from other directions [7]. For an improved performance, these methods are only effective when the number of sources is less than the number of microphones i.e. the over-determined case. Their performance is also limited at higher levels of reverberation, and thus, are not practically

very useful [8]. In blind source separation (BSS) using independent component analysis (ICA), an unmixing matrix is estimated assuming the mixed sources to be statistically independent. However, only determined or over-determined cases could be solved [9]. Computational auditory scene analysis (CASA) based methods follow a different approach in that they are inspired by the human auditory processing. They aim to model the fundamental cues that humans utilize in performing the separation task and typically utilize one or two microphones (i.e. are monaural or binaural). They generally exploit the time-frequency representation of observations and aim to estimate time-frequency masks to segregate individual sources from the mixture [10]. Assuming that the sources do not overlap in the time-frequency domain, these techniques are capable of solving the under-determined case i.e. more sources than microphones. Monaural cues i.e. pitch, onset/offset, and binaural cues i.e. interaural level and phase differences are typically used to identify the time-frequency points belonging to a certain source, and generate either hard (binary) or soft (probabilistic) time-frequency masks [11]. The masks are applied to the mixture to reconstruct the sources.

The performance of current source separation systems in realistic reverberant conditions is very limited. Reverberation distorts the cues, such as, the interaural level and phase difference, which are typically exploited by the separation systems. The assumptions on which the different techniques are based are also weakened due to reverberation. For instance, sparsity, which is usually exploited in time-frequency CASA-based methods, which assumes that signals are sparse in the time-frequency domain. Reverberation smears and increases energy across time. As such the signal becomes less sparse in the time-frequency domain, thus, causing degradation of the performance of separation algorithms. Separation performance further deteriorates when the number of sources in the mixtures increases.

1.2 Aims and Objectives

In many applications, information about the locations of the sound sources may be known *a priori*, or it may have been estimated through independent video processing. Can the source locations be used to advantage? Can this "informed" approach better tackle reverberation and the case of multiple speakers? To answer these and other similar questions, this thesis aims to develop multiple signal processing techniques for informed source separation in enclosed reverberant environments. The location information in this work however is assumed to be derived from video processing but this is not the subject of this thesis, further details can be found in [12, 13]. Such location information is used in all the contribution chapters, i.e. Chapters 3-6, whereas the room spatial characteristics are employed in Chapter 5. In the evaluation studies later in Chapter 3 of the thesis, the effect of estimation errors in such location information is also studied. Complexity issues and real-time implementation are outside of the scope of this thesis.
The aims of this thesis are summarized as follows:

Exploit multi-microphone array based method combined with binary time-frequency masking to segregate sources in reverberant environments

A multi-microphone beamforming method with binary time-frequency based post-processing is studied. The beamformer, utilizing the known source locations, provides an estimate of the speech sources. The source estimates are further enhanced by exploiting binary time-frequency masking. The aim of binary masking is to suppress any energy from the interfering source that has remained in the estimate of the target source obtained by the beamformer. Since the binary masks tend to generate unwanted musical noise, cepstral processing is also incorporated.

Study a two-microphone model-based method that generates soft time-frequency masks for under-determined reverberant source separation

To further the separation performance when the level of reverberation is high, and to be able to solve the under-determined problem, a two-microphone model-based approach is pursued. Inspired by the human auditory processing, the combined probabilistic models of the interaural level and phase differences and the mixing vectors are used. Since the source locations are assumed to be known, they are utilized within the modeling. Parameters of the models are estimated using the expectation-maximization algorithm. Soft (probabilistic) masks are obtained from the posterior probabilities to separate the sources from their reverberant mixtures.

Investigate modeling of the properties of the enclosure using a spatial covariance model

To utilize additional spatial properties of the enclosure, such as the reverberation time, and the wall reflective properties, a spatial covariance model is studied. The spatial covariance model is used in conjunction with the aforementioned models and is shown to improve the separation.

Explore a pre-processing stage and a novel cascade structure for binaural dereverberation based on amplitude spectral subtraction

To tackle high levels of reverberation, a dereverberation based pre-processing is studied. It is based on amplitude spectral subtraction. The pre-processing is evaluated both as a single stage and also as a cascade structure.

The objectives of this study include

- Developing efficient algorithms that are able to separate multiple sounds from their reverberant mixtures by exploiting the source locations.
- Publishing the work in leading journals and conferences in the area.

1.3 Organization of this thesis

Chapter 2 gives background for the related topics which will be studied in the later part of the thesis. It begins by describing sound production and its propagation within rooms. The room impulse response, reverberation time and other important related parameters are then introduced. Time-frequency representation of signals is studied before reviewing the different approaches to the source separation problem, including blind source separation, beamforming, and computational auditory scene analysis. The different performance evaluation measures are also discussed.

Chapter 3 describes a multi-microphone array based approach combined with binary time-frequency masking. Exploiting the knowledge of the locations of the sources, a robust least squares frequency invariant data independent beamformer is designed. A binary time-frequency masking based post-processing is then introduced. The estimated sources by the robust beamformer are further refined using the binary masks. To smooth the binary masks, since they tend to produce musical noise, cepstral based smoothing is applied.

Chapter 4 illustrates a two-microphone based algorithm inspired by the human auditory processing. It presents the probabilistic models of the interaural level and phase difference and mixing vectors. The models utilize the information of the locations of the sources. The models are combined and their parameters are estimated using the expectation-maximization algorithm. Experimental evaluation then follows which are conducted in varying scenarios.

Chapter 5 studies the spatial covariance model. The spatial covariance model exploits the spatial characteristics of the enclosure such as its reverberation time and wall reflection properties. The spatial covariance model is combined with the models explained in Chapter 4, with the aim to further

the separation performance in highly reverberant scenarios.

Chapter 6 investigates pre-processing based on dereverberation. Single-microphone spectral subtraction based dereverberation methods are studied first, and then extended to the binaural context. The two-microphone based dereverberation is utilized as a pre-processing stages before source separation. To further suppress that late reverberation, a new cascade structure is then studied. The cascade structure is also used as a pre-processor. A variety of experiments are performed in the dereverberation-only, and joint dereverberation and source separation processing contexts.

Chapter 7 summarizes the findings and the conclusions and discusses directions for future work.

Chapter 2

BACKGROUND AND LITERATURE REVIEW OF SOUND SOURCE SEPARATION IN REVERBERANT ENVIRONMENTS

2.1 Introduction

This chapter provides some background and a brief insight into the relevant topics discussed in the later chapters. Although the areas could be discussed in more detail, the focus here was to provide coverage which is sufficient for a reasonable overall understanding of the area, and not shallow enough to skip essential concepts. Further detail can be obtained through the extensive list of references provided.

The chapter begins by introducing sound production and propagation in enclosed environments. After discussing the room impulse response, im-

portant parameters such as the reverberation time, direct-to-reverberation ratio, and the critical distance are defined. A section explaining the time-frequency representation of signals follows next. The different approaches to the source separation problem are then briefly reviewed, including blind source separation, beamforming, and computational auditory scene analysis. Since the computational auditory scene analysis based methods are more relevant to this thesis, they are reviewed in relatively more detail. Different performance evaluation measures are then explained which are used to test the performance of the algorithms developed, as detailed in the forthcoming chapters.

2.2 Room Acoustics

Sound is produced by the physical vibrations of the sound source and propagates as a pressure wave through air (or another medium). Sound waves emitted in an enclosed environment are subject to multiple reflections and diffractions with wall surfaces and objects within the enclosure, before being received by a sensor (ear or microphone), as depicted in Fig. 2.1. Reflections of the source signals are sensitive to characteristics of the geometry of the environment, and the materials and objects within it. Thus, the received sound signal will be a mixture of the delayed and attenuated versions of the original source signal (along with the direct path signal). The propagation of sound and the reflections for a certain source-receiver position i.e. the room acoustic properties, can be fully described by the **room impulse response (RIR)**.

The RIR is composed of three main parts, namely, the direct-path, early reflections and late reverberation, illustrated in Fig. 2.2. The direct-path of

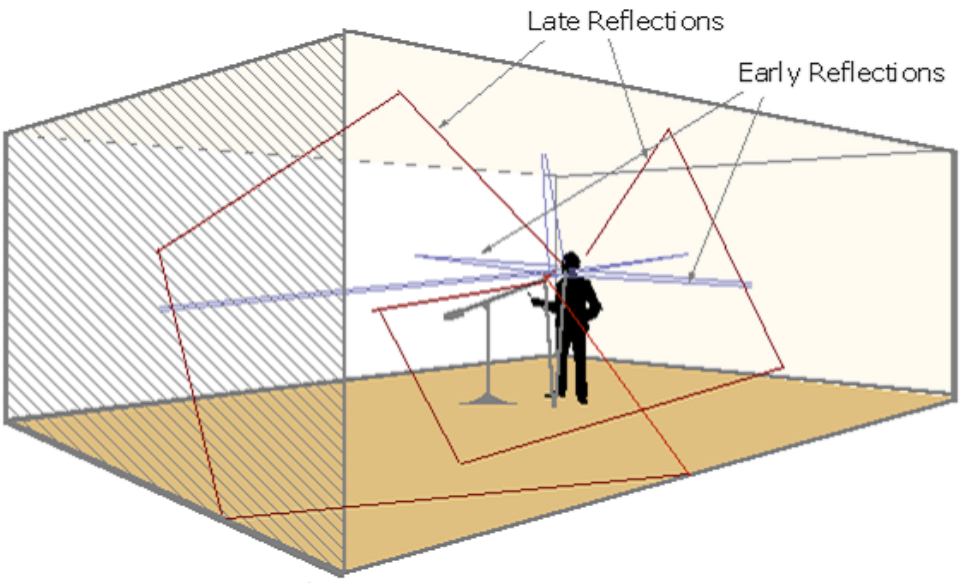


Figure 2.1. Multi-path reflections of sound waves in a room environment, from, SAE Institute (www.sae.edu).

the RIR in Fig. 2.2 is shown in blue, the early reflections in green (described as all energy between 10-50 ms here), and late reverberation in red. The direct-path signal is the sound received directly from the source without any reflections, and travels the shortest distance. Since the direct-path sound propagates from the original direction of the source, it has accurate information of the location of the source [14]. Early reflections arrive after the direct-path and there is evidence that they also improve intelligibility. Late reverberation starts after the early reflections and typically begins in the range of 50-100 milliseconds [15]. There is evidence of a perceptual mechanism in humans, termed as the precedence effect that aids in localizing sounds within reverberant environments [16].

Reverberation time (RT60) is an important parameter in room acous-

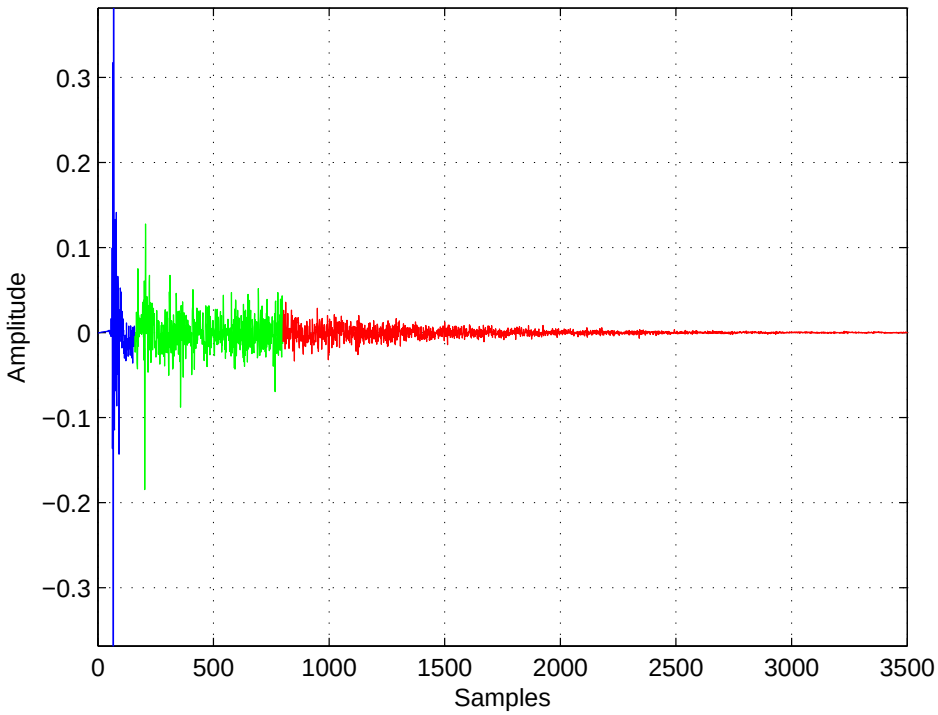


Figure 2.2. An illustration of an RIR with the direct-path, early reflections and late reverberation signals in blue, green and red respectively. Samples beyond 50 ms (800 samples at a sampling rate of 16000 Hz) are considered as late reverberation.

tics. It is the time taken by the sound signal power to decrease by 60 dB from the time when the sound source is switched off [17]. Studies by Sabine [15] indicate that the RT60 is directly proportional to the volume of the room and inversely proportional to the amount of absorption. If the volume of the room is denoted by "Vol", and α_{Sabine} and A denote the absorption coefficient and total absorption area respectively, the RT60, in seconds, can be estimated as [17]

$$RT60^{Sabine} = \frac{24\ln(10)}{c}\frac{Vol}{\alpha_{Sabine}A} \tag{2.2.1}$$

where c is the speed of sound in air. An alternative equation to estimate the

RT60 (in seconds) is also given by Eyring [17] as

$$RT60^{Eyring} = \frac{24\ln(10)}{c}\frac{Vol}{\ln(1-\alpha_{Eyring})A} \tag{2.2.2}$$

where α_{Eyring} is the Eyring sound absorption coefficient. Theoretically, from both the equations, it could be observed that the RT60 is independent of the distance of the sound source from the receiver [18]. The RIR is also characterized by another important parameter that compares the energies of its different components, called the **direct-to-reverberation energy ratio (DRR)**. It is the ratio of the energy of the direct-path signal (and usually some early reflections) to the remaining reverberant part [15]. When a sound source is at a position from the receiver that the direct-path energy is equal to the reverberant part energy, it is said to be at the critical distance [17].

The performance of current source separation systems at medium or higher RT60s ($>$ 300 ms) is limited. The late reflections within a room arrive with perceptible delay at the receiver and distort the information contained in the sound [19]. Even state-of-the-art source separation methods [9, 20–22] fail to overcome this problem. New techniques are thus required that could mitigate the effects of reverberation and thereby improve the separation performance.

2.3 Time-frequency Representation

Time-frequency representation is a very useful way to analyze (and process) speech signals that provides a representation of the signal over both time and frequency. The short-time Fourier transform (STFT) is typically used to transform the signal into the time-frequency domain.

Fig. 2.3 shows the speech waveform of an example utterance taken from

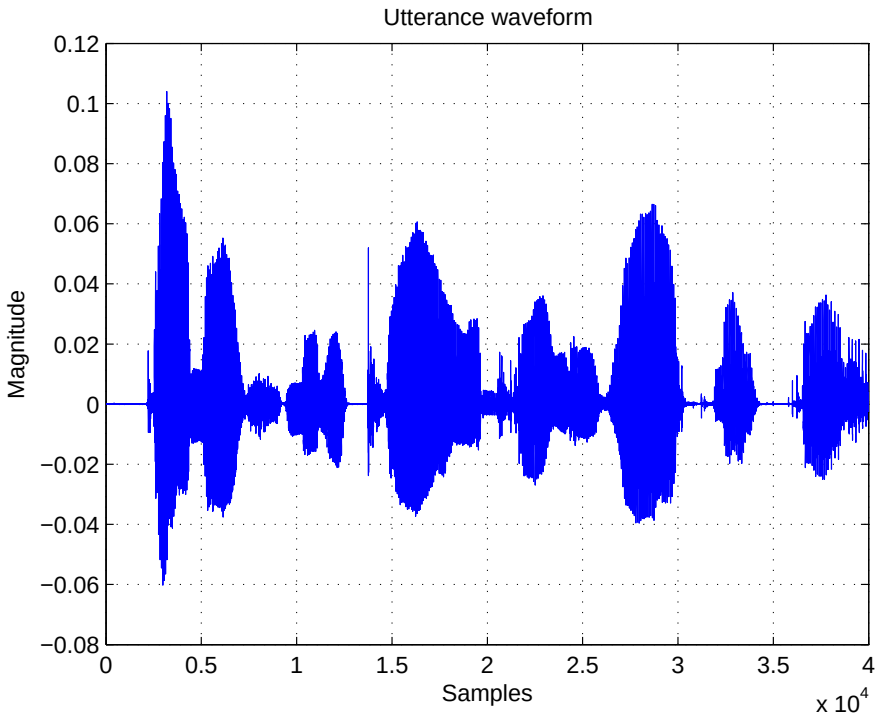


Figure 2.3. The waveform of a 2.5 second long example utterance "Don't ask me to carry an oily rag like that", sampled at 16 kHz.

the TIMIT database [23] where a female speaker says, "Don't ask me to carry an oily rag like that". The spectrogram, the magnitude squared of the STFT coefficients, of the same utterance is provided in Fig. 2.4, with time on the horizontal axis and frequency on the vertical axis. The analysis window in this example was 32 ms (512-point at sampling frequency of 16000 Hz). This means that the whole utterance was divided into chunks of size 512 each, and a 512-point fast Fourier transform (FFT) was taken. Each FFT thus represents the spectral activity over the 32 ms duration of the signal, giving us the variation of the spectrum of the signal over time. It can be observed from the spectrogram that most of the time-frequency points contain insignificant energy, indicating that this signal representation is sparse [24].

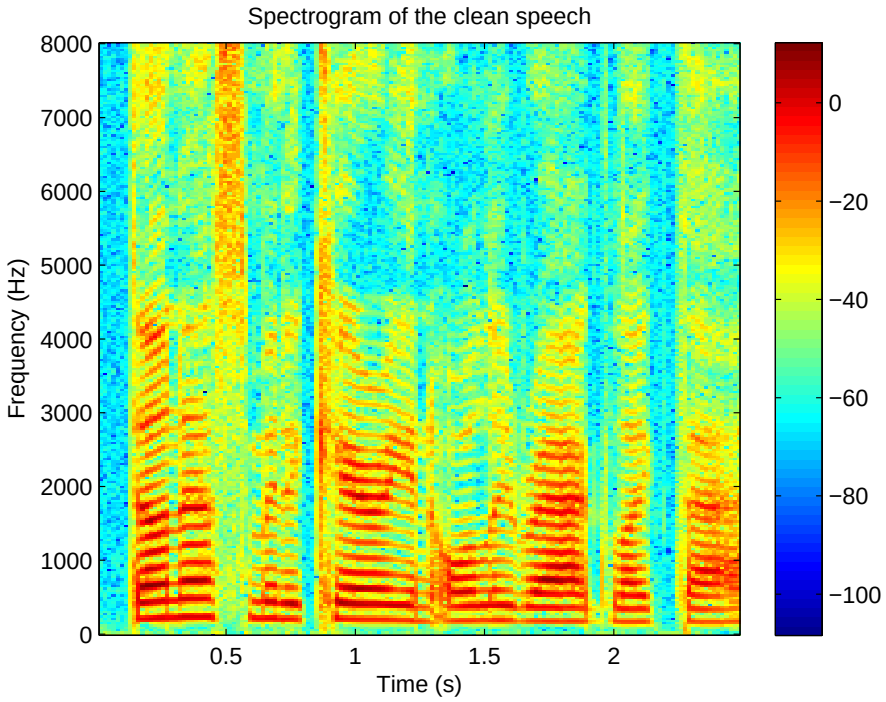


Figure 2.4. Spectrogram of the utterance "Don't ask me to carry an oily rag like that" with an analysis window of 32 ms.

The speech signal is convolved with a room impulse response and the spectrogram of the reverberant speech is shown in Fig. 2.5. RT60 of the room was 320 ms. It can be seen that the spectrogram is considerably blurred and time-frequency points with no or less energy are now filled by reverberation energy.

It is the late reverberation that causes temporal smearing of the signal and significantly degrades the performance of many signal processing applications [17,25]. There is evidence of certain perceptual mechanisms that help humans to adapt to different reverberant conditions [19]. Although humans with normal hearing do well in tackling the reverberation challenge [19], reverberation remains a challenge to machine-based processing. Different

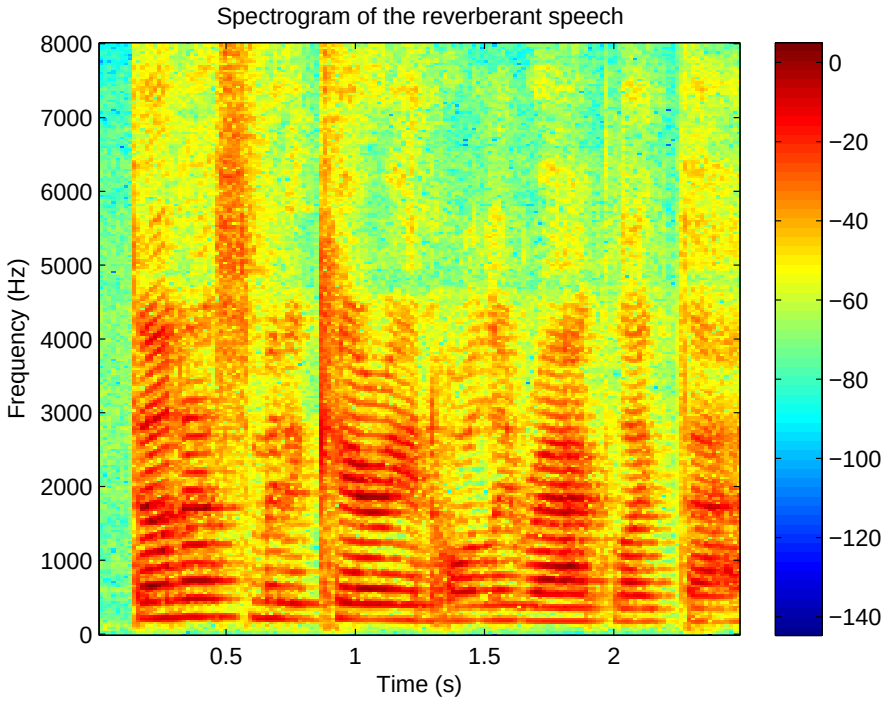


Figure 2.5. Spectrogram of the reverberant speech. The utterance in Fig. 2.4 was convolved with the room impulse response with an RT60 of 320 ms.

dereverberation methods have been proposed to mitigate the effects of reverberation [17]. Sound source separation systems also tend to compensate for the distortions caused due to reverberation, but generally still provide poor performance in the presence of reverberation equivalent to realistic levels, such as when RT60 $>$ 300 ms. This motivates the development of source separation algorithms that are relatively more robust to reverberation. Some of the different approaches to source separation are discussed next.

2.4 Blind Source Separation

Blind source separation (BSS) algorithms attempt to separate the source signals without the prior knowledge of sources or the mixing process. Depending on how the signals are mixed, algorithms can be classified as instan-

taneous, anechoic and reverberant (or convolutive) [26]. In the instantaneous mixing case each source signal appears within all the mixture channels at the same time with differing intensity. The anechoic mixing differs from the instantaneous case in that each source signal reaches the microphone with a delay [27]. The anechoic mixing model is sometimes referred to as an instantaneous mixing model with delays [26]. Mixing is reverberant (or echoic) when there are multiple reflective paths between each source and each microphone. The source separation task is challenging when source signals arrive at microphones from multiple directions and with different delays.

Time-domain convolutive BSS is computationally demanding because of the convolution calculation associated with the length of the room impulse response. Time-domain methods generally also have low convergence speeds, which motivates the transformation to the frequency-domain. Since convolution in the time-domain corresponds to multiplication in the frequency domain, the separation problem is simplified and instantaneous mixtures are obtained at each frequency bin. However, the main downside to the frequency-domain approach is the permutation problem (arbitrary order of sources). Most instantaneous BSS algorithms yield source estimates with scaling ambiguities and arbitrary order of sources. Applying such algorithms independently to each frequency bin and combining them can potentially lead to unintelligible and incorrect source estimates. The arbitrary scaling which occurs at each frequency bin is usually overcome by restricting the demixing matrix or the source estimates to be normalized [26]. On the other hand, the arbitrary order of sources in each frequency bin can lead to a total loss of the source separation achieved in the frequency domain when combined incorrectly.

The ratio of the number of sources to the number of microphones also influences the complexity of the separation process. A mixture is termed as determined when the number of microphones is equal to the number of

sources; over-determined when the number of microphones is larger than the number of sources; and under-determined (or over-complete) when the number of microphones is smaller than the number of sources [27]. Source separation is generally more difficult in the under-determined case.

In convolutive source separation (CSS), for I audio sources recorded by N microphones, the noise-free convolutive audio mixtures obtained can be described mathematically as

$$x_m(t_s) = \sum_{j=1}^{I} \sum_{p=1}^{P} h_{mj}(p) s_j(t_s - p + 1) \tag{2.4.1}$$

where s_j is the source signal from a source $j = 1, ..., I$, x_m is the received signal by microphone $m = 1, ..., N$, and $h_{mj}(p)$, $p = 1, \ldots, P$, is the p-th tap coefficient of the impulse response from source j to microphone m and t_s is the discrete time index.

In time-domain CSS, the sources are estimated using a set of unmixing filters such that

$$y_j(t_s) = \sum_{m=1}^{N} \sum_{q=1}^{Q} w_{jm}(q) x_m(t_s - q + 1) \tag{2.4.2}$$

where $w_{jm}(q)$, $q = 1, \ldots, Q$, is the q-th tap weight from microphone m to source j.

The CSS problem in the time-domain can be converted to multiple complex-valued instantaneous problems in the frequency-domain by using a T-point windowed short-time Fourier transformation (STFT), provided $T >> P$. The time-domain signals $x_m(t_s)$, are converted into time-frequency domain signals $x_m(\omega, t)$, where ω and t are respectively, frequency and time frame indices. The N observed mixed signals can be described as a noise-free vector in the time-frequency domain as

$$\mathbf{x}(\omega, t) = \mathbf{H}(\omega)\mathbf{s}(\omega, t) \tag{2.4.3}$$

where $\mathbf{x}(\omega, t)$ is an $N\times1$ observation column vector for frequency bin ω, $\mathbf{H}(\omega)$ is an $N\times I$ mixing matrix, $\mathbf{s}(\omega, t)$ is an $I\times1$ speech sources vector, and the source separation can be described as

$$\mathbf{y}(\omega, t) = \mathbf{W}(\omega)\mathbf{x}(\omega, t) \tag{2.4.4}$$

where $\mathbf{W}(\omega)$ is $I\times N$ separation matrix. By applying an inverse STFT (ISTFT), $\mathbf{y}(\omega, t)$ can be converted back to the time-domain as

$$y(t_s) = ISTFT(\mathbf{y}(\omega, t)) \tag{2.4.5}$$

BSS methods could broadly be classified as based on second-order statistics (SOS) or higher-order statistics (HOS). In SOS-based separation algorithms the sources are separated on the basis of decorrelation rather than independence and assume that the sources are statistically non-stationary or have a minimum phase mixing system [28].

Utilizing SOS, Parra and Spence [21] exploited non-stationarity of speech and proposed a solution to the source permutation problem. Separation was performed in the frequency domain. They used a multiple decorrelation approach and least-squares optimization to estimate the mixing/unmixing matrix as well as to estimate the signal and noise powers. They proposed to impose a smoothness constraint on the unmixing filters that forces the frequency bins to align. It is achieved by constraining the filter length in the time-domain to be much less than the frame size of the Fourier transform [28]. Many researchers have focussed on tackling the source permutation problem [29–32].

Another statistical technique that uses HOS is independent component

analysis (ICA). In the ICA model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also termed as sources, can be estimated by ICA. Typical assumptions of ICA can be summarized as follows: sources are assumed to be statistically independent of each other; all but one of the sources must have *non-Gaussian* distribution; the mixing matrix is usually assumed to be square and invertible (and the number of sources is equal to the number of mixtures, a determined problem) [28]. ICA generally suffers from permutation, scaling and data length problems.

Kim et al. in [33] proposed independent vector analysis (IVA), which preserves the higher-order dependencies and structures of signals across different frequencies to overcome the permutation problem in ICA. IVA exploits a dependency model which captures inter-frequency dependencies. The inter-frequency dependencies depend on a modified model for the source signal prior. The IVA method defines each source prior as a multivariate super-Gaussian distribution. Thus, it can potentially preserve the higher-order dependencies and structures of frequency components. Moreover, the permutation problem can be potentially avoided leading to an improved separation performance [33].

Taking the effects of reverberation and longer room impulse responses into perspective, Araki et al. in [22] studied the poor performance of frequency-domain BSS at higher reverberations. They reported that it was not very useful to be constrained by the condition where the Fourier transform frame size is greater than the filter length of the room impulse response. They also showed that both short and long frames fail: for a longer frame size, the number of samples in each frequency is small, therefore, the zero-mean and independence assumptions collapse and correct estimation of statistics

become challenging. For the case of a short frame, failure results because the frame size will not cover the reverberation. For instance, an RT60 of 500 ms correspond to the impulse response filter length of 4000, with a sampling rate of 8 kHz. So, the Fourier transform (FT) frame sizes of 1024 or 2048 are short and do not cover the entire reverberation profile. Whereas if the frame size increased to 4096, there will be insufficient samples at each frequency to apply a learning algorithm and the independence assumptions will collapse giving a poor separation performance. The authors concluded that there existed an optimum frame size that was determined by the trade-off between covering the entire reverberation and maintaining the independence assumption.

2.5 Beamforming

Beamforming techniques tackle the source separation problem from a spatial viewpoint. A beamformer, or a spatial filter, is a processor whose objective is to estimate the signal arriving from a desired direction in the presence of noise and interfering signals [34]. Fig. 2.6 illustrates a beamformer's beam pattern, where a source of interest is accepted by forming the main lobe towards it, while interferers are nulled from other directions.

In a delay-and-sum beamformer, with microphones arranged in a linear array, a sound source of interest from the far field arrives at the microphones with a delay and a particular angle relative to the array. If suitable delays are applied, all the advanced signals could be time-aligned and their sum would lead to the cancelation of any uncorrelated noise. It is frequency-dependent and the frequency selectivity generally depends on the size of the array and the distance between the microphones. Beamforming methods generally require a large number of microphones for an improved performance, and typically need prior information about the source directions.

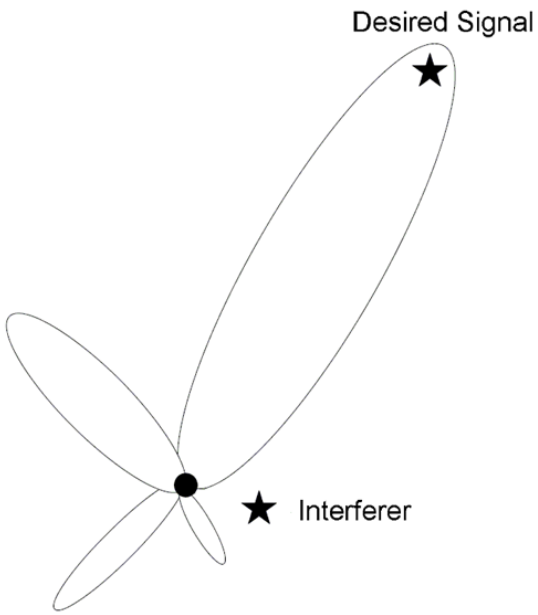


Figure 2.6. An illustration of a beam pattern with the main lobe pointed towards the desired source and a null towards the interferer.

A large number of microphones are required for a beamformer to achieve separation, in contrast, humans use only two ears to perform the same task. Further issues with beamforming are array geometry, as a uniform linear array will not provide 360 degrees azimuth response whereas a circular array can overcome this generally with more microphones. Moreover, the spacing between microphones is critical for a broadband signal such as speech as it will determine the limitations of the response of the array for example due to spatial aliasing [7, 35].

Computational auditory scene analysis based methods aim to mimic the abilities of the human hearing system, but utilize mixtures from either one or two microphones. These techniques are discussed in detail in the next section.

2.6 Computational Auditory Scene Analysis

Computational Auditory Scene Analysis (CASA)-based source separation methods are inspired by the human auditory processing and exploit the cues that humans make use of within the auditory scene analysis [36]. These methods generally utilize mixtures from one microphone (monaural) or two microphones (binaural), and typically exploit the time-frequency signal representation, also referred to as time-frequency (TF) masking or ideal binary mask (IBM).

TF masking relies on the assumption of signal sparseness i.e. the majority of the samples of each signal are almost zero and thus the sources rarely overlap [24]. A TF mask (or filter) is based on a TF representation of a signal, commonly obtained by a short-time Fourier transform (STFT) [37]. Broadly speaking, masks could either be binary (hard) or soft (probabilistic). Speech sources can be perfectly demixed via binary TF masks provided the TF representations of the sources do not overlap [38], a condition that Yilmaz and Rickard [38] term *W-disjoint orthogonality*. Let $S_1(\omega, t)$ and $S_2(\omega, t)$ be the STFT of two speech signals $s_1(t_s)$ and $s_2(t_s)$ respectively. Then the W-disjoint orthogonality (WDO) assumption can be written as

$$S_1(\omega, t)\ S_2(\omega, t) = 0, \forall\ \omega, t \tag{2.6.1}$$

where t denotes the time index and ω is the frequency index. Speech signals have generally been found to have sparse time-frequency representations and satisfy a weakened form of eq. (2.6.1) in that the product of their TF representations is almost always small [38].

Roman et al. [39] and Yilmaz and Rickard [38] provided a study for binaural speech separation. In [39] authors used spatial localization cues: interaural time differences (ITD) and interaural intensity differences (IID) for speech separation. Their work was motivated by the way in which the human

auditory system performs the speech separation task. In [38] the authors introduced the concept of approximate W-disjoint orthogonality. They showed that ideal binary TF masks do exist that could separate multiple speech signals from a single mixture. The Degenerate Unmixing Estimation Technique (DUET) technique [38,40] calculates a two-dimensional histogram of the observed interaural level and time differences, and finds its peaks which would correspond to the number of sources. They approximated masks when two anechoic mixtures were given, assumed that the interaural cues were constant at all frequencies, and that there was no spatial aliasing, which limits its use in practical reverberant situations.

The TF masking based methods have further been developed and are combined with either ICA or beamforming in several studies, discussed in the following sections.

2.6.1 ICA and TF Masking

Kolossa and Orglmeister [41] proposed *non-linear post-processing* in the form of TF masking applied to the output of the frequency-domain ICA. Tests were based only for the special case of two sources and sensors. Initially, signal estimates were obtained by applying ICA to the mixtures. Direction of arrival information was used for permutation correction. The output was then further enhanced by exploiting the approximate WDO of speech signals. The authors claimed that the algorithm was applicable for demixing an arbitrary number of sources as long as the approximate disjoint orthogonality requirement was met. In their proposed *post-processing* method, the magnitudes of the ICA outputs at each frequency bin and at each time frame were compared. Assuming WDO, only one output would be dominant at any given frame and bin. Thus, bins with greatest magnitudes were held and others were set to zero. The combined ICA and TF masking method was applied to in-car, reverberant (RT60 of 300 ms) and artificial speech

recordings. SNR improvement of 15 dB for the in-car case was claimed. The post-processing was tested in conjunction with two ICA algorithms and one beamformer. It was shown that the non-linear post-processing added between 1 dB and 6 dB (3.8 dB on average) to the output SNR. It was concluded that TF masking can significantly improve separation if used as a post-processing step for frequency-domain ICA algorithms.

Araki et al. [42] proposed a solution for under-determined source separation by combining a sparseness approach and ICA. They first extracted one source using binary TF masking and then applied frequency-domain ICA to separate the remaining mixtures. They considered the case with two sensors (microphones) and three speech sources. The speech sources were assumed to be mutually independent and sufficiently sparse in the TF domain. They employed the TF approach because they claim speech signals are more sparse in the TF domain than in the time-domain. The authors pointed out that in [40] the signal sparsity assumption was used to extract signals using a binary TF mask, but the method results in discontinuous zero-padding of the extracted signals and thus are severely distorted (*musical noise* is introduced). The authors claim to have overcome the musical noise problem. They remove only one source with a binary mask in the first stage and separate the remaining sources by ICA in the second stage. Tests were also performed in reverberant conditions with RT60 of 130 ms and 200 ms claiming separation with little distortion.

Araki et al. [43], in a later work, used a continuous (soft) mask instead of a binary mask (which they used earlier), and reported that the signals extracted through binary masks contained loud musical noise. They considered the under-determined case with more sources (I) than sensors (N). The non-binary continuous mask was based on a *directivity pattern*. As they had done previously, in the first stage they remove I-N sources by utilizing the directivity pattern of a null beamformer (which generates nulls towards

the given I-N directions) and employ $N \times N$ ICA at the second stage to separate the remaining sources. Experimental results were given for I = 3, N = 2 and I = 4, N = 2. For I = 3, N = 2 when RT60 = 0 ms, they mentioned that the method by Yilmaz and Rickard [38] gave unsatisfactory signal-to-distortion ratio and a large level of musical noise was also present. While they claimed that with their proposed method they obtained high signal-to-distortion ratio values with no serious deterioration in separation performance. The performance of all methods was worse in the reverberant case with RT60 = 130 ms (compared with the results when RT60 is 0 ms). However, the authors claimed to be able to obtain higher SDR without musical noise compared with the method by Yilmaz and Rickard in a reverberant environment.

Saruwatari et al. [6] proposed a two-stage real-time algorithm by combining a *single-input multiple-output (SIMO)* ICA technique and binary TF masking. In the first stage, the SIMO ICA is used to generate mutiple SIMO signals at each microphone. A binary mask is introduced in the second stage to efficiently reduce the remaining error in ICA. They also considered reverberation and claimed that their method outperformed the conventional ICA and binary masking techniques.

Sawada et al. [44] combined ICA and phase-based TF masking to extract certain dominant sources of interest that were assumed to be close to the sensors, to have dominant power and be non-Gaussian. Unlike their previous work, they initially apply ICA to remove independent components and obtain basis vectors. A TF masking stage follows that reduces the residuals caused by ICA (in the under-determined case). It was claimed that the basis vector normalization and clustering can be used to determine the number of target sources and align the permutation ambiguity of ICA.

Araki et al. [45] presented a new sparse source separation method for non-linearly arranged sensors by utilizing the k-means clustering algorithm

(a commonly used unsupervised learning algorithm) and binary TF masking. Experiments were performed for under-determined conditions with RT60 of 128 ms and 300 ms. The distance (R) between sensors was varied i.e. R = 50, 110, 170 cm. Separation results were shown with two sensors, two-dimensional three sensors, four sensors. It was concluded that the direct-to-reverberant ratio was important for current sparse source separation; and sparse source separation in reverberant conditions was still an open problem.

Pederson et al. in [46] and [47] used an iterative method by combining instantaneous ICA and binary masking to segregate each signal (using only two microphones). Their algorithm flows as follows: a two-input two-output ICA algorithm is applied to the input mixtures, not knowing the number of sources in the mixtures. The estimated outputs of ICA are re-scaled and transformed to the frequency-domain by the use of STFT. Binary TF masks are then determined for each TF unit by comparing the amplitudes of the two spectrograms. Then each of the two binary masks are applied to the original microphone mixtures in the TF domain. After the application of masks the sources are reconstructed in the time-domain by the inverse STFT. A stopping criterion is devised to stop further processing when the signal consists of only one source or when the mask is too sparse. With this iterative algorithm the authors claim to separate successfully mixtures having up to seven speech sources and to have achieved high signal-to-noise ratio (SNR) gains in reverberant conditions (with RT60 of 400 ms). The method proposed in [47] was compared with other methods i.e. with DUET [38] in the instantaneous and convolutive cases and results were given. Their method gave better Δ SNR compared to the instantaneous DUET, while the convolutive DUET gave similar results.

Kolossa et al. [48] combined ICA and TF masking together with uncertainty-based decoding techniques to separate the source of interest when multiple speakers are simultaneously active. They mentioned that by using TF mask-

ing, part of the information of the original signals might be lost along with the interfering sources, thus each estimated mask is considered uncertain. A complex Gaussian uncertainty model was used to estimate the uncertainty in the spectrum domain. A linear four-microphone array was used and experiments were performed in noisy conditions with RT60 of approximately 160 ms.

Sawada et al. [49] proposed a frequency-domain two-stage convolutive source separation method that could also be applied to the under-determined case. In the first stage the expectation-maximization (EM) algorithm is used in which frequency-domain samples of the mixtures are clustered (in a frequency bin-wise manner) into each source. The second stage aligns the permutation ambiguities introduced by the first stage. They claim to obtain good results with this two-stage method even in reverberant conditions. Experimental results were also provided for the under-determined case with reverberation (varied from 130 ms to 450 ms) of four speakers and three microphones. The proposed method was shown to perform best compared to three other BSS methods.

Jan et al. [50] devised a multi-stage approach by combining ICA and ideal binary masking (IBM) to separate convolutive speech mixtures from two microphones. They also apply post-filtering in the cepstral domain. Firstly, they separate the signals from the two-microphones recordings using ICA. They then estimate the IBM by comparing the energies of the corresponding time-frequency units of the separated sources obtained from the first stage. Lastly, they employ cepstral smoothing to reduce the musical noise introduced by TF masking. They evaluated their algorithm for simulated reverberant mixtures as well as real recordings claiming increased efficiency and improved signal quality. Detailed results were provided for a separation example with two sources and sensors with varying Fourier transform frame lengths, RT60s and microphone noise. The proposed algorithm was also

compared with two other methods [47] [50] and provided results in which it outperformed both.

2.6.2 Beamforming and TF Masking

Some studies have focussed on combining beamforming and TF masking to further enhance the separation. Roman and Wang [10] and Roman, Srinvasan and Wang [51] established a method for two-microphone sound separation of mixtures contaminated with interferences and reverberation by utilizing adaptive beamforming. The adaptive beamformer, having known the source directions, first cancels the target source. Then the TF units that were highly attenuated in the first stage (to have likely originated from the target location) are set to unity to get an estimate of the IBM.

Boldt et al. [52] use two cardioids (first-order differential beamformers) to calculate the IBM. Having the information of the directions of target and interfering signals, both the cardioids that are pointing in opposite directions provide the basis for IBM estimation. A theoretical derivation was provided and it was shown that it is possible to calculate the IBM without having access to the unmixed signals.

In [53] Beh et al. proposed a two-stage algorithm to separate two sound sources by combining matched beamforming and TF masking. The beamformer estimates the sources and then the residual interference is suppressed by TF masking. The locations of the sources were assumed to be known and to estimate the impulse response the beamformer uses a least-squares method. The beamwidth of the beamformer was controlled to preserve the original source content to a maximum. The output of the beamformer still contained unwanted acoustic content which was further reduced by using TF masking.

2.6.3 Other related work

Aarabi et al. [54] proposed a multi-microphone TF masking technique that uses both the magnitude and phase information of the TF *blocks* (units) for comparison. They assume that the direction (or the *time-delay of arrival*, as they call it) of the target speaker is known. They mentioned that the popular source separation techniques (ICA, beamforming, and others) are not specifically designed to deal with speech signals. Utilizing certain characteristics of speech could greatly enhance the signal separation problem. They claimed that their algorithm was capable of preserving speech features from the direction of interest and degrading features from other directions. The two noisy mixtures from two microphones were first transformed into frequency-domain representations. A phase-error was derived for each TF unit based on the information from the two microphones. Each TF unit for each microphone was given a value between zero and one. The TF units with smaller phase-error were '*rewarded*' by larger value '1' and TF units with large phase-errors were '*punished*' by a small value '0'.

Later in 2004, Aarabi and Shi [55] based their two-microphone algorithm upon phase-error based filters which depend only on the phase of the signals. First, TF phase-error filters are obtained. The time difference of arrival (TDOA) of sources and phases of microphone signals were assumed to be known. The individual TF units were *rewarded* or *punished* based on the observed and expected phases of those units. Their aim was to maintain the spectral structure of the sources thus preserving the main contents of the speech source. Soft masking was utilized and experiments were performed both in anechoic and low reverberant (RT60 = 100 ms) conditions. The authors mentioned that the SNR gain simulations were useful but could not truly portray the effectiveness of the speech enhancement technique. A better way was to test the output on a speech recognition system. A speaker-independent digit recognition system was used for testing.

Takenouchi and Hamada [56] applied TF masking to an equilateral triangle array to obtain three delay estimates for each microphone pair. Cermak et al. [57] proposed a three-stage algorithm employing TF binary masking, beamforming and non-linear post-processing. They claim that their method removes the musical noise (introduced by binary TF masking) and suppresses the interference in all time-frequency slots.

Given the binaural mixtures, Mouba and Marchand [58] used expectation-maximization based clustering, where the interaural level and time differences at each TF point are mapped to an azimuth angle to separate the sources. Mandel et al. [59] model the interaural spatial parameters as Gaussian distributions and use expectation-maximization to estimate the model parameters. The posterior probabilities, after a fixed number of iterations, are used to construct probabilistic masks for each source, with the assumption that total number of sources are known *a priori*.

The different methods described above are able to perform source separation in constrained scenarios i.e. with either no or very low levels of reverberation, and consider the simple case of mixtures of only two sound sources. The performance of even the state-of-the-art methods in realistic reverberant and multi-speaker environments is limited. New techniques need to be developed that could tackle the reverberation problem well and provide improved performance in multi-source scenarios. This thesis focuses on the development of such algorithms, by exploiting the knowledge of the locations of the sound sources that could either be known a priori or calculated by a video processing system. Estimating these locations is not within the scope of this thesis, but further details can be found in [12, 13, 60, 61].

Evaluating the performance of source separation systems is discussed next.

2.7 Performance Evaluation Metrics

An important task in the evaluation of source separation algorithms is to have suitable subjective and/or objective evaluation metrics to quantify how well the algorithms have performed. In the speech separation context, typically there are two things to focus within the processed speech: the overall speech quality and the intelligibility. The overall speech quality is generally the notion of the listener relating to the perceived speech in that how well it sounds. Whereas, the intelligibility has to do with perceiving the content of the utterance in that what is being uttered. In general, speech rated as of good quality is highly intelligible and vice versa; however, speech perceived as of bad quality may give a high intelligibly score [62]. Subjective listening tests are the most accurate way of performance evaluation, but they are expensive, require intensive labour and thus are time-consuming. Objective measures have therefore been developed. Different evaluation measures are used in different domains depending on the type of processing involved and the distortions produced due to that processing [63]. The evaluation metrics used in this thesis are discussed as follows. The main motivation for using these specific metrics was because of their usage by the wider research community within this research area and their suitability for the different algorithms developed in this thesis.

2.7.1 Objective Measures

Signal-to-distortion ratio

The Signal-to-distortion ratio (SDR) which is the ratio of the energy in the original signal to the energy in interference from other signals and other artifacts proposed in [64] is used as an evaluation metric throughout this thesis. The implementation provided in BSS_EVAL toolbox is utilized. Consider the anechoic original time-domain signals be represented as $s_i(t_s), \cdots, s_I(t_s)$,

the anechoic target signal denoted as $s_t(t_s)$, and the estimated target as $\hat{s}_t(t_s)$. The SDR is expressed in terms of the three time-domain signals produced by projecting the estimated signal onto the space of the original signals i.e. the target signal $s_{targ}(t_s)$, the error caused by interference, $e_{intf}(t_s)$, and the error because of the artifacts, $e_{artf}(t_s)$. Let $P(\cdot)$ be the projection operator and τ_{max} be the maximum number of samples utilized in the shifting process, the three signals can be expressed as [65]

$$s_{targ}(t_s) = P(\hat{s}_t, s_t, \tau_{max}) \tag{2.7.1}$$

$$e_{intf}(t_s) = P(\hat{s}_t, \{s_i\}, \tau_{max}) - P(\hat{s}_t, s_t, \tau_{max}) \tag{2.7.2}$$

$$e_{artf}(t_s) = \hat{s}_t - P(\hat{s}_t, \{s_i\}, \tau_{max}) \tag{2.7.3}$$

SDR can be written as

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{intf} + e_{artf}\|^2} \tag{2.7.4}$$

where $\|\cdot\|^2$ denote square of the vector 2-norm (the sum of squares of all entries). Late reverberation from the sources and any other unexplained noise (including musical noise) is considered as the artifact error.

The signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR), defined below, are also used in Chapter 3.

$$SIR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{intf}\|^2} \tag{2.7.5}$$

$$SAR = 10\log_{10}\frac{\|s_{target}\|^2 + \|e_{intf}\|^2}{\|e_{artf}\|^2} \tag{2.7.6}$$

In contrast to SDR, the SIR metric does not penalize reverberation. In Chapter 6 where dereverberation methods are studied, the SNR and the segmental SNR (segSNR) are also used for evaluation. The SNR is defined

as

$$SNR = 10\log_{10}\frac{\sum_{t_s=1}^{T} s_t^2(t_s)}{\sum_{t_s=1}^{T}(s_t(t_s) - \hat{s}_t(t_s))^2} \tag{2.7.7}$$

where T is the length of the signal. The segSNR is a frame based measure which is obtained by averaging frame level estimates of SNR [63, 66].

Performance Index

Adopting Performance Index (PI) as an evaluation metric is motivated by assessing the performance at each frequency bin to provide an insight into the separation achieved by the frequency-domain convolutive source separation algorithm. Utilizing the matrices $\mathbf{H}$ and $\mathbf{W}$ in eq. (2.4.3) and (2.4.4), the matrix $\mathbf{G}$ is obtained as, $\mathbf{G} = \mathbf{WH}$. Assuming that the number of source signals equals the number of mixtures, the PI as a function of matrix $\mathbf{G}$ is written as [28, 67, 68]

$$\begin{aligned} PI(\mathbf{G}) = &\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{m}\frac{abs(G_{ik})}{max_k abs(G_{ik})} - 1\Big)\Big] \\ &+\Big[\frac{1}{m}\sum_{k=1}^{m}\Big(\sum_{i=1}^{n}\frac{abs(G_{ik})}{max_i abs(G_{ik})} - 1\Big)\Big] \end{aligned} \tag{2.7.8}$$

where G_{ik} is the ik-th element of $\mathbf{G}$. Lower bound for the PI is zero while the upper bound is the function of the normalization factor. PI with a value zero means superior separation performance. The algorithm detailed in Chapter 3 is evaluated with this criterion.

Perceptual Evaluation of Speech Quality

The Perceptual Evaluation of Speech Quality (PESQ) measure is an international telecommunication union (ITU-T) standard originally designed for the assessment of speech quality within telephony applications. PESQ compares the original and the processed (separated) signals after transforming them to a representation that is inspired by psychoacoustics. PESQ is used

in Chapter 6 of this thesis. Precisely, the implementation provided in [69] is used.

2.7.2 Subjective Measures

Mean opinion score

In many applications, the ultimate goal of the sound processing algorithms is an enhanced human listening experience. In Mean opinion score (MOS), the algorithm performance is subjectively measured by conducting listening experiments involving human subjects. MOS tests for voice are specified by the ITU-T recommendation P.800 with the following scale (Table 2.1). Subjects listen to the processed signals and give their opinions. The arithmetic mean of a collection of these opinion scores is termed as the mean opinion score. MOS is used in Chapter 3 of this thesis.

Table 2.1. Opinion scale.

Category rating	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

2.8 Summary

This chapter provided background of the important issues relating to sound source separation in reverberant enclosures. It highlighted the hazard posed by reverberation and discussed some approaches to the source separation problem. CASA-based methods, that aim to model the cues that humans make use of while performing the source segregation task, were reviewed in detail. Different methods used on their own or in conjunction with TF masking were then reviewed followed by a description of different perfor-

mance evaluation measures. The following points highlight some limitations of the current separation algorithms and demand for an improved performance, specifically in reverberant scenarios.

1. Most of the works, for instance, [38], [56], [46], [47] consider an anechoic environment (no reflections occur) and thus their mixing systems are either anechoic or instantaneous. Instantaneous BSS does not take signal propagation delay and reverberation into account. They can not model real-world scenarios that are convolutive.

2. Room reverberation poses great threat to the source separation problem. Even the most sophisticated algorithms are practically ineffective with medium or high level of reverberation i.e. with $RT60 > 300ms$ or $RT60 > 500$ ms respectively. Since, a realistic average-sized office room may have an RT60 of 500 ms or more, there is a need for more robust techniques that work well in reverberant conditions.

3. Most earlier works have focussed on scenarios with two speakers only e.g. [47], [38]. Robust algorithms need to be developed to separate more than two speech sources in order to be applicable in practical situations.

4. TF masking is mostly exploited in the under-determined area. There are instances where multiple-microphone algorithms need to be used e.g beamforming. Work is required to incorporate TF masking in these conditions to enhance the separation process.

The above-mentioned points provide sufficient motivation for the development of new algorithms that are more efficient in real-world reverberant environments. The rest of this thesis will aim to develop such algorithms with the assumption that the locations of the sound sources are known. In the following chapter, a multi-microphone based method is proposed that also utilizes binary time-frequency masking.

Chapter 3

BEAMFORMING AND BINARY TIME-FREQUENCY MASKING FOR SOURCE SEPARATION

3.1 Introduction

This chapter presents a novel multi-microphone source separation approach which exploits spatial beamforming and binary time-frequency masking. Typically, for sound sources measured in reverberant rooms, for instance with reverberation time over 300 ms, the performance of audio-only blind source separation (BSS) methods is limited. Therefore, in the proposed approach, the source location information is utilized to facilitate a robust least squares frequency invariant data independent (RLSFIDI) beamformer. The convex optimization approach in the beamformer design also allows compensation for the possible uncertainties in source location and direction of arrival estimates. Sources separated by the RLSFIDI beamformer are further enhanced by applying a binary time-frequency masking technique as a post-filtering process. The RLSFIDI beamformer design for linear array configurations in a 3-D room environment is explained in the following sec-

tion.

3.2 Robust Least Squares Frequency Invariant Data Independent Beamformer

The least squares approach is a suitable choice for data independent beamformer design [7], by assuming the over-determined case with $N > I$, which provides greater degrees of freedom. The over-determined least squares problem for the beamformer design for one of the sources is obtained as

$$\min_{\mathbf{w}(\omega)} ||\mathbf{H}^T(\omega)\mathbf{w}(\omega) - \mathbf{r}_d(\omega)||_2^2 \tag{3.2.1}$$

where $\mathbf{r}_d(\omega)$ is an Ix1 desired response vector and can be designed from a 1D window e.g. the Dolph-Chebyshev or Kaiser windows [70], $\mathbf{w}^T(\omega)$ is one of the beamformer weight vectors which corresponds to one row vector of $\mathbf{W}(\omega)$ in (2.4.4), and $(\cdot)^T$ and $||\cdot||_2$ denote respectively the transpose operation and the Euclidean norm.

A frequency-invariant beamformer design can be obtained by assuming the same coefficients for all frequency bins i.e. $\mathbf{r}_d(\omega) = \mathbf{r}_d$ [71]. If the wavelengths of the low frequencies of the source signals are greater than twice the spacing between the microphones then this design leads to spatially white noise [70]. In audio-only (unimodal) CSS systems there are no priori assumptions on the source statistics of the mixing system. Assuming that the sound source locations are known, the mixing filter is formulated as $\mathbf{H}(\omega) = [\mathbf{d}(\omega, \theta_1, \phi_1), ..., \mathbf{d}(\omega, \theta_I, \phi_I)]$, where $\mathbf{d}(\cdot)$ denotes the beamformer response vector and θ and ϕ are the elevation and azimuth angles. The elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the center of the microphone array are calculated as $r_i = \sqrt{(u_{x_i} - u'_{x_m})^2 + (u_{y_i} - u'_{y_m})^2 + (u_{z_i} - u'_{z_m})^2}$, $\theta_i = \tan^{-1}(\frac{u_{y_i} - u'_{y_m}}{u_{x_i} - u'_{x_m}})$, $\phi_i = \sin^{-1}(\frac{u_{y_i} - u'_{y_m}}{r_i Sin(\theta_i)})$, where u_{x_i}, u_{y_i} and u_{z_i} are the 3-D locations of the speaker i, while u'_{x_m}, u'_{y_m} and u'_{z_m} are Cartesian coordinates

of the center of the microphone array.

The 3-D positions of N-microphone array, with the sensors equally spaced, are written in matrix form as

$$\mathbf{U}^{'} = \begin{bmatrix} u^{'}_{x_1} & u^{'}_{y_1} & u^{'}_{z_1} \\ \vdots & \vdots & \vdots \\ u^{'}_{x_N} & u^{'}_{y_N} & u^{'}_{z_N} \end{bmatrix} \tag{3.2.2}$$

where the Cartesian coordinates of the m-th sensor (microphone) are in the m-th row of matrix $\mathbf{U}^{'}$.

The beamformer response $\mathbf{d}(\omega, \theta_i, \phi_i)$ for frequency bin ω and for source of interest (SOI) $i = 1, ..., I$, can be derived [72] as

$$\mathbf{d}(\omega, \theta_i, \phi_i) = \begin{bmatrix} \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).u^{'}_{x_1} + \sin(\theta_i). \\ \sin(\phi_i).u^{'}_{y_1} + \cos(\theta_i).u^{'}_{z_1})) \\ \vdots \\ \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).u^{'}_{x_N} + \sin(\theta_i). \\ \sin(\phi_i).u^{'}_{y_N} + \cos(\theta_i).u^{'}_{z_N})) \end{bmatrix} \tag{3.2.3}$$

where $\kappa = \omega/c$ and c is the speed of sound in air at room temperature i.e 343 m/s.

To design the beam pattern which allows the SOI, and to better block the interferences in the least squares problem in (3.2.1), the following constraints are used

$$\begin{aligned} &\mid \mathbf{w}^H(\omega)\mathbf{d}(\omega, \theta_i + \Delta\theta, \phi_i + \Delta\phi) \mid \; = \; 1 \\ &\mid \mathbf{w}^H(\omega)\mathbf{d}(\omega, \theta_j + \Delta\theta, \phi_j + \Delta\phi) \mid \; < \; \varepsilon \qquad \forall\omega \end{aligned} \tag{3.2.4}$$

where θ_i, ϕ_i and θ_j, ϕ_j, $j = 1, ..., I$ except i, are respectively, the angles of arrival of the SOI and interference, and $\Delta\theta$ and $\Delta\phi$ have angular ranges defined by $\alpha_1 \leq \Delta\theta \leq \alpha_2$ and $\alpha_3 \leq \Delta\phi \leq \alpha_4$, where α_1, α_3 and α_2, α_4 are lower and upper limits respectively, and ε is the bound for interference.

The white noise gain (WNG) is a measure of the robustness of a beamformer and a robust superdirectional beamformer can be designed by constraining the WNG. Superdirective beamformers are extremely sensitive to small errors in the sensor array characteristics and to spatially white noise. The errors due to array characteristics are nearly uncorrelated from sensor to sensor and affect the beamformer in a manner similar to spatially white noise. The WNG is also controlled here by adding the following constraint

$$\mathbf{w}^H(\omega)\mathbf{w}(\omega) \leq \frac{1}{\gamma} \qquad \forall\omega \tag{3.2.5}$$

where γ is the bound for the WNG.

The constraints in (3.2.4) for each discrete pair of elevation and azimuth angles, and the respective constraint for WNG in (3.2.5) are convex [70]. And the unconstrained least squares problem in (3.2.1) is a convex function, therefore convex optimization [73] is used to calculate the weight vector $\mathbf{w}(\omega)$ for each frequency bin ω.

Finally, $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), ..., \mathbf{w}_I(\omega)]^T$ is placed in the equation, $\mathbf{y}(\omega, t) = \mathbf{W}(\omega)\mathbf{x}(\omega, t)$, to estimate the sources. These estimated sources are further enhanced by applying the binary time-frequency masking technique, discussed in the following section.

3.3 Post-Processing: Binary TF Masking

As mentioned above, the RLSFIDI beamformer accepts the target signal from a certain direction and suppresses interferences and reflections, but

the removal of interference is not perfect, therefore the ideal binary mask (IBM) technique is used as a post-processing stage. The block diagram of combining the output of the RLSFIDI beamformer and TF masking is shown in Fig. 3.1. The separated time-domain speech signal $y_i(t_s)$ of speaker i is converted into the time-frequency domain $y_i(\omega, t)$, where ω is the normalized frequency index. By using a T-point windowed discrete short-time Fourier transformation the spectrograms are obtained as

$$y_i(\omega, t) = STFT(y_i(t_s)) \quad i = 1, ..., I \tag{3.3.1}$$

where t and ω respectively represent time and frequency bin indices.

From the above TF representations, binary masks are estimated by comparing the amplitudes of the spectrograms [2,74]. The binary masks for three audio sources are estimated as

$$BM_1(\omega, t) = \begin{cases} 1, & \text{if } |y_1(\omega, t)| > \tau |y_2(\omega, t)| \quad \& \quad |y_1(\omega, t)| > \tau |y_3(\omega, t)| \\ 0, & \text{otherwise} \qquad\qquad\qquad \forall (\omega, t) \end{cases} \tag{3.3.2}$$

$$BM_2(\omega, t) = \begin{cases} 1, & \text{if } |y_2(\omega, t)| > \tau |y_3(\omega, t)| \quad \& \quad |y_2(\omega, t)| > \tau |y_1(\omega, t)| \\ 0, & \text{otherwise} \qquad\qquad\qquad \forall (\omega, t) \end{cases} \tag{3.3.3}$$

$$BM_3(\omega, t) = \begin{cases} 1, & \text{if } |y_3(\omega, t)| > \tau |y_1(\omega, t)| \quad \& \quad |y_3(\omega, t)| > \tau |y_2(\omega, t)| \\ 0, & \text{otherwise} \qquad\qquad\qquad \forall (\omega, t) \end{cases} \tag{3.3.4}$$

where τ is a parameter to control how much of the interfering signals should be removed at each iteration [2, 74].

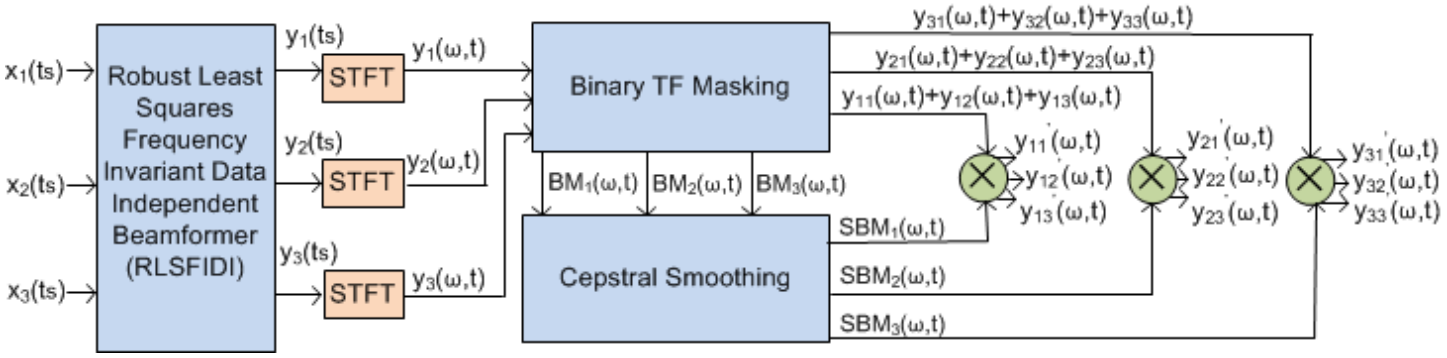


Figure 3.1. The RLSFIDI beamformer with binary TF masking as a post-processing technique provides estimates of the speech sources $y_i(t_s)$, where $i = 1, .., I$. The separated speech signals are transformed into the TF domain $y_i(\omega, t)$, using the STFT. Binary masks $\mathrm{BM}_i(\omega, t)$ are then estimated by comparing the energies of the individual TF units of the source spectrograms. The cepstral smoothing stage follows that smooths the estimated binary masks and $SBM_i(\omega, t)$ is obtained. The smoothed binary masks $SBM_i(\omega, t)$ are utilized to enhance the signals separated by the RLSFIDI beamformer.

Each of the three binary masks are then applied to the original mixtures in the time-frequency domain in order to enhance the separated signals as

$$y_i(\omega, t) = BM_i(\omega, t)x_i(\omega, t) \quad i = 1, 2, 3. \tag{3.3.5}$$

The enhanced signals are transformed to the time-domain by applying an inverse short-time Fourier transform (ISTFT).

This binary mask based TF technique considerably improves the separation performance of the RLSFIDI beamformer by reducing the interferences to a much lower level which ultimately provides better estimates of the separated speech signals. However, a problem with the binary masking is the introduction of errors in the estimation of the masks i.e. fluctuating musical noise [74]. To overcome the musical noise a cepstral smoothing technique [74, 75] is used.

3.3.1 Cepstral smoothing technique

In the cepstral smoothing the estimated IBM is first transformed into the cepstral domain, and different smoothing levels, based on the speech production mechanism, are then applied to the transformed mask. The smoothed mask is converted back to the spectral domain. In this method the musical artifacts within the signals can be reduced. The broadband structure and pitch information of the speech signal are also well preserved without being noticeably affected by the smoothing operation [74]. The estimated masks in (3.3.2), (3.3.3) and (3.3.4) can be represented in the cepstral domain as:

$$BM_i^c(l, t) = DFT^{-1}\{\ln(BM_i(\omega, t)) \mid_{\omega=0,..,T-1}\} \quad i = 1, 2, 3 \tag{3.3.6}$$

where l is the quefrency bin index; DFT and ln denote the discrete Fourier transform and the natural logarithm operator respectively; T is the length

of the DFT and after applying smoothing, the resultant smoothed mask is given as:

$$BM_i^s(\omega,t) = \beta_l BM_i^s(l,t-1) + (1-\beta_l)BM_i^c(l,t) \tag{3.3.7}$$

where β_l controls the smoothing level and is selected according to different values of quefrency l

$$\beta_l = \begin{cases} \beta_{env} & \text{if } l \in \{0,...,l_{env}\}, \\ \beta_{pitch} & \text{if } l = l_{pitch}, \\ \beta_{peak} & \text{if } l \in \{(l_{env}+1),...,T\} \setminus l_{pitch} \end{cases} \tag{3.3.8}$$

where l_{env} and β_{pitch} are respectively quefrency bin indices for the spectral envelope and the structure of the pitch harmonics in $BM_i(\omega,t)$, and $0 \leq \beta_{env} < \beta_{pitch} < \beta_{peak} \leq 1$. The symbol "\" excludes l_{pitch} from the quefrency range $(l_{env}+1),...,T$. The details of the principle for the range of β_l and the method to calculate β_{peak} are described in [74]. The final smoothed version of the spectral mask is given as:

$$SBM_i(\omega,t) = \exp(DFT\{BM_i^s(\omega,t)\mid_{l=0,...,T-1}\}). \tag{3.3.9}$$

The smoothed mask is then applied to the segregated speech signals in (3.3.5) as follows:

$$\overline{y}_i(\omega,t) = SBM_i(\omega,t)y_i(\omega,t). \tag{3.3.10}$$

Finally, by applying the ISTFT, $\overline{y}_i(\omega,t)$ is converted back to the time-domain. The experimental results based on objective and subjective evaluations are presented in the following section.

3.4 Experiments and Results

Simulations are performed in a room with dimensions $4.6 \times 3.5 \times 2.5\ m^3$. Audio signals sampled at 8 KHz were used. Room impulse responses were generated by the image method [76] for reverberation time (RT60) of 300, 450, and 600 ms. The RT60 was controlled by varying the absorption coefficient of the walls. The source image method assumes point sources which radiate isotropic pressure waves [76]. This is an assumption which allows generation of synthetic impulse responses. In reality the sound emitted by a human is directional therefore from Chapter 4 we also include evaluations with real room impulse responses.

A linear array configuration of sixteen equally spaced microphones, $N = 16$, was used. The distance between the microphones was 4 cm. The other important variables were selected as: STFT length $T = 1024$ & 2048 and filter lengths were $Q = 512$ & 1024, the Hamming window was used with the overlap factor set to 0.75. Duration of the speech signals was 7 seconds, $\tau = 1$, $\varepsilon = 0.1$, $\gamma = -10$ dB, for SOI $\alpha_1 = +5$ degrees and $\alpha_2 = -5$ degrees, for interferences $\alpha_1 = +7$ degrees and $\alpha_2 = -7$ degrees, speed of sound $c = 343$ m/s, $l_{env} = 8$, $l_{low} = 16$, and $l_{high} = 120$, and parameters for controlling the smoothing levels were $\beta_{env} = 0$, $\beta_{pitch} = 0.4$, $\beta_{peack} = 0.8$.

Note that the locations (and thus the direction of arrivals (DOAs)) estimated from the video recordings may contain errors, so in the simulations, the exact DOAs of the sources are perturbed by zero-mean Gaussian noise with a standard deviation of 3 degrees, which corresponds approximately to the average of that for the three speakers given in Fig. 5 of [61]. Such a simulation set-up is assumed throughout Chapters 3-6.

Evaluation Criteria: The objective evaluation of the algorithms include performance index (PI) [77], signal-to-interference-noise ratio (SINR) and $\Delta\text{SINR} = \text{SINR}_o - \text{SINR}_i$, percentage of energy loss (PEL), percentage of

noise residue (PNR) [2]; signal-to-distortion ratio (SDR), signal-to-interference (SIR) ratio, and signal-to-artifact ratio (SAR) [78]. For a signal separated using a binary time-frequency mask, the PEL and PNR measures are mathematically written as [2, 79]

$$PEL = \frac{\sum_{t_s=1}^{T}(e_1^t(t_s))^2}{\sum_{t_s=1}^{T}(I^t(t_s))^2} \tag{3.4.1}$$

$$PNR = \frac{\sum_{t_s=1}^{T}(e_2^t(t_s))^2}{\sum_{t_s=1}^{T}(y^t(t_s))^2} \tag{3.4.2}$$

where $y^t(t_s)$ represents the estimated signal obtained from RLSFIDI beamformer and $I^t(t_s)$ is the resynthesized signal obtained after applying the smoothed estimated masks; $e_1^t(t_s)$ is the signal present in $I^t(t_s)$ but absent in $y^t(t_s)$ and similarly $e_2^t(t_s)$ is the signal present in $y^t(t_s)$ but absent in $I^t(t_s)$ [2].

SINR_i is the ratio of the desired signal to the interfering signal taken from the mixture. SINR_o is the ratio of the desired signal resynthesized from the ideal binary mask to the difference of the desired resynthesized signal and the estimated signal [2]. The separation of the speech signals is evaluated subjectively by listening tests. Mean opinion scores (MOS tests for voice are specified by ITU-T recommendation P.800) are also provided.

In the first set of simulations, two tests were performed on mixtures with an RT60 of 130 ms, which were separated by the original independent vector analysis (IVA) based method [1] and the RLSFIDI beamformer. From the known source locations, the respective elevation and azimuth angles were obtained and were used by the RLSFIDI beamformer. The resulting performance indices of the first test are shown in Fig. 3.2(a) and the performance of the original IVA method for the same test is shown in Fig. 3.2(b). The other objective evaluations for both tests are shown in Table 3.1. These separations were also evaluated subjectively with MOS [STD]= 4.1 [0.15]

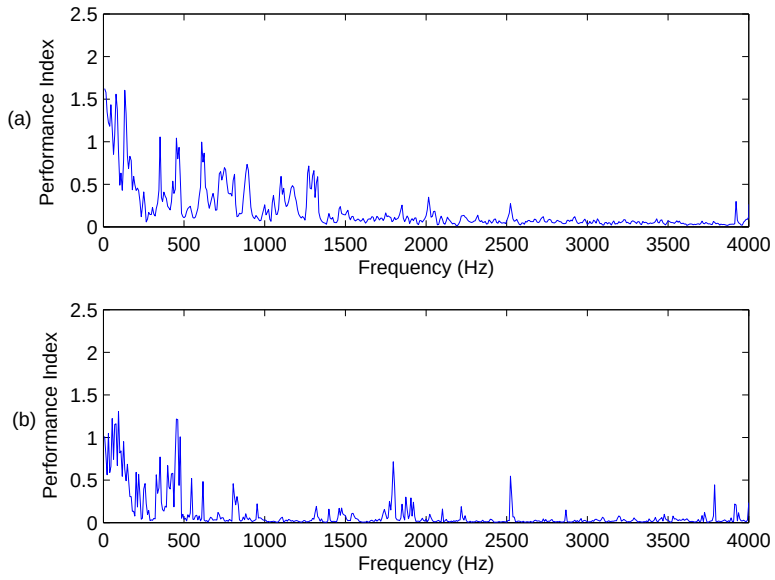


Figure 3.2. Performance index at each frequency bin for (a) the RLSFIDI beamformer and (b) the original IVA method [1], length of the signals is 7 s. A lower PI refers to a superior method. The performance of the IVA method is better than the RLSFIDI beamformer at RT60 = 130 ms.

and 4.2 [0.13] for the RLSFIDI beamformer and IVA methods respectively. The performance of the higher-order statistics based IVA method at RT60 = 130 ms with data length = 7 s is better than the RLSFIDI beamformer. The output of the RLSFIDI beamformer was further enhanced by the IBM technique. The masks of clean, estimated and enhanced speech signals are shown in Figs. 3.3, 3.4 & 3.5 respectively. The highlighted areas, compared with the corresponding ones in Figs. 3.3, 3.4 & 3.5 show how the post-filtering technique improves the speech signals separated by the RLSFIDI beamformer at the post-filtering process stage. In particular, the regions highlighted in Fig. 3.5 resemble closely the original sources in the regions shown in Fig. 3.3; the IBM technique has removed the granular noise shown in the regions highlighted in Fig 3.4. The post-filtering enhanced the separated speech signals as shown in Table 3.2.

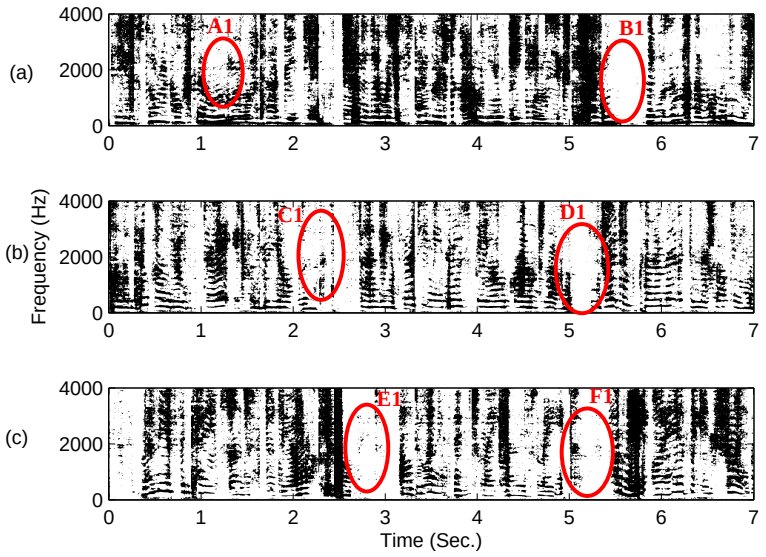


Figure 3.3. Ideal binary masks (IBMs) [2] of the three original speech signals used in the experiment at RT60 = 130 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding ones in Figs. 3.4 & 3.5 show how the post-filtering technique improves the output of the RLSFIDI beamformer.

In the second set of simulations, two tests are performed on the mixtures of length = 7 s for RT60 = 300, 450 & 600 ms, which were separated by the RLSFIDI beamformer and the IVA method [1]. The respective objective evaluations for each RT60 are shown in Table 3.3, which affirms the statement in [80] that with long impulse responses the separation performance of CSS algorithms +(based on second-order and higher-order statistics) is highly limited. For the condition $T > P$, the DFT length was also increased, $T = 2048$, but there was no significant improvement observed because the number of samples in each frequency bin was reduced to $truncate(7Fs/T) = 27$.

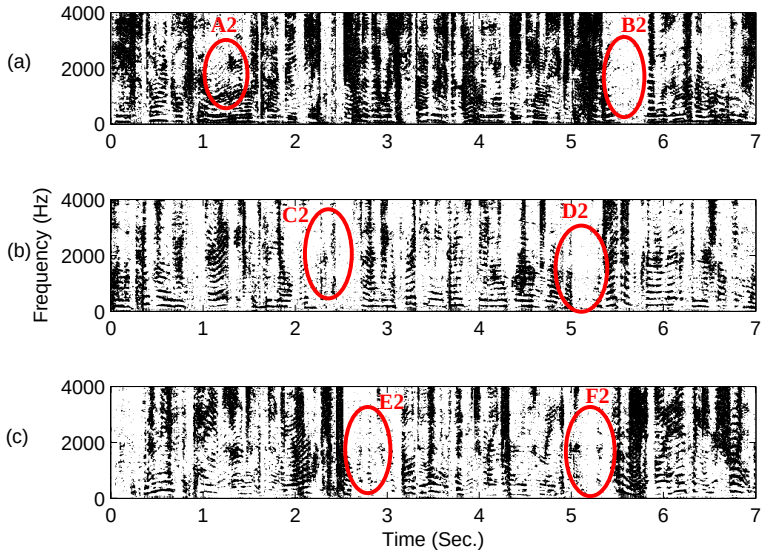


Figure 3.4. Binary masks of the speech signals separated by the RLS-FIDI beamformer at RT60 = 130 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding original speech signals in Fig. 3.3 show that a considerable amount of interference from the other sources still exists when the ΔSINR = 14.97 dB.

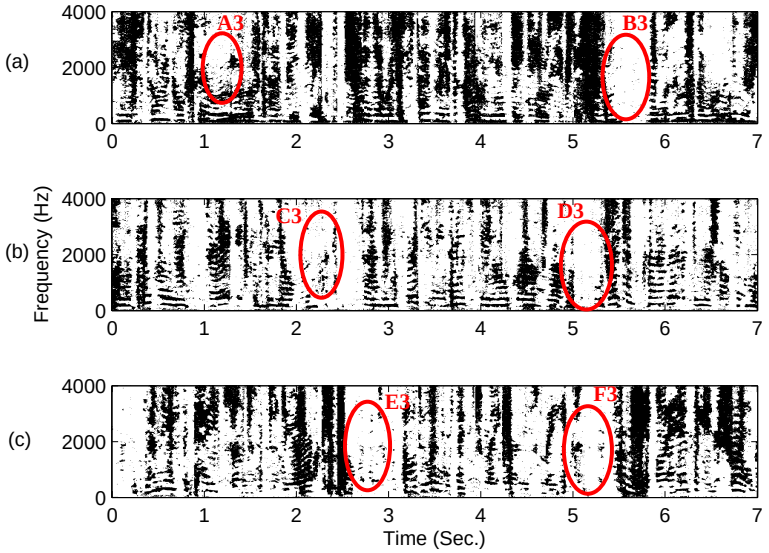


Figure 3.5. Binary masks of the three enhanced speech signals by the IBM TF masking technique at RT60 = 130 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding ones in Figs. 3.3 & 3.4 show the post-filtering processing stage improves the output of the RLSFIDI beamformer. For these enhanced signals PEL = 10.15%, PNR = 11.22%, and SINR = 16.83 dB.

Table 3.1. Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer and the original IVA method [1] at RT60 = 130 ms, the length of the signals is 7 s. Results (in decibels) are provided for two different mixtures.

	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
RLSFIDI	14.97	6.38	0.78	11.59	14.88	7.39	12.24	7.18	2.58	20.41
beamformer	15.03	7.42	0.53	11.81	8.63	7.00	12.87	14.13	2.43	18.66
IVA	16.49	2.35	4.44	5.90	2.90	6.00	7.84	13.04	10.61	10.99
method	16.35	2.40	4.51	6.10	3.30	6.06	8.14	13.32	10.69	11.07

Table 3.2. Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer after post-processing at RT60 = 130 ms, the length of the signals is 7 s. Results (in decibels) are provided for two different mixtures.

	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
RLSFIDI	16.83	6.59	7.90	8.04	14.30	14.37	15.54	7.55	9.17	9.02
beamformer	16.97	6.84	7.99	7.51	15.35	17.02	15.36	7.63	8.65	8.41

Table 3.3. Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer without post-processing and the original IVA method [1], in decibels, for different RT60s. Results are provided for two different mixtures.

RT60 (ms)	Method	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
300	RLSFIDI beamformer	11.25	2.93	0.90	6.91	9.31	4.87	8.97	4.54	4.35	11.66
		11.17	5.60	0.90	6.89	6.71	4.77	8.99	12.91	4.45	11.69
	IVA	12.02	0.96	3.49	5.44	1.74	5.08	7.28	10.99	9.70	10.81
		12.20	0.98	3.45	5.96	1.78	5.05	7.91	10.97	9.75	11.05
450	RLSFIDI beamformer	7.76	-0.77	0.45	4.26	4.94	3.94	7.75	1.78	4.51	7.52
		7.95	3.76	0.45	4.20	5.54	3.94	7.62	9.57	4.51	7.54
	IVA	6.55	-1.94	2.13	2.52	-0.37	4.26	4.70	6.44	7.62	7.75
		6.78	-0.86	2.98	4.09	0.43	5.31	6.33	7.81	7.90	8.93
600	RLSFIDI beamformer	6.30	-3.47	0.53	3.31	2.65	3.37	7.00	-0.37	3.9	4.85
		6.46	0.82	-0.12	2.24	4.19	3.37	6.87	4.89	3.96	4.88
	IVA	5.26	-5.26	-1.30	0.36	-2.88	3.99	5.77	3.17	1.67	2.80
		5.40	-5.14	-0.35	1.21	-3.44	2.80	4.04	4.82	4.35	4.89

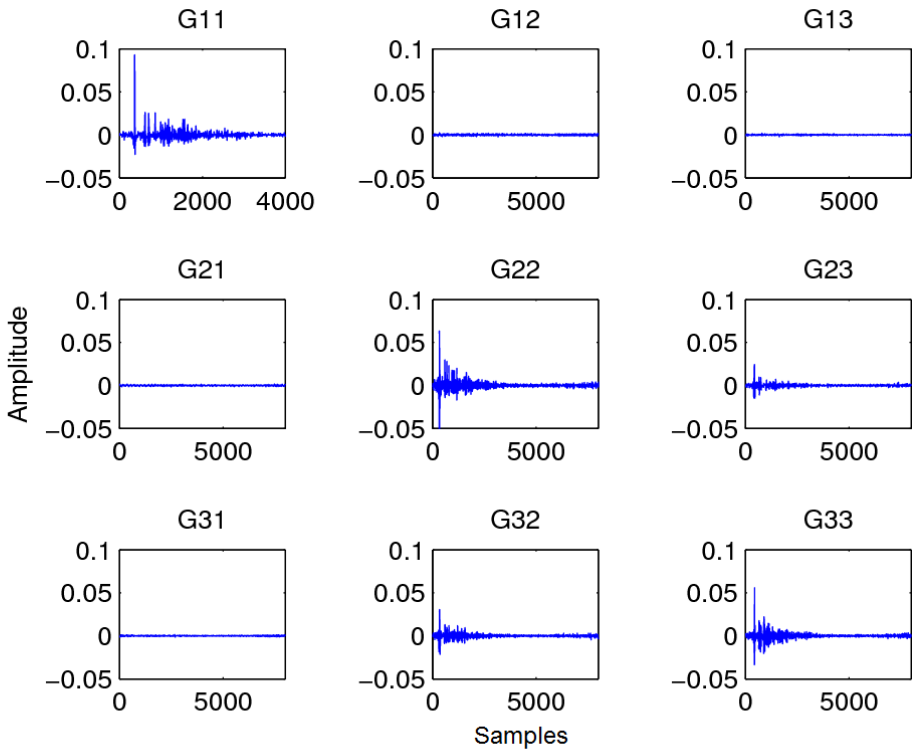


Figure 3.6. Combined impulse response $G = WH$ by the original IVA method. The reverberation time RT60 = 300 ms and SIR improvement was 12.2 dB.

The improved performance of the RLSFIDI beamformer over the original IVA method, specifically, at RT60 = 300 ms (Table 3.3) when ΔSINR of IVA method is higher than the RLSFIDI beamformer, is investigated in Figs. 3.6 & 3.7. Actually, the CSS method removed the interferences more effectively, therefore, the ΔSINR is slightly higher. However, the separated speech signals are perceptually not of an improved quality, because the reverberations are not well suppressed. According to the "law of the first wave front" [81], the precedence effect describes an auditory mechanism which is able to give greater perceptual weighting to the first wave front of the sound (the direct path) compared to later wave fronts arriving as reflections from surrounding surfaces. On the other hand, beamforming accepts the direct path and also suppresses the later reflections therefore the MOS is better. For comparison,

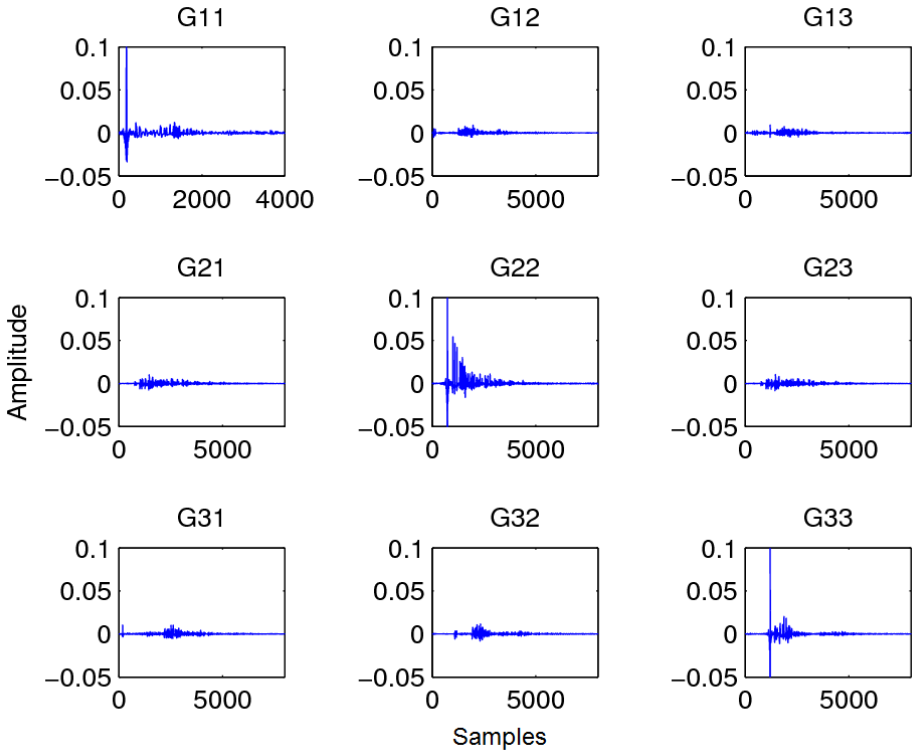


Figure 3.7. Combined impulse response $G = WH$ by the RLSFIDI beamformer. The reverberation time RT60 = 300 ms and SIR improvement was 11.2 dB.

a typical room impulse response for RT60 = 300 ms is shown in Fig. 3.8.

In the final set of simulations, the separated speech signals by the RLSFIDI beamformer for each value of RT60 were further enhanced by applying the IBM technique. The respective objective evaluations for each RT60 are shown in Table 3.4. To show the performance of TF masking as a post-processing stage, the results for RT60 = 300 ms for the first test are presented. The ideal binary masks (IBMs) of the three clean speech sources are shown in Fig. 3.9. In Fig. 3.10 the estimated binary masks (BM_s) of the output signals obtained from the RLSFIDI beamformer are shown. These binary masks are applied on the spectrograms of the three selected microphones and masks of the enhanced speech signals are shown in Fig. 3.11.

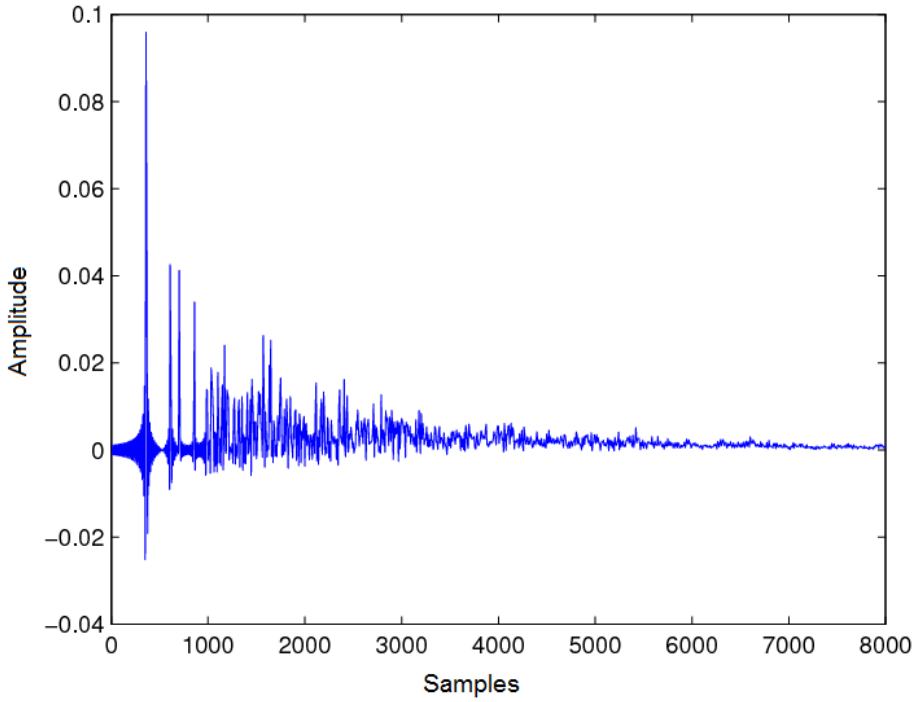


Figure 3.8. A typical room impulse response for reverberation time RT60 = 300 ms is provided for comparison.

For comparison, two regions are shown in one of the three speech signals, which are marked as $G_1, H_1, I_1, J_1, K_1, L_1$ in the IBMs, $G_2, H_2, I_2, J_2, K_2, L_2$ in the SBM_s, and $G_3, H_3, I_3, J_3, K_3, L_3$ in the final separated signals. From the highlighted regions, it can be observed that the interference within one source that comes from the other is reduced gradually in the post-processing stage. The listening tests are also performed for each case and MOSs are presented in Table 3.5, which indicates that at higher RT60 the performance of the RLSFIDI beamformer is better than the IVA algorithm. The proposed solution not only improves the performance at lower RT60s but also at higher RT60 when the performance of conventional CSS algorithms is limited.

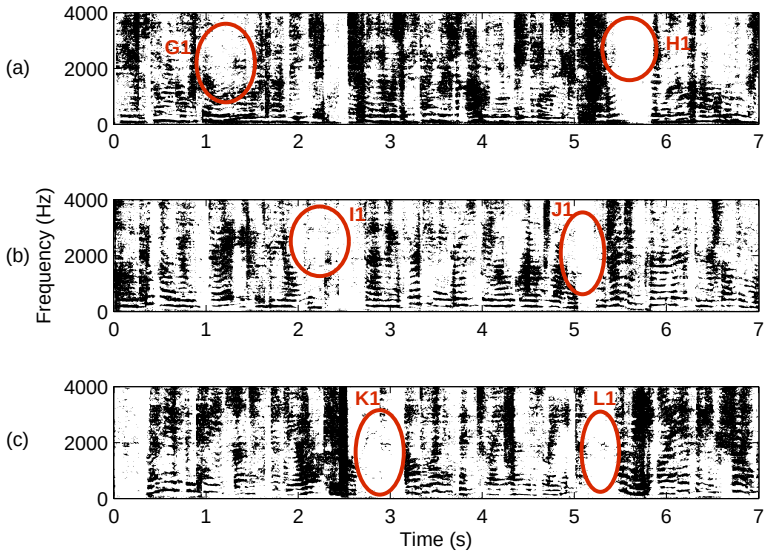


Figure 3.9. Ideal binary masks (IBMs) [2] of the three original speech signals used in the experiment at RT60 = 300 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding ones in Figs. 3.10 & 3.11 show how the post-filtering technique improves the output of the RLSFIDI beamformer.

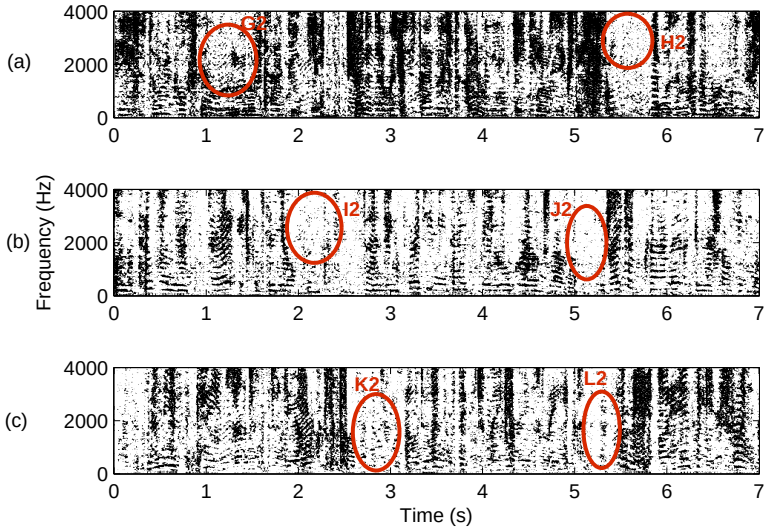


Figure 3.10. Binary masks of the speech signals separated by the RLS-FIDI beamformer at RT60 = 300 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding original speech signals in Fig. 3.9 show that a considerable amount of interference from the other sources still exists when the ΔSINR = 11.25 dB.

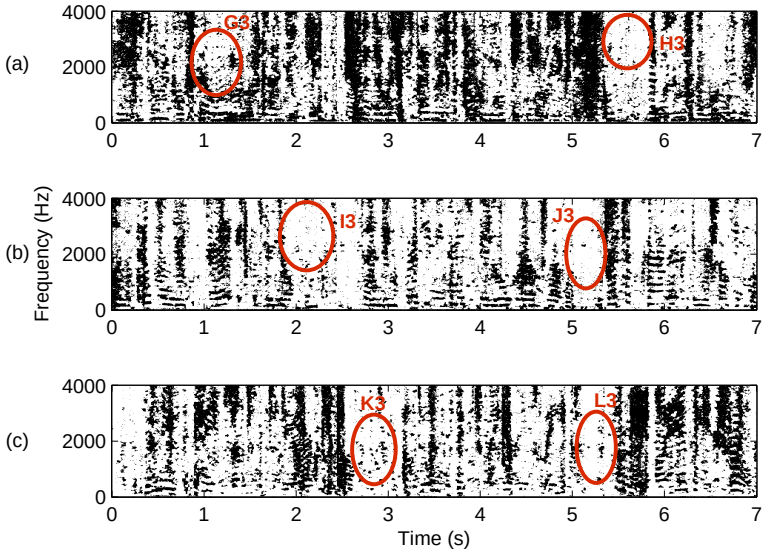


Figure 3.11. Binary masks of the three enhanced speech signals by the IBM TF masking technique at RT60 = 300 ms; (a) speaker 1, (b) speaker 2 and (c) speaker 3. The highlighted areas, compared with the corresponding ones in Figs. 3.9 & 3.10 show the post-filtering processing stage improves the output of the RLSFIDI beamformer. For these enhanced signals PEL = 24.82 %, PNR = 28.04 %, and ΔSINR = 12.18 dB.

Table 3.4. Final results: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer after post-processing for different reverberation. Results (in decibels) are provided for two different mixtures.

RT60 (ms)	Method	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
300	RLSFIDI beamformer	12.18	3.41	6.35	6.71	8.60	12.74	13.05	5.53	7.71	8.07
		12.36	5.22	5.80	6.50	13.00	11.47	16.54	5.05	7.48	7.05
450	RLSFIDI beamformer	8.86	-0.07	4.03	4.57	4.50	9.8	10.39	3.35	5.77	6.27
		9.76	4.86	3.85	4.09	10.19	9.62	15.32	2.88	5.64	4.56
600	RLSFIDI beamformer	7.59	-2.48	2.34	3.25	0.42	10.09	8.91	3.42	3.55	5.15
		7.91	-1.50	1.90	2.41	6.08	8.02	14.34	0.29	3.76	2.86

Table 3.5. Subjective evaluation: MOS for the RLSFIDI beamformer with and without post-processing and the IVA method, for different reverberation times.

RT60 (ms)	Mean opinion score (MOS [STD])		
	RLSFIDI beamformer	IVA Method	Proposed Method
300	3.9 [0.16]	3.5 [0.17]	4.0 [0.21]
450	3.3 [0.19]	3.1 [0.15]	3.7 [0.15]
600	3.1 [0.20]	2.9 [0.31]	3.3 [0.15]

3.5 Summary

In this chapter a beamforming based method combined with a post-processing scheme based on binary time-frequency masking for the separation of multiple sources in a reverberant environment was studied. Cepstral processing was also utilized to smooth the masks. The beamformer exploited the knowledge of the sound source locations (and thus the directions of arrival of the sources to the microphone array). A robust least squares frequency invariant data independent (RLSFIDI) beamformer was implemented with a linear array configuration. The performance of the RLSFIDI beamformer was further enhanced by applying a binary TF masking, or ideal binary masking (IBM) technique in the post-filtering stage. The proposed approach was shown to provide better separation than the IVA method.

Although the proposed beamforming method combined with the binary time-frequency masking achieves considerable separation improvement at low (and mildly medium) reverberation levels, the performance at high levels of reverberation is still limited. Further, this performance is achievable only with sixteen microphones in the array; reducing the number of sensors will generally deteriorate the separation performance. These limitations provide strong motivation to pursue new methods that require lesser number of sensors and are relatively more robust to reverberation.

Additionally, the time-frequency masking based post-processing in this chapter utilized binary or hard masks. A disadvantage of such masks is the introduction of musical noise due to estimation errors. To alleviate this problem, more flexible, soft or probabilistic masks need to be used.

To achieve the aforementioned objectives, in the proceeding chapter, a two-microphone based source separation method is proposed that generates soft time-frequency masks in order to separate sources from their acoustic mixtures. The method, inspired by human auditory processing, is based

on the probabilistic modeling of three cues, the interaural level difference (ILD), the interaural phase difference (IPD) and the mixing vectors. The sound source location information is also utilized within the modeling. The parameters for the models are estimated using EM. The algorithm generates probabilistic time-frequency masks that are used to isolate the individual sources.

Chapter 4

INFORMED MODEL-BASED SOURCE SEPARATION IN REAL REVERBERANT ROOMS

4.1 Introduction

This chapter describes an informed model-based source separation algorithm that utilizes observations from only two microphones. Given the reverberant mixtures, containing at least two sources, the interaural level difference (ILD), interaural phase difference (IPD), and the mixing vectors are modeled probabilistically. The sound source location estimates (assumed to be known, potentially obtained using information from video) are utilized in the probabilistic modeling. Direction vectors towards each source in the mixture are calculated using the source location estimates as described in Section 4.3.1. The direction vectors are used as the mean parameter of the mixing vector model. The source location estimates are also utilized in the overall algorithm initialization. The optimum parameters of the probabilistic models are estimated by the expectation-maximization (EM) algorithm as detailed in Section 4.4. The EM algorithm, after a fixed number of itera-

tions, generates soft time-frequency masks. The probabilistic time-frequency masks are applied to the reverberant mixtures to reconstruct the individual sources. As discussed earlier, it is assumed that the number of sources "I" and their locations are estimated through video processing and are known *a priori*. It is further assumed that the source signals are sparse and that they do not overlap in the time-frequency domain [82] [24] [38]. In this work and the remainder of the thesis two and three sources are considered. However, the method may work if the number of sources in the mixture further increase as the separation in the time-frequency space may still be possible but confirming this is left as future work. The sparsity assumption would weaken as the sources grow in number and thus force the method to fail.

4.2 The ILD, IPD, and Mixing vector models

Consider a stereo-recorded speech signal with the left and right sensor (ears or microphones) mixture signals denoted as $l(t_s)$ and $r(t_s)$. The mixtures are sampled with the sampling frequency f_a (sampling period $T_a = 1/f_a$) and hence are available at discrete time indices t_s for processing. The convolutive mixing model for the left and right sensors respectively, as shown in Fig. 4.1, can be written as $l(t_s) = \sum_{i=1}^{I} s_i(t_s) * h_{li}(t_s)$, and $r(t_s) = \sum_{i=1}^{I} s_i(t_s) * h_{ri}(t_s)$, where $s_i(t_s)$ denote the speech sources, $h_{li}(t_s)$ and $h_{ri}(t_s)$ are the impulse responses associated with the enclosure from source i to the left and right sensors respectively, and $*$ denotes the discrete time convolution operation. The time domain signals are then converted to the TF domain using the short-time Fourier transform (STFT). The interaural spectrogram is obtained by taking the ratio of the STFT of the left and right channels at each time frame t and frequency ω [58] as, $\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}$. Thus, the observed interaural spatial cues are $\alpha(\omega, t)$, the ILD, measured in dB, and $\phi(\omega, t)$, the IPD. Since the sources are assumed to be physically

stationary, the corresponding room impulse responses (RIRs) are assumed to be time invariant. Because of the phase wrapping, the IPD observations, $\angle(\frac{L(\omega,t)}{R(\omega,t)})$, are constrained to be in the range $[-\pi, \pi)$ and thus cannot be assigned to a source directly.

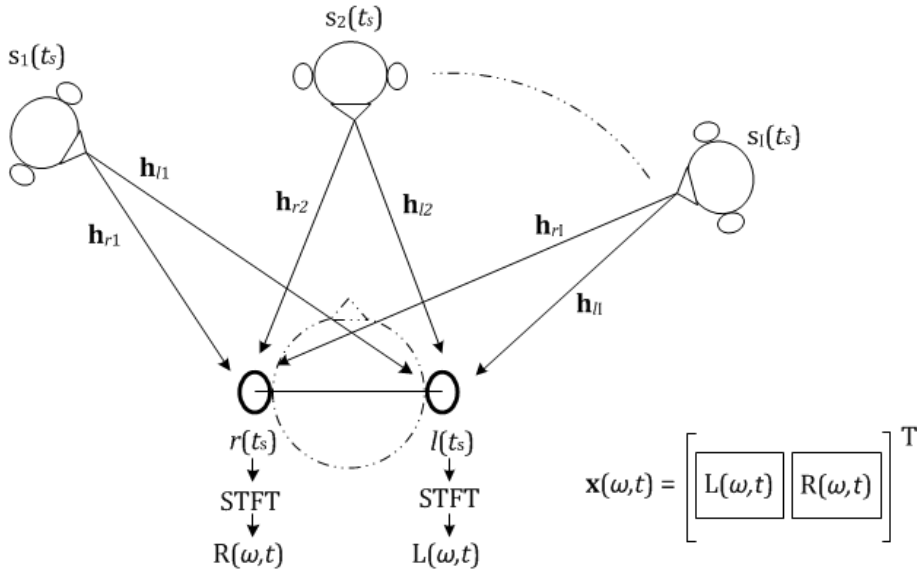


Figure 4.1. Signal notations. The left and right sensor convolutive mixtures are transformed to the TF-domain to obtain $\mathrm{L}(\omega, t)$ and $\mathrm{R}(\omega, t)$, and $\mathbf{x}(\omega, t)$ is formed by concatenating $\mathrm{L}(\omega, t)$ and $\mathrm{R}(\omega, t)$ as shown in the bottom righthand part of the image.

A source positioned at a certain location is modeled with a frequency-dependent interaural time difference (ITD) $\tau(\omega)$, and a frequency-dependent ILD as in [59]. The recorded IPD for each TF point, cannot always be mapped to the respective τ due to spatial aliasing. The model also requires that τ and the length of $h(t_s)$ should be smaller than the Fourier transform window. With the inter-microphone distance kept approximately the same as the distance between the two ears of an average-sized human head (around

0.17 m), the delay is much smaller than the Fourier analysis window of 1024 samples at a sampling frequency of 16 KHz (64 ms). Any portion of $h(t_s)$ over one window length is considered part of the noise. A top-down approach as described in [83] is thus adopted that makes it possible to map a τ to a recorded IPD at any desired group of frequencies. The phase residual error, the difference between the recorded IPD and the predicted IPD (by a delay of τ samples), in the interval $[-\pi, \pi)$ is defined as, $\hat{\phi}(\omega, t; \tau) = \angle(\frac{L(\omega,t)}{R(\omega,t)} e^{-j\omega\tau})$. The phase residual is modeled with a Gaussian distribution denoted as $p(\cdot)$ with mean $\xi(\omega)$ and variance $\sigma^2(\omega)$ that are dependent on frequency,

$$p(\phi(\omega, t)|\tau(\omega), \sigma^2(\omega)) = \mathcal{N}(\hat{\phi}(\omega, t; \tau)|\xi(\omega), \sigma^2(\omega)). \tag{4.2.1}$$

The ILD is also modeled with a Gaussian distribution with mean $\mu(\omega)$ and variance $\eta^2(\omega)$,

$$p(\alpha(\omega, t)|\mu(\omega), \eta^2(\omega)) = \mathcal{N}(\alpha(\omega, t)|\mu(\omega), \eta^2(\omega)). \tag{4.2.2}$$

The STFTs of the left and right channels are concatenated to form a new mixture $\mathbf{x}(\omega, t)$ as shown in Fig. 4.1. Assuming the W-disjoint orthogonality (WDO) property [38] of speech signals, the signals are sparse in the TF domain and only one source is dominant at each TF point, the STFT of the recordings $\mathbf{x}(\omega, t)$ at each time t and frequency ω can be written as [84],

$$\mathbf{x}(\omega, t) = \sum_{i=1}^{I} \mathbf{h}_i(\omega) s_i(\omega, t) \tag{4.2.3}$$

and approximated as

$$\mathbf{x}(\omega, t) \approx \mathbf{h}_d(\omega) s_d(\omega, t) \tag{4.2.4}$$

where $\mathbf{h}_d(\omega) = [h_{ld}(\omega), h_{rd}(\omega)]^T$ is the mixing vector from the dominant source $s_d(\omega, t)$ to the left and right sensor at that TF point, assumed to

be time invariant. The vector $\mathbf{x}(\omega, t)$ is normalized to have a unit norm to eliminate the effects of source scaling. The mixing vectors are modeled for each source with a Gaussian model as [84], [85]

$$p(\mathbf{x}(\omega,t)|\mathbf{d}_i(\omega),\varsigma_i^2(\omega)) = \\ \frac{1}{\pi\varsigma_i^2(\omega)}\exp\Big(-\frac{\|\mathbf{x}(\omega,t)-(\mathbf{d}_i^H(\omega)\mathbf{x}(\omega,t)).\mathbf{d}_i(\omega)\|^2}{\varsigma_i^2(\omega)}\Big) \tag{4.2.5}$$

where $\mathbf{d}_i(\omega)$ is the direction vector of the direct-path of the source signal i which will be derived using the source location estimates obtained from the video measurements, $\varsigma_i^2(\omega)$ is the variance of the model, $(\cdot)^H$ is the Hermitian transpose, and $\|\cdot\|$ indicates the Euclidean norm operator. In [85], [84], and [86] the authors proposed the use of an eigen decomposition of a sample covariance matrix to define unit norm vectors $\mathbf{d}_i(\omega)$ to represent the source directions in the probabilistic modeling of the mixing vectors. This approach, however, will be sensitive to estimation errors due to short data lengths, statistical non-stationarity in the audio scene and background noise. In contrast, in the proposed method the direction vectors are estimated through vision on the basis of a plane wave assumption, as discussed in Section 4.3.1 which thereby overcomes these shortcomings. Due to the comparatively accurate estimation of the mean parameter of the mixing vector model, and thus the improved posterior probability, the resulting TF masks for all sources that are found through the probabilistic modeling will then be enhanced, as explained in Section 4.4.2. The estimation of the parameter $\mathbf{d}_i(\omega)$ is described next.

4.3 Source Location Information and the Combined Models

As mentioned earlier, it is assumed that the source locations are known. These locations could potentially be estimated using the visual modality. Once the 3-D locations of the speakers are available, the mean parameter

$\mathbf{d}_i(\omega)$ is calculated as follows.

4.3.1 Parameter d_i Calculation

After estimating the 3-D position of each speaker i, the elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the coordinates of the center of the microphones, $p_{x_c}^{'}$, $p_{y_c}^{'}$ and $p_{z_c}^{'}$, are calculated as

$$\theta_i = \tan^{-1}\left(\frac{p_{y_i} - p_{y_c}^{'}}{p_{x_i} - p_{x_c}^{'}}\right) \tag{4.3.1}$$

and

$$\phi_i = \sin^{-1}\left(\frac{p_{y_i} - p_{y_c}^{'}}{r_i \sin(\theta_i)}\right) \tag{4.3.2}$$

where $r_i = \sqrt{(p_{x_i} - p_{x_c}^{'})^2 + (p_{y_i} - p_{y_c}^{'})^2 + (p_{z_i} - p_{z_c}^{'})^2}$. The direct-path weight vector $\mathbf{d}_i(\omega)$ for frequency bin ω and for source of interest $i = 1, ..., I$, can then be derived [72] as

$$\mathbf{d}_i(\omega) = \begin{bmatrix} \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).p_{x_1}^{'} + \sin(\theta_i). \\ \sin(\phi_i).p_{y_1}^{'} + \cos(\theta_i).p_{z_1}^{'})) \\ \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).p_{x_2}^{'} + \sin(\theta_i). \\ \sin(\phi_i).p_{y_2}^{'} + \cos(\theta_i).p_{z_2}^{'})) \end{bmatrix} \tag{4.3.3}$$

where $p_{x_j}^{'}$, $p_{y_j}^{'}$ and $p_{z_j}^{'}$ for $j = 1, 2$ are the 3-D positions of the sensors and $\kappa = \omega/c_s$ and c_s is the speed of sound in air at room temperature. The vector $\mathbf{d}_i(\omega)$ is normalized to unity length before it is used in the model.

4.3.2 Combining the Models

To obtain enhanced time-frequency masks for each static source the video-initialized IPD and ILD models, and the model for the mixing vectors that utilize the direct-path weight vector in Eq. (4.3.3) obtained with the aid of

video are used in conjunction. Since the sources are differently distributed in the mixture spectrograms, in terms of their IPD, ILD and their mixing, the parameters of the above models cannot be obtained directly from those mixtures. It is a hidden maximum-likelihood parameter estimation problem and thus the expectation-maximization algorithm is employed for its solution. Considering the models to be conditionally independent, they are combined given their corresponding parameters as

$$\begin{aligned} p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)|\widetilde{\Theta}) = \mathcal{N}(\alpha(\omega,t)|\mu(\omega),\eta^2(\omega)) \\ . \mathcal{N}(\hat{\phi}(\omega,t)|\xi(\omega),\sigma^2(\omega)) \\ . \mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{d}(\omega),\varsigma^2(\omega)) \end{aligned} \tag{4.3.4}$$

where $\widetilde{\Theta}$ denotes all of the model parameters. It is emphasized that it is only the noise in the measurements of ILD and IPD that is assumed to be conditionally independent and this same assumption is adopted as in [59] for the measurement related to the source direction vector. However, the conditional independence assumption offers particular advantage in algorithm development; namely, at each iteration of the EM algorithm, the parameters can be updated separately. As in [59], the dependence between ILD and IPD is introduced through prior assumptions on the mean values of the model parameters. Since the ILD and IPD may have dependence on source direction, the assumption of the conditional independence amongst the noise components may only be an approximation. Modeling such dependence is beyond the scope of this study, but is an interesting point for further investigation.

4.4 Model Parameters and Expectation-Maximization

4.4.1 Model Parameters

All of the model parameters $\widetilde{\Theta}$ can be collected as a parameter vector

$$\widetilde{\Theta} = \{\mu_i(\omega), \eta_i^2(\omega), \xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega), \mathbf{d}_i(\omega), \varsigma_i^2(\omega), \psi_{i\tau}\} \tag{4.4.1}$$

where μ_i, $\xi_{i\tau}$, and $\mathbf{d}_i$ and η_i^2, $\sigma_{i\tau}^2$, and ς_i^2 are respectively the means and variances of the ILD, IPD, and mixing vector models. The subscript i indicates that the parameters belong to the source i, and τ and ω show the dependency on delay and frequency. The parameter $\mathbf{d}_i(\omega)$ is included since it is used within the EM algorithm but highlight that since it is obtained from the video it remains constant throughout the algorithm. The parameter $\psi_{i\tau}$ is the mixing weight, i.e. the estimate of the probability of any TF point belonging to source i at a delay τ. Note that $\psi_{i\tau}$ is obtained from the hidden variable $z_{i\tau}(\omega, t)$ that qualifies the assignment of a TF unit to source i for the delay τ [59]. The hidden variable is an important variable and is unity if the TF point belongs to both source i and delay τ and zero otherwise. In more detail, the probability of $z_{i\tau}(\omega, t)$ is equivalent to $\psi_{i\tau}$ which is the estimate of the joint probability of a TF point being from source i at a delay τ. Since discrete values of τ are pre-defined, $\psi_{i\tau}$ is a two-dimensional matrix of the probability of being in each discrete state. $z_{i\tau}(\omega, t)$ is not explicitly calculated. The parameter $\psi_{i\tau}$ is computed in the expectation step of the EM algorithm. $\psi_{i\tau}$ is estimated by placing a Gaussian with its mean at each cross-correlation peak and a standard deviation of one sample [59].

The log value of the likelihood function ($\mathcal{L}$) given the observations can

be written as

$$\begin{aligned}\mathcal{L}(\widetilde{\Theta}) &= \sum_{\omega,t} \log p(\alpha(\omega,t), \phi(\omega,t), \mathbf{x}(\omega,t)|\widetilde{\Theta}) \\ &= \sum_{\omega,t} \log \sum_{i,\tau} [\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)) \\ &\qquad . \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega)) \\ &\qquad . \mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{d}_i(\omega), \varsigma_i^2(\omega)).\ \psi_{i\tau}\ ]\end{aligned} \tag{4.4.2}$$

and the maximum likelihood solution is the parameter vector which maximizes this quantity.

4.4.2 The Expectation-Maximization Algorithm

The algorithm is initialized using the estimated locations of the speakers provided by video. In the expectation step (E-step) the posterior probabilities are calculated given the observations and the estimates of the parameters as

$$\begin{aligned}\epsilon_{i\tau}(\omega,t) = \psi_{i\tau}\ &. \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)) \\ &. \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega)) \\ &. \mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{d}_i(\omega), \varsigma_i^2(\omega))\end{aligned} \tag{4.4.3}$$

where $\epsilon_{i\tau}(\omega,t)$ is the expectation of the hidden variable. In the maximization step (M-step), the parameters are updated using the observations and $\epsilon_{i\tau}(\omega,t)$ from the E-step. The IPD and ILD parameters and $\psi_{i\tau}$ are re-estimated as in [59]. The mean parameter of the mixing vectors $\mathbf{d}_i(\omega)$ is obtained through video as discussed in Section 4.3.1 and $\varsigma_i^2(\omega)$ is updated as [84]

$$\varsigma_i^2(\omega) = \frac{\sum_{t,\tau} \epsilon_{i\tau}(\omega,t).\|\mathbf{x}(\omega,t) - (\mathbf{d}_i^H(\omega)\ \mathbf{x}(\omega,t)).\mathbf{d}_i(\omega)\|^2}{\sum_{t,\tau} \epsilon_{i\tau}(\omega,t)}. \tag{4.4.4}$$

The mixing vector model starts contributing from the second iteration, as in the first iteration the occupation likelihood $\epsilon_{i\tau}(\omega, t)$ is calculated using only the ILD and IPD models. The initial value of $\varsigma_i^2(\omega)$ is computed after the first iteration using $\epsilon_{i\tau}(\omega, t)$. Since the algorithm is initialized with source locations estimates from video and $\epsilon_{i\tau}(\omega, t)$ contains the correct order of the sources the permutation problem is bypassed. The probabilistic masks for each source can be formed as $M_i(\omega, t) \equiv \sum_\tau \epsilon_{i\tau}(\omega, t)$. The time domain source estimates are obtained by applying the TF masks to the mixtures and taking the inverse STFT. The efficacy of the proposed approach is experimentally verified in the next section. A brief summary of the proposed scheme is given in Algorithm 1.

Algorithm 1 Brief summary of the proposed audio-visual source separation approach

Input: Synchronized audio-visual measurements
Output: Separated speech sources

1: Obtain the speaker locations when the sources are judged physically stationary
2: Calculate parameter $\mathbf{d}_i$ as in Section 4.3.1
3: Initialize the EM algorithm in Section 4.4.2 with speaker locations and PHAT-Histogram
4: Run the EM algorithm as in Section 4.4.2 to generate time-frequency masks for all sources
5: Apply the time-frequency masks to the mixtures to reconstruct the sources

4.5 Experimental Evaluation in a Room Environment

The performance of the proposed algorithm is evaluated in two main sets of experiments and is compared with five other algorithms, two are audio-only and three are audio-visual. Firstly, mixtures of two sources are simulated with varying reverberation times (RT60s) using synthetic room impulse responses (RIRs), different model complexities and separation angles, and

three sources with varying separation angles utilizing real RIRs. Comparisons are provided in all of the above scenarios with two other state-of-the-art audio-only algorithms to highlight the advantage of the audio-visual approach to source separation. Secondly, experiments are performed for varying RT60s for both two and three source mixtures and the proposed method is compared with three other state-of-the-art audio-visual algorithms.

4.5.1 Common Experimental Settings

Room Layout

The room setting is shown in Fig. 4.2. Experiments were performed for mixtures of both two and three speech sources. The desired source was located in front of the sensors at 0° azimuth and the interferer was positioned at one of the six different azimuths between 15° and 90° i.e. $[15°, 30°, 45°, 60°, 75°, 90°]$ for the case of two speakers. In the three-speaker case the third source was located symmetrically with the same azimuth, as shown for approximately 60° in Fig. 4.2.

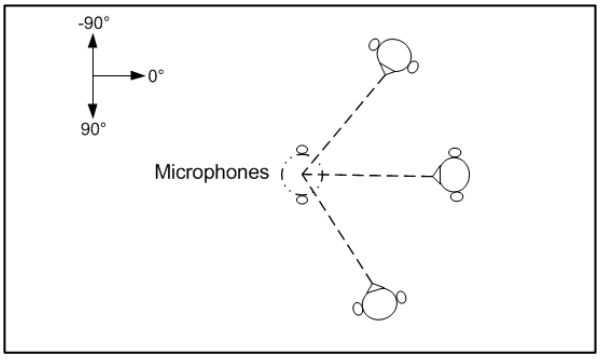


Figure 4.2. The room layout showing one of the approximate positions of the sources and the sensors.

Table 4.1. Different Parameters Used In Experiments

STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (real) or 160-600 ms (image method)
Room dimensions	[9 5 3.5] m
Source signal duration	2.5 s (TIMIT)
Sensor spacing	0.17 m

Speech Data and Room Impulse Responses

Speech signals from the TIMIT acoustic-phonetic continuous speech corpus [23] were used. Utterances were randomly chosen to form mixtures with different combinations i.e. male-male, male-female, and female-female. The first (16k×2.5) samples of the TIMIT speech sources were used and were normalized to unity variance before convolving with the RIRs. The real RIRs were used from [4] which were measured in a real classroom with an RT60 of approximately 565 ms. The center location was used in the experiments with the sensor-to-speaker distance of 1 m. The image method [3] was also used to evaluate the proposed algorithm for varying RT60s.

Evaluation of Separation Performance

The signal-to-distortion ratio (SDR) as in [64] was used to evaluate the performance of the algorithm in cases where the original speech sources were available. SDR is the ratio of the energy of the original signal to the energy from interferers, other noise energy and artifacts.

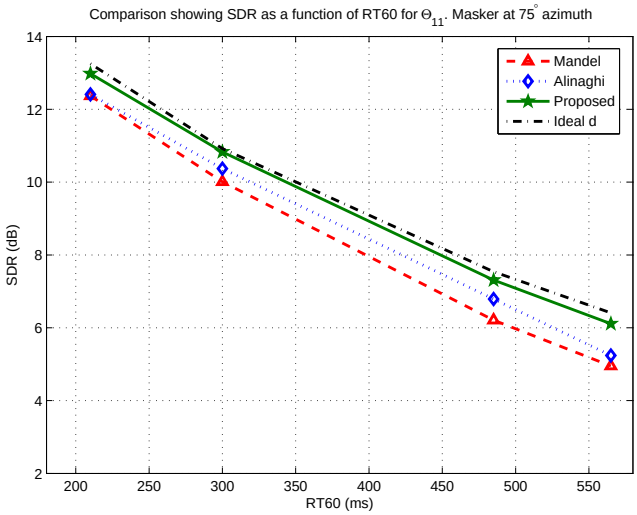
4.5.2 Results and Comparison With Other Audio-Only Algorithms

Extensive experiments were conducted to test the robustness and consistency of the proposed algorithm. The common parameters used in all experiments are given in Table 4.1. As mentioned earlier, to emphasize the advantage of the multimodal approach over audio-only methods in realistic multi-speaker

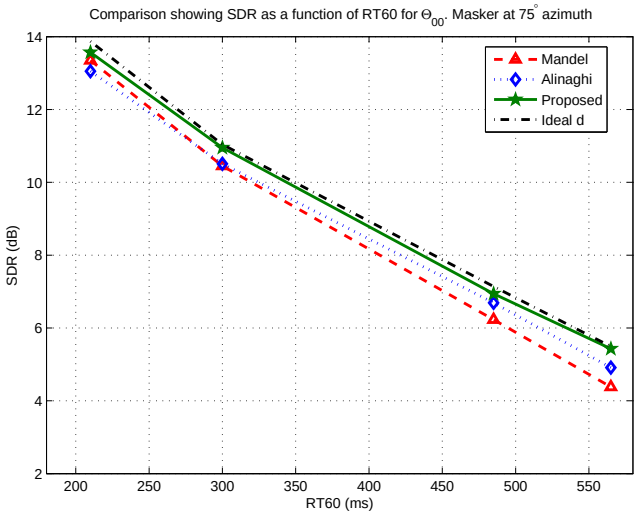
environments the results are compared with [59], referred to as Mandel, and [86], termed as Alinaghi.

Different model complexities, for ILD and IPD, were evaluated similar to [59]. For instance, the ILD and IPD model complexity of Θ_{00} will have no ILD contribution and an IPD model with zero mean and a standard deviation that varies only by source, whereas Θ_{11} will have a frequency-independent ILD model and an IPD model with a frequency-independent mean and a standard deviation that varies by source and τ, while $\Theta_{\Omega\Omega}$ uses the full frequency-dependent ILD and IPD model parameters. And $\Theta^{G}_{\Omega\Omega}$ has parameters similar to $\Theta_{\Omega\Omega}$ but includes a garbage source and an ILD prior as described in [59].

In Fig. 4.3, the two model complexities Θ_{11} and Θ_{00} for two sources were simulated with an interferer at 75°. The speech files from the TIMIT dataset were convolved with the RIRs generated using the image method [3] to obtain the reverberant mixtures. The RT60 was varied to evaluate performance of the algorithms at different levels of reverberation. A curve that corresponds to the model which uses the ideal $\mathbf{d}_i$ vector found from the known source locations has also been included in the results. The curve provides an upper bound for performance improvement for the algorithm. The results indicate the improved performance of the proposed technique over [59] and [86]. In Fig. 4.3(a), for RT60 of 210 ms the proposed algorithm gives an output of 12.98 dB, Mandel's algorithm gives 12.37 dB and Alinaghi 12.41 dB. As the RT60 increases the proposed algorithm still performs best, for example at 565 ms it is 6.11 dB, which is 1.16 dB higher than Mandel and 0.87 dB higher than the method by Alinaghi. In Fig. 4.3(b), with a simpler model Θ_{00}, at an RT60 of 210 ms the proposed method outputs 13.57 dB, compared to Mandel, 13.35 dB, and Alinagi, 13.05 dB. At the maximum RT60 of 565 ms the proposed algorithm gives an output of 5.43 dB, 1.05 dB higher than Mandel and 0.52 dB higher than Alinaghi. The ILD cues fade away with



(a)



(b)

Figure 4.3. Comparison of performance at different RT60s. The interferer was located at 75° azimuth. Synthetic RIRs using [3] were used to simulate varying RT60s. The Θ_{11} (a) and Θ_{00} (b) modes are under consideration.

increasing reverberation and thus the direct-path direction vector obtained by video information in the proposed algorithm contributes to better model the mixing vectors and improve the separation performance.

In Fig. 4.4 (a) the proposed algorithm was evaluated for all the model complexities. Real RIRs from [4] were utilized to form acoustic mixtures in

this set of experiments. The results indicate that the proposed algorithm's performance is consistently best for all models. In [86] the authors reported that their algorithm showed significant improvement over [59] with simpler models but the improvement diminished with the increasing model complexity as confirmed in Fig. 4.4 (a), specifically when the ILD model started contributing. In contrast, the performance of the proposed algorithm is clearly shown not to deteriorate with increasing complexity and shows consistent improvement over all the models. The average improvement across the models in the Alinaghi method over the Mandel method is 1.53 dB, whereas for the proposed method is 2.39 dB. In Fig. 4.4 (b) the SDR as a function of the separation angle between the speakers for the Θ_{11} model is shown. Comparatively, over all angles the proposed algorithm that utilizes the estimate of the source direct-path direction vector, by exploiting visual information, yields an average improvement of 1.53 dB whereas Alinaghi's method gives 0.75 dB. Results in Fig. 4.5 show SDR as a function of separation angle i.e. between 15° and 90° for mixtures of three speakers with the most complex frequency-dependent mode $\Theta_{\Omega\Omega}$ using real RIRs. The two interferers on either side of the target were positioned symmetrically with the same azimuth. The interferer to the left was simulated by reversing the order of the sensors. At the minimum separation angle of 15° the proposed algorithm gives an output of 2.16 dB, whereas Mandel, 0.9 dB, and Alinaghi, 1.43 dB. The results indicate that the method in [86] offers improvement over [59] at smaller separation angles from 15° to 45° but no significant improvement at larger separation angles. The proposed algorithm, in contrast, shows consistent improvement over all separation angles, specifically in the difficult scenario with smaller separation angles, over both [59] and [86] in the three-speaker reverberant case confirming the suitability of the audio-visual approach in multi-speaker realistic settings, and the value of adding visual information in audio source separation.

4.5.3 Results and Comparison with Other Audio-Visual Methods

The proposed approach is next compared with three other audio-visual algorithms, the beamforming based method in [87] which is referred to as Naqvi, the technique in [88], which is termed as Maganti and the scheme in [89] using robust beamforming, referred to as RLSFIDI. Similar to the proposed work, these audio-visual methods employ the visual modality to estimate the speaker locations which are then utilized within the algorithms.

The multimodal approach to BSS [87] uses the visual modality to enhance the separation of both static and moving sources. The speaker positions estimated by a 3-D tracker are used to initialize the frequency domain BSS algorithm for the physically stationary speakers and beamforming if the speakers are moving. The algorithm's performance is reasonable at low reverberation when the direct path signal is strong but deteriorates at higher RT60s when the direct-to-reverberant ratio (DRR) is low. The beamformer is also generally limited to the determined and overdetermined cases and achieves improved performance with larger number of audio sensors.

In [88] an audio-video multispeaker tracker is proposed to localize sources and then separate them using microphone array beamforming. A postfiltering stage is then applied after the beamforming to further enhance the separation. The overall objective of the system is automatic speech recognition which lies outside the scope of the proposed work, thus, the output of the speech enhancement part is compared.

In [89] a robust least squares frequency invariant data independent beamformer is implemented. The MCMC-PF based tracker estimates the direction of arrival of the sources using visual images obtained from at least two cameras. The robust beamformer, given the spatial knowledge of the speakers, uses a convex optimization approach to provide a precise beam for the desired source. To control the sensitivity of the beamformer a white noise constraint is used. The scheme provides significant improvement at lower

RT60s but the performance degrades as reverberation increases. The original code used in [89] is employed in the comparison.

In contrast, in [90] a speech source is separated by utilizing its coherence with the speaker's lip movements. Parameters describing a speaker's lip shape are extracted using a face processing system. The authors provide results for separation of simple vowel-plosive combinations from other meaningful utterances and acknowledge that separating complex mixtures would be increasingly difficult. In the extension of their work in [91], the spectral content of the sound that is linked with coherent lip movements is exploited and assessment is provided on two audio-visual corpora, one having vowel-plosive utterances similar to their previous work and the second containing meaningful speech spoken by a French speaker. They discuss the determined case and the underdetermined case with two sensors and three sources but reported that performance was limited as the phonetic complexity increased. These works, as in [92, 93], require the speakers to be right in front of the camera(s), with the face clearly visible so that facial cues can be observed. The proposed approach is more general, in that only head localization information is required and therefore audio-visual recordings with low resolution can be processed. Hence the methods in [90–93] are not included in the comparison.

Results

The experimental results in Fig. 4.6 provide the average SDR (dB) as a function of RT60 for ten random mixtures of two sources for the proposed method and the three other audio-visual methods i.e. Naqvi, Maganti, and RLSFIDI. The masker was positioned at -15 degrees azimuth i.e. the minimum and most challenging separation angle in the earlier simulations. The other algorithms were each evaluated with two, four and eight microphones at all RT60s. The proposed algorithm gives better separation, using only

two microphones, than all the other algorithms at all RT60s except at 160 ms where the RLSFIDI outperforms the proposed method with four microphones. The Naqvi and Maganti methods adopt the general trend by improving the separation as the number of microphones is increased, since the increased number of filter coefficients provides better interference removal. The postfiltering stage in Maganti's scheme refines the output further from its previous beamforming stage by exploiting sparsity of the speech sources. Masking postfilters are obtained by retaining the maximum filter output values at each frequency bin. The final postfilter is then applied to the beamformer output. This scheme considerably improves the performance over that of Naqvi for all number of microphones and all RT60s in terms of the SDR, but introduces musical noise which was observed when the reconstructed source was listened to. In the RLSFIDI method the designed unmixing filters used are frequency invariant and data independent thus the source statistics and RT60 are not considered. Also, since the physical separation between the sources is only 15°, the increased spatial selectivity of the RLSFIDI design appears to deteriorate the separation performance at higher RT60s. In summary, the RLSFIDI method with eight microphones has the best performance among the three competing techniques below RT60 of around 450 ms and Maganti with eight microphones above 450 ms.

The results in Fig. 4.7 show the average SDR (dB) as a function of RT60 for ten random mixtures for the proposed method and the three other audio-visual methods when separating three sources. Each of these three algorithms was run by using four and eight microphones. Having three sources in the mixture, the case of only two-microphones becomes underdetermined and solution is not possible through the beamformers in Naqvi, Maganti, and RLSFIDI, unlike the proposed algorithm which can handle the underdetermined case too. The improved spatial selectivity of the RLSFIDI design again explains this advantage but this degrades with increasing

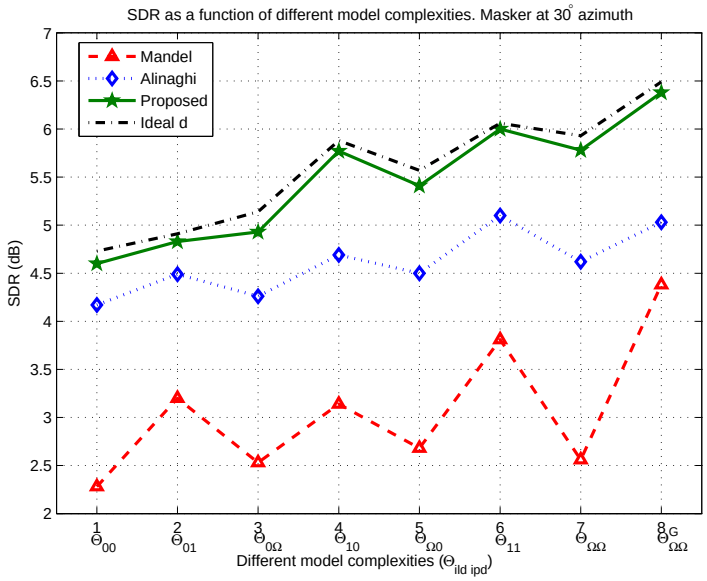
RT60. All the algorithms follow this general trend of degraded performance with increased RT60. For 160 ms, 210 ms and 300 ms utilizing the eight microphones mixture RLSFIDI performs best. This is the strength of the RLSFIDI method that at lower RT60s, with reduced reflections, and hence fewer reflections from the interfering source and overall reverberation leak through the precise beam formed for the desired source, the separation performance is greatly enhanced. This behaviour changes as the RT60 increases beyond 300 ms, where even increasing the number of microphones does not stop the deterioration in the separation performance of the beamformer. In Fig. 4.8, as an example, the beam patterns for the RLSFIDI beamformer are provided using four and eight microphones for the case of three sources. The sources are positioned at $-45°$, $0°$, and $45°$. The beam towards the desired source becomes more precise as the number of microphones is increased. Note, that for Fig. 4.6 the masker is at $-15°$ which explains why separating three sources can be better with beamforming.

4.6 Summary

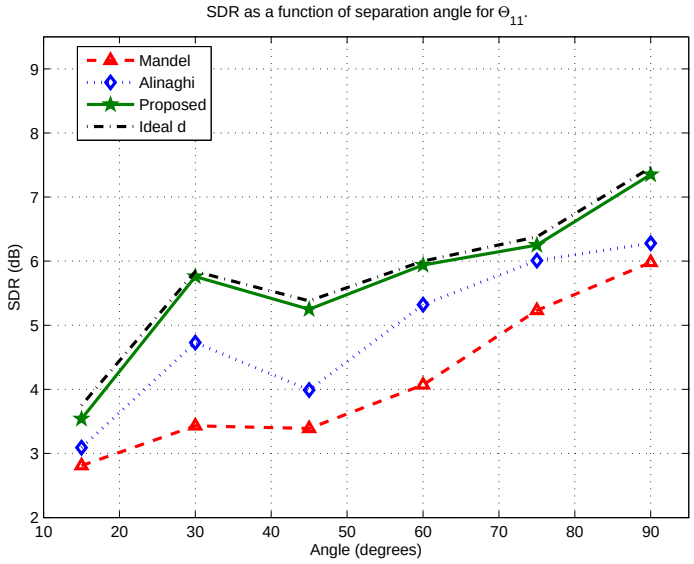
This chapter explained a source separation algorithm that utilizes visual contribution in terms of the source location estimates. By utilizing this visual information, it has been confirmed that more accurate TF masks can be obtained which give improved source estimates, particularly in highly reverberant multi-speaker environments. The proposed algorithm has been experimentally tested in a variety of settings including real room impulse responses confirming its robustness over two other audio-only methods and three similar audio-visual algorithms in both the two-speaker and three-speaker cases.

Two further questions remain: can additional cues associated with the spatial properties of the sources and the enclosure enhance the separation performance, specifically when the level of reverberation is high? Can the knowledge of the properties of the room, alongside knowing the source locations, such as its total wall area, reflective characteristics of the wall surfaces, and the reverberation time be used to achieve additional advantage in highly reverberant scenarios?

To address these questions, in the following chapter, the spatial covariance model, a model that utilizes the knowledge of the spatial properties of the sources and the room is investigated. The model is evaluated when used in conjunction with the ILD and IPD models, and also when used in combination with the ILD, IPD, and mixing vector models.



(a)



(b)

Figure 4.4. In (a) the performance at different model complexities $\Theta_{ild\ ipd}$ for two sources with the interferer at 30° azimuth is shown. The graph in (b) indicates results at different separation angles for model Θ_{11}. The position of the interferer was varied in steps of 15° between 15° to 90°. Real binaural RIRs from [4] were used. Results were averaged over five random mixtures. The proposed method yields a considerable improvement at all modes and separation angles.

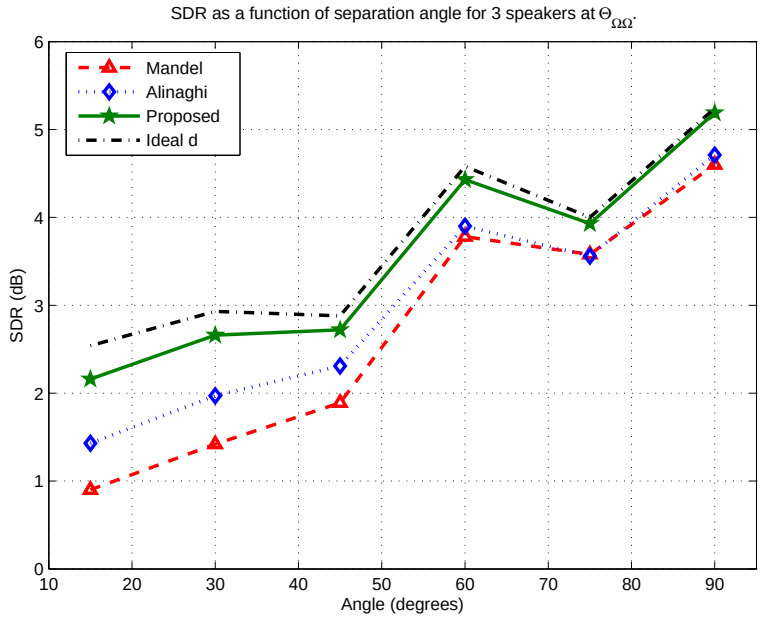


Figure 4.5. Results of the three-speaker case at different separation angles using the real RIRs at the $\Theta_{\Omega\Omega}$ mode. The interferers were located symmetrically to both sides of the target source. Results indicate that our proposed method performs best at all separation angles.

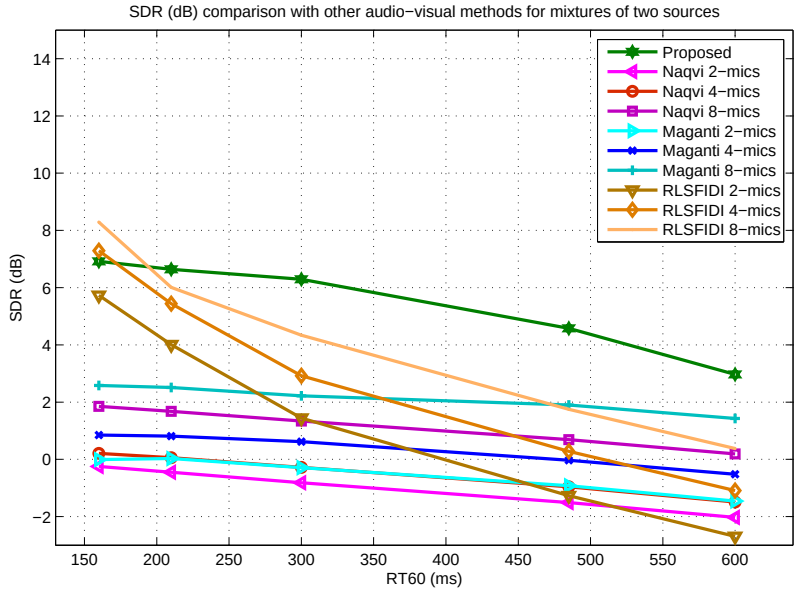


Figure 4.6. Comparison of SDR (in decibels) performance as a function of RT60 using the proposed algorithm utilizing two microphones and the Naqvi, Maganti and RLSFIDI methods employing two, four and eight microphones for mixtures of two sources.

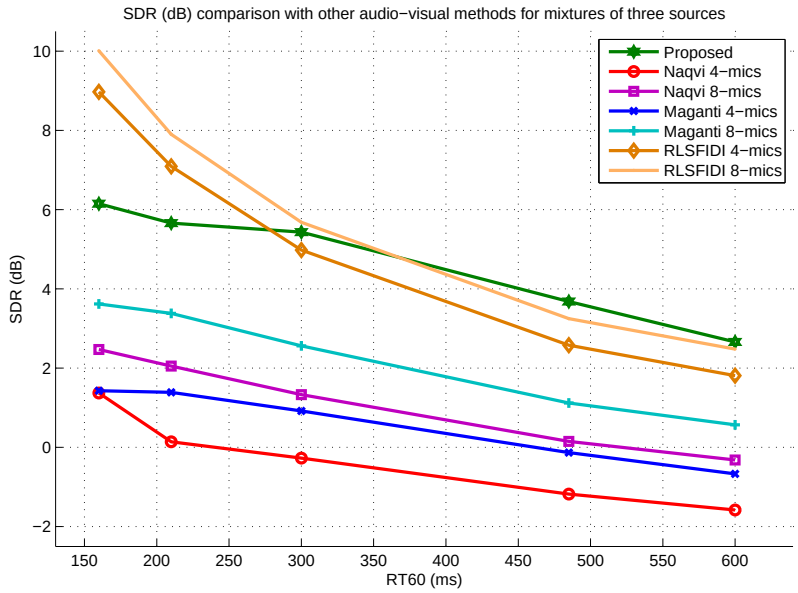
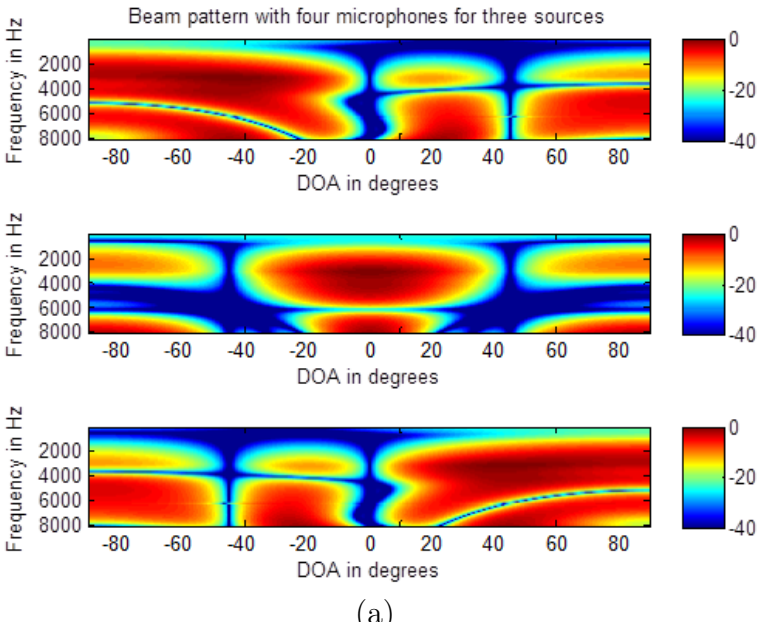
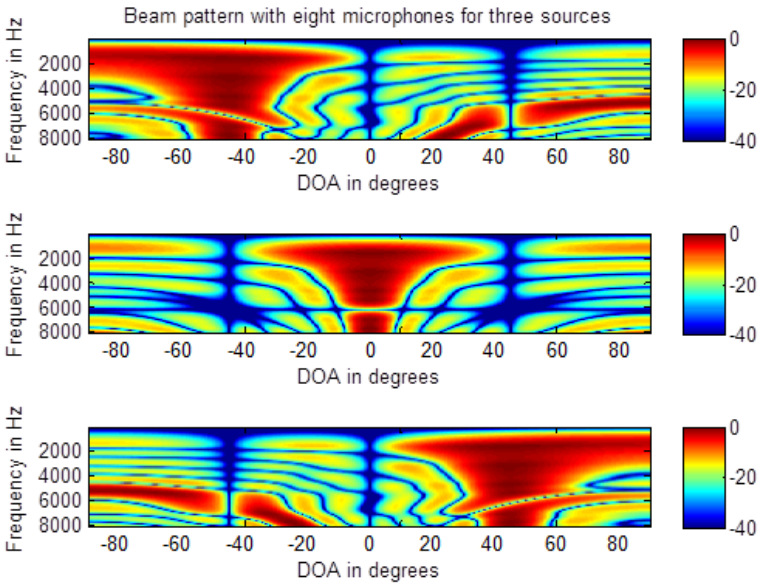


Figure 4.7. Comparison of SDR (in decibels) performance as a function of RT60 using the proposed algorithm utilizing two microphones and the Naqvi, Maganti and RLSFIDI methods employing four and eight microphones for mixtures of three sources.



(a)



(b)

Figure 4.8. Beam patterns achieved by the RLSFIDI beamformer with four microphones in (a) and eight microphones in (b) for the case of three sources. It is clearly visible that as the number of sensors is increased the beam for the desired source becomes more precise strictly allowing the desired source and forming a null towards the interferer. With fewer microphones the interferers and reverberation leak through with the desired source degrading the separation performance.

Chapter 5

INFORMED SPATIAL COVARIANCE MODEL: MODELING SPATIAL PROPERTIES OF THE SOURCES AND THE ROOM

5.1 Introduction

This chapter investigates modeling the spatial characteristics of the sound sources and the enclosure to mitigate the degradation caused by the high level of reverberation. It aims to model the contribution of individual sources to both the mixture channels (left and right microphones) with a zero-mean Gaussian distribution. The covariance of the distribution is modeled by exploiting the location information of the sources, the reflective attributes of the wall surfaces, the area and the reverberation time of the room. The model operates in the time-frequency (short-time Fourier transform) domain and is fused with models of the interaural cues discussed in Chapter 4 to further the separation performance, specifically in the cases when the room is highly reverberant.

Consider, for instance, the case when there are speakers in a meeting or teleconference room and the enclosure is highly reverberant; performance of the current source separation methods in such environments is very limited. Can additional cues assist in improving the separation performance in such acoustically hostile environments? This chapter thus addresses this question by incorporating the spatial covariance model to the ILD and IPD models discussed in Chapter 4. The spatial covariance model utilizes the knowledge of the locations of the speakers and properties of the room, which are assumed to be known as before.

Similar to Chapter 4, the optimal parameters of the combined models are estimated in a maximum-likelihood sense through the expectation-maximization (EM) algorithm. The estimation of a parameter in the spatial covariance model makes use of the known speaker locations. In the E-step, the posterior probabilities are calculated whereby TF points are assigned to sources using the observations and the initial values of the parameters. In the M-step, the parameters of the models are updated based on the measurements and the probabilities from the E-step. The combined algorithm generates TF masks that are used to separate the individual sources.

5.2 The Spatial Covariance Model

Given the two-channel reverberant mixtures, $l(t_s)$ and $r(t_s)$, a new signal $\mathbf{x}(t_s)$ is formed by concatenating them. The contribution of "I" sources to both the left and right channels can also be represented as [94]

$$\mathbf{x}(t_s) = \sum_{i=1}^{I} \mathbf{img}_i(t_s) \tag{5.2.1}$$

where $\mathbf{img}_i(t_s) = [img_{li}(t_s), img_{ri}(t_s)]^T$ is the spatial image of the i^{th} source to the left and right channels. Assuming the sources are uncorrelated, $\mathbf{x}(\omega, t)$, the short-time Fourier transform (STFT) of $\mathbf{x}(t_s)$, is modeled as a zero-mean

Gaussian distribution with the covariance matrix [94]

$$\mathbf{R}_{\mathbf{x}}(\omega, t) = \sum_{i=1}^{I} v_i(\omega, t)\mathbf{R}_i(\omega), \tag{5.2.2}$$

where $v_i(\omega, t)$ is the time-varying scalar variance and $\mathbf{R}_i(\omega)$ is the time-invariant covariance matrix utilizing the spatial properties of the source i and the enclosure. From results based on statistical room acoustics [95], it is assumed that the impulse response is the sum of the direct path and the diffuse part. The diffuse propagation of sound is due to reverberation. Reverberation increases the spatial spreading of the source due to multiple reflections with wall surfaces and other objects in the room. The spatial covariance of the source i, $\mathbf{R}_i(\omega)$, is thus estimated as the sum of the direct path direction vector and the covariance matrix of the reverberant part [94] [95]

$$\mathbf{R}_i(\omega) = \mathbf{d}_i(\omega)\mathbf{d}_i^H(\omega) + \sigma_{rev}^2 \begin{bmatrix} 1 & \Omega(d_{lr}, \omega) \\ \Omega(d_{lr}, \omega) & 1 \end{bmatrix} \tag{5.2.3}$$

where $\mathbf{d}_i(\omega)$ is the direct-path direction vector, σ_{rev}^2 is the variance of the reverberant part and $\Omega(d_{lr}, \omega)$ depends on the distance between left and right sensors d_{lr} and the frequency ω. The intensity of the reverberation observed at both the microphones is assumed to have diffuse characteristics with the same power,

$$\Omega(d_{lr}, \omega) = \frac{\sin(2\pi\omega d_{lr}/c)}{2\pi\omega d_{lr}/c} \tag{5.2.4}$$

where c is the speed of sound in air at room temperature. The variance of the reverberant part is given by [94]

$$\sigma_{rev}^2 = \frac{4\beta^2}{A(1-\beta^2)}, \tag{5.2.5}$$

where A is the total wall area and β is the wall reflection coefficient estimated from the room reverberation time (RT60), assumed to be known *a priori*, using Eyring's formula [95] as

$$\beta = \exp\big(-\frac{13.82}{L_{xyz}cRT60}\big) \tag{5.2.6}$$

where L_{xyz} is computed using the x, y and z dimensions of the rectangular room as, $L_{xyz} = (\frac{1}{L_x} + \frac{1}{L_y} + \frac{1}{L_z})$.

To estimate the scalar variance $v_i(\omega, t)$ for the computation of the covariance matrix $\mathbf{R}_{\mathbf{x}}(\omega, t)$ in Eq. 5.2.2, the method in [96] is followed where it is obtained by minimizing the sum over all TF units (ω, t) of the Kullback-Leibler (KL) divergence between the theoretical covariance matrix $\mathbf{R}_{\mathbf{x}}(\omega, t)$ and the covariance matrix of the observed mixture $\mathbf{R}_{\mathbf{x}}^{obs}(\omega, t)$. The variance for each source is then given as, assuming only a single source is active at each TF point,

$$v_i(\omega, t) = \frac{1}{2}\ \mathrm{tr}[\mathbf{R}_i^{-1}(\omega)\mathbf{R}_{\mathbf{x}}^{obs}(\omega, t)] \tag{5.2.7}$$

where tr[.] is the trace operator. The covariance matrix of the observed mixture is calculated as [96]

$$\mathbf{R}_{\mathbf{x}}^{obs}(\omega, t) = \frac{\Sigma_{\omega' t'}\Delta(\omega' - \omega, t' - t)\mathbf{x}(\omega', t')\mathbf{x}^H(\omega', t')}{\Sigma_{\omega' t'}\Delta(\omega' - \omega, t' - t)} \tag{5.2.8}$$

where Δ is a two-dimensional window describing the weighting in the neighbourhood of the TF point under consideration.

The probability distribution of the model is given as [97]

$$\begin{aligned} P(\mathbf{x}(\omega, t)|\{v_i(\omega, t), \mathbf{R}_i(\omega), \forall\ i\}) = \\ \frac{1}{\det(\pi\mathbf{R}_{\mathbf{x}}(\omega, t))}\exp(-\mathbf{x}^H(\omega, t)\mathbf{R}_{\mathbf{x}}^{-1}(\omega, t)\mathbf{x}(\omega, t)) \end{aligned} \tag{5.2.9}$$

where $(\cdot)^H$ is the Hermitian transpose, and the mean is assumed zero.

To accomplish the calculation of the spatial covariance matrix in Eq. (5.2.3), the direct-path direction vector is estimated using the spatial locations of the sources as

$$\mathbf{d}_i(\omega) = [h_{li}, h_{ri}]^T \tag{5.2.10}$$

where

$$h_{li} = \exp(\frac{-j\omega}{c}(\sin(\theta_i).\cos(\phi_i).p^{'}_{x_l} + \sin(\theta_i).\sin(\phi_i).p^{'}_{y_l} + \cos(\theta_i).p^{'}_{z_l})) \tag{5.2.11}$$

and

$$h_{ri} = \exp(\frac{-j\omega}{c}(\sin(\theta_i).\cos(\phi_i).p^{'}_{x_r} + \sin(\theta_i).\sin(\phi_i).p^{'}_{y_r} + \cos(\theta_i).p^{'}_{z_r})). \tag{5.2.12}$$

Here $[p_{x_i}, p_{y_i}, p_{z_i}]$ is the location estimate of speaker i, $p^{'}_{x_m}$, $p^{'}_{y_m}$ and $p^{'}_{z_m}$, wherein m is the left or right sensor index, are the 3-D positions of the sensors and c is the speed of sound in air at room temperature. The elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the center of the microphones of each speaker i are computed respectively as $\theta_i = \tan^{-1}\left(\frac{p_{y_i} - p^{'}_{y_c}}{p_{x_i} - p^{'}_{x_c}}\right)$ and $\phi_i = \sin^{-1}\left(\frac{p_{y_i} - p^{'}_{y_c}}{r_i \sin(\theta_i)}\right)$, where $r_i = \sqrt{(p_{x_i} - p^{'}_{x_c})^2 + (p_{y_i} - p^{'}_{y_c})^2 + (p_{z_i} - p^{'}_{z_c})^2}$, while $p^{'}_{x_c}$, $p^{'}_{y_c}$ and $p^{'}_{z_c}$ are coordinates of the center of the microphones.

5.3 Incorporating the Spatial Covariance Model

The spatial covariance model is incorporated in the source separation framework described in Chapter 4 in two different contexts: firstly, it is combined with the interaural level difference (ILD) and the interaural phase difference (IPD) models; secondly, it is combined with the ILD, IPD, and mixing vector models.

5.3.1 The Combined ILD, IPD and Spatial Covariance Models

The spatial covariance model is first combined with the ILD and IPD models. The spatial covariance model which utilizes the spatial properties of the sources and the room is believed to further the separation performance, specifically when the level of reverberation is high. The combined models given their parameters can be written as

$$\begin{aligned} p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)|\widehat{\Theta}) &= \mathcal{N}(\alpha(\omega,t)|\mu(\omega),\eta^2(\omega)) \\ &. \mathcal{N}(\hat{\phi}(\omega,t)|\xi(\omega),\sigma^2(\omega)).\ \mathcal{N}(\mathbf{x}(\omega,t)|0,\mathbf{R}_{\mathbf{x}}(\omega,t)), \end{aligned} \tag{5.3.1}$$

wherein conditional independence is assumed between the noise models. The parameters of the combined models are estimated, similar to the previous chapter, using the expectation-maximization (EM) algorithm and can be collected as

$$\widehat{\Theta} = \{\mu_i(\omega), \eta_i^2(\omega), \xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega), v_i(\omega,t), \psi_{i\tau}\} \tag{5.3.2}$$

where μ_i, $\xi_{i\tau}$, and η_i^2, $\sigma_{i\tau}^2$ are respectively the means and variances of the ILD, IPD models, and v_i is the scalar variance. The subscript i indicates that the parameters belong to the source i, and τ and ω describe the dependency on delay and frequency.

The log likelihood function ($\mathcal{L}$) given the observations can be written as

$$\begin{aligned} \mathcal{L}(\widehat{\Theta}) &= \sum_{\omega,t} \log p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)|\widehat{\Theta}) \\ &= \sum_{\omega,t} \log \sum_{i,\tau} [\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega),\eta_i^2(\omega)) \\ &. \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)).\ \mathcal{N}(\mathbf{x}(\omega,t)|0,\mathbf{R}_{\mathbf{x}}(\omega,t)).\ \psi_{i\tau}\ ]. \end{aligned} \tag{5.3.3}$$

Similar to the preceding chapter, the EM algorithm is initialized with the known estimated locations of the speakers. In the expectation step (E-

step) of the EM algorithm, the posterior probabilities are computed given the observations and the estimates of the parameters as

$$\begin{aligned}\tilde{\epsilon}_{i\tau}(\omega,t) = \psi_{i\tau}\ .\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega),\eta_i^2(\omega)) \\ .\ \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)).\ \mathcal{N}(\mathbf{x}(\omega,t)|0,\mathbf{R}_{\mathbf{x}}(\omega,t))),\end{aligned} \tag{5.3.4}$$

where $\tilde{\epsilon}_{i\tau}(\omega,t)$ is the expectation of the hidden variable $m_{i\tau}(\omega,t)$, which is unity if the TF point belongs to both source i and delay τ and zero otherwise. In the maximization step (M-step), the parameters of the models are updated using the observations and $\tilde{\epsilon}_{i\tau}(\omega,t)$ from the E-step. The IPD residual model parameters are estimated as

$$\xi_{i\tau}(\omega) = \frac{\sum_t \hat{\phi}(\omega,t;\tau)\tilde{\epsilon}_{i\tau}(\omega,t)}{\sum_t \tilde{\epsilon}_{i\tau}(\omega,t)} \tag{5.3.5}$$

$$\sigma_{i\tau}^2(\omega) = \frac{\sum_t (\hat{\phi}(\omega,t;\tau) - \xi_{i\tau}(\omega))^2 \tilde{\epsilon}_{i\tau}(\omega,t)}{\sum_t \tilde{\epsilon}_{i\tau}(\omega,t)}. \tag{5.3.6}$$

The ILD model parameters are updated as

$$\mu_i(\omega) = \frac{\sum_{t,\tau} \alpha(\omega,t)\tilde{\epsilon}_{i\tau}(\omega,t)}{\sum_{t,\tau} \tilde{\epsilon}_{i\tau}(\omega,t)} \tag{5.3.7}$$

$$\eta_i^2(\omega) = \frac{\sum_t (\alpha(\omega,t) - \mu_i(\omega))^2 \sum_\tau \tilde{\epsilon}_{i\tau}(\omega,t)}{\sum_{t,\tau} \tilde{\epsilon}_{i\tau}(\omega,t)}. \tag{5.3.8}$$

The parameter $\psi_{i\tau}$ is initialized using a PHAT histogram [59]. The spatial covariance matrix of the ith source $\mathbf{R}_i(\omega)$ is obtained using Eq. (5.2.3) whereas the parameter $v_i(\omega,t)$ is estimated as in Eq. (5.2.7).

The spatial covariance model starts contributing from the second iteration, as in the first iteration when calculating $\tilde{\epsilon}_{i\tau}(\omega,t)$, the source i with delay τ is assumed dominant at the corresponding TF unit, and is calculated using only the ILD and IPD models. Also, since $\tilde{\epsilon}_{i\tau}(\omega,t)$ contains the correct

order of the sources as in [59] the permutation problem is avoided. The TF mask for each source can be obtained as $\tilde{M}_i(\omega,t) \equiv \sum_\tau \tilde{\epsilon}_{i\tau}(\omega,t)$. The masks are applied to the mixtures to obtain the individual sources.

5.3.2 The Combined ILD, IPD, Mixing Vector and Spatial Covariance Models

The spatial covariance model is also used in combination with the ILD, IPD and the mixing vector models. The new parameter set is given as

$$\breve{\Theta} = \{\mu_i(\omega), \eta_i^2(\omega), \xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega), v_i(\omega,t), \mathbf{d}_i(\omega), \varsigma_i^2(\omega), \psi_{i\tau}\} \tag{5.3.9}$$

where μ_i, $\xi_{i\tau}$, and $\mathbf{d}_i$ and η_i^2, $\sigma_{i\tau}^2$, and ς_i^2 are respectively the means and variances of the ILD, IPD, and mixing vector models and v_i is the scalar variance. The log likelihood function ($\breve{\mathcal{L}}$) is now written as

$$\begin{aligned}\breve{\mathcal{L}}(\breve{\Theta}) \quad &= \sum_{\omega,t} \log \sum_{i,\tau} [\ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)) \\ . \ \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega)). \ &\mathcal{N}(\mathbf{x}(\omega,t)|0, \mathbf{R}_{\mathbf{x}}(\omega,t)). \\ &\mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{d}_i(\omega), \varsigma_i^2(\omega)).\ \psi_{i\tau}\ ],\end{aligned} \tag{5.3.10}$$

assuming conditional independence between the noise models. The EM algorithm iterates similarly with the combined posterior probabilities computed as

$$\begin{aligned}\breve{\epsilon}_{i\tau}(\omega,t) = \psi_{i\tau}\ . \ &\mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)) \\ . \ \mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega)). \ &\mathcal{N}(\mathbf{x}(\omega,t)|0, \mathbf{R}_{\mathbf{x}}(\omega,t))), \\ &\mathcal{N}(\mathbf{x}(\omega,t)|\mathbf{d}_i(\omega), \varsigma_i^2(\omega)).\end{aligned} \tag{5.3.11}$$

All the corresponding parameters are estimated as explained in Section 5.3.1 and Chapter 4. The mixing vector and spatial covariance models start contributing from the second iteration. The TF masks for the individual sources

are then obtained as

$$\breve{M}_i(\omega, t) \equiv \sum_{\tau} \breve{\epsilon}_{i\tau}(\omega, t). \tag{5.3.12}$$

Masks are applied to the mixtures to reconstruct individual sources.

5.4 Experimental Evaluation in a Room Environment

Experiments are performed with the spatial covariance model used in both the contexts explained in Section 5.3.1 and Section 5.3.2. The room settings are similar as in Chapter 4. Speech data were chosen from the TIMIT [23] database. The first 40,000 (16k×2.5) samples of the TIMIT sources were used and were normalized to unity variance. The source image method [3] was used to evaluate the different models for varying RT60s. The different reverberation times under consideration are: 160 ms, 300 ms, 485 ms and 600 ms. The signal-to-distortion ratio (SDR) [64] was used to measure the separation performance of the algorithms.

Detailed experiments were performed ranging from mixtures simulated with varying reverberation times (RT60s), sources with varying separation angles, and different model complexities. The spatial covariance model used in conjunction with the ILD and IPD models, termed as IIM+SC, is compared with the ILD and IPD models, referred to as IIM, and the combined ILD, IPD, mixing vectors and the spatial covariance models, termed as IIMM+SC. The common parameters used in all experiments are given in Table 5.1.

Table 5.1. Common Parameters Used In Simulations

STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	160-600 ms (image method)
Room dimensions	[9 5 3.5] m
Source signal duration	2.5 s (TIMIT)
Sensor spacing	0.17 m

In the first set of experiments the target source was positioned at 0° azimuth while the interferer at 75°. The level of reverberation was then varied from 160 ms to 600 ms, and the separation performance measured for the three different model complexities, Θ_{00}, Θ_{11} and $\Theta_{\Omega\Omega}$. The different model complexities [59], $\Theta_{ILD\ IPD}$, mean that the parameters of the ILD and IPD models are either frequency-dependent or are fixed across frequency.

5.4.1 Results

Fig. 5.1 shows results for the models with the complexity Θ_{00} with the interferer located at 75°. With this complexity the mean of the ILD model is zero and the standard deviation is ∞, while the IPD model has a mean zero and a frequency-independent standard deviation. The results indicate that combining the spatial covariance model to the ILD and IPD models the separation improves consistently over all RT60s. The separation, nevertheless, is best over all RT60s when the maximum cues are utilized, the IIMM+SC method. As the level of reverberation increase, the advantage of the IIM+SC method over the IIM technique also increase, for instance, at an RT60 of 160 ms it is 0.42 dB and at 600 ms it is 0.71 dB better than the IIM method. A similar trend is followed by the IIMM+SC method, in that it is 1.19 dB and 1.81 dB better than the IIM technique at 160 ms and 600 ms respectively.

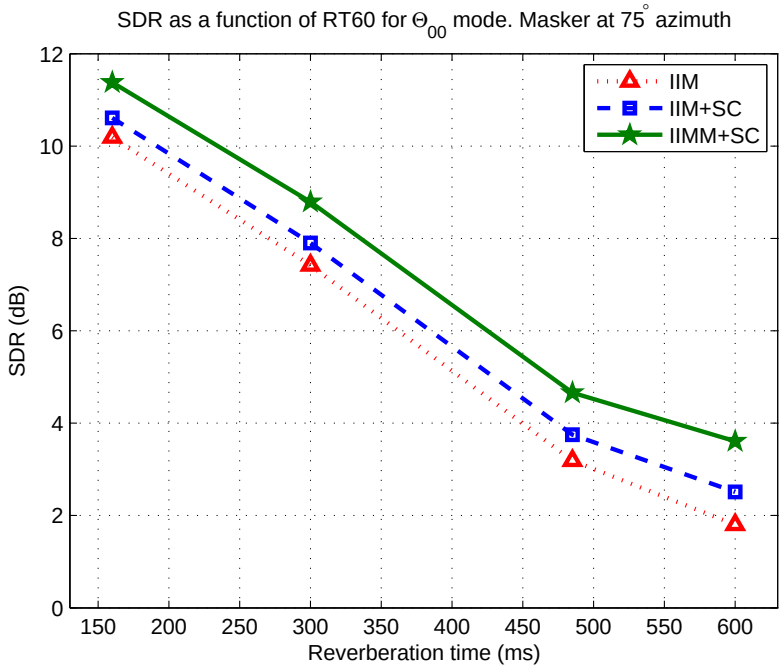


Figure 5.1. SDR (dB) for the Θ_{00} model. The interference is placed at 75°

Fig. 5.2 provides the separation results in terms of the SDR (dB) over a range of RT60s for the complexity Θ_{11}. Within this complexity the means and the variances of both the ILD and IPD models are frequency-independent. All the methods follow the similar trend of degrading performance as the level of reverberation increases and improved performance while exploiting more cues. At the RT60 of 300 ms, the IIMM+SC is 1.12 dB and 0.64 dB better than the IIM and IIM+SC methods, while at 600 ms it is 1.47 dB and 0.98 dB better respectively.

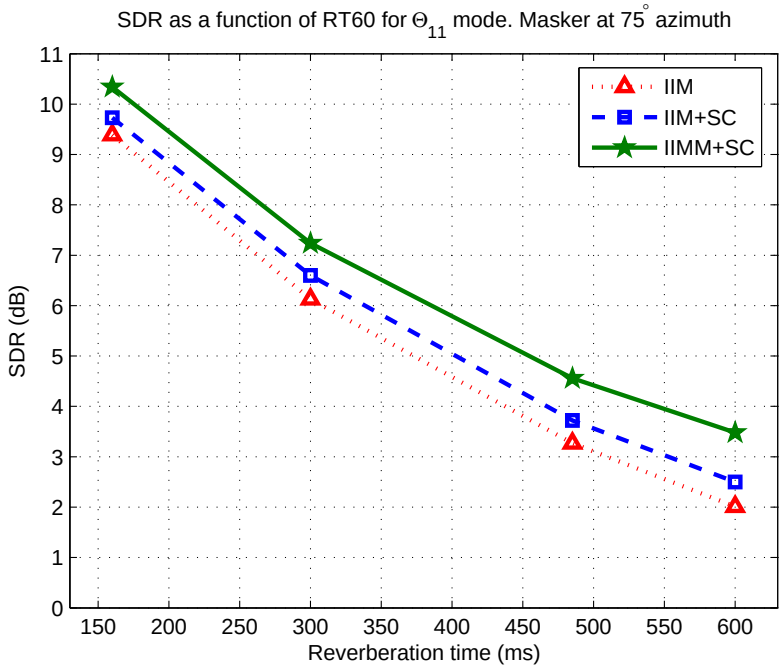


Figure 5.2. SDR (dB) for the Θ_{11} model. The interference is placed at 75°

In Fig. 5.3 results are shown for the $\Theta_{\Omega\Omega}$ complexity with the masker positioned at 75°. The means and variances of the ILD and IPD models in this complexity depend on frequency. At 160 ms, the IIM method has an SDR of 7.51 dB, whereas, the IIM+SC and IIMM+SC are 7.80 dB and 8.44 dB respectively. As the reverberation increases the separation performance degrades, in that at 600 ms, the IIM, IIM+SC and IIMM+SC techniques have an SDR of 1.55 dB, 2.01 dB, and 2.88 dB respectively.

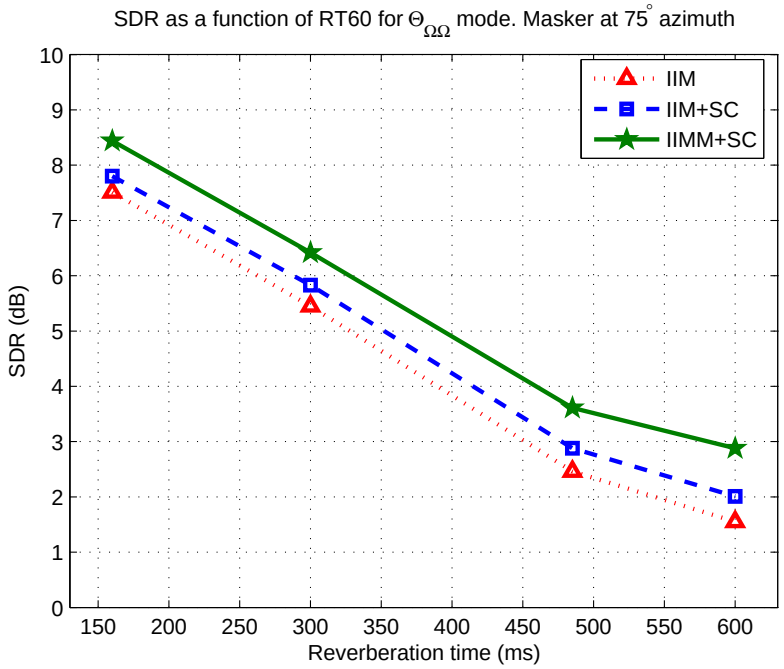


Figure 5.3. SDR (dB) for the $\Theta_{\Omega\Omega}$ model with the interference placed at 75°

Figures 5.4, 5.5, and 5.6 depict the scenario, over the range of RT60s, when the target and interferer are separated by 15°. Separation in this scenario is particularly challenging since the sources are close together and the interaural cues become indistinct, hence, greatly degrading the separation performance.

In Fig. 5.4, for the Θ_{00} complexity, it can be observed that all the methods in general perform worse when the separation between the sources is smaller than if they are well apart. At 300 ms the IIM method has an

SDR of 1.81 dB whereas in the similar complexity but with the separation angle of 75° the SDR was 7.42 dB. The IIM+SC improves the separation of IIM by around 0.6 dB, while the IIMM+SC method by 2.47 dB at 300 ms. As the RT60 increases, the contribution by the spatial covariance model also slightly increase and at 600 ms the IIM+SC is 0.8 dB better than the IIM method, while the IIMM+SC performs best by improving separation around 2.87 dB over the IIM method.

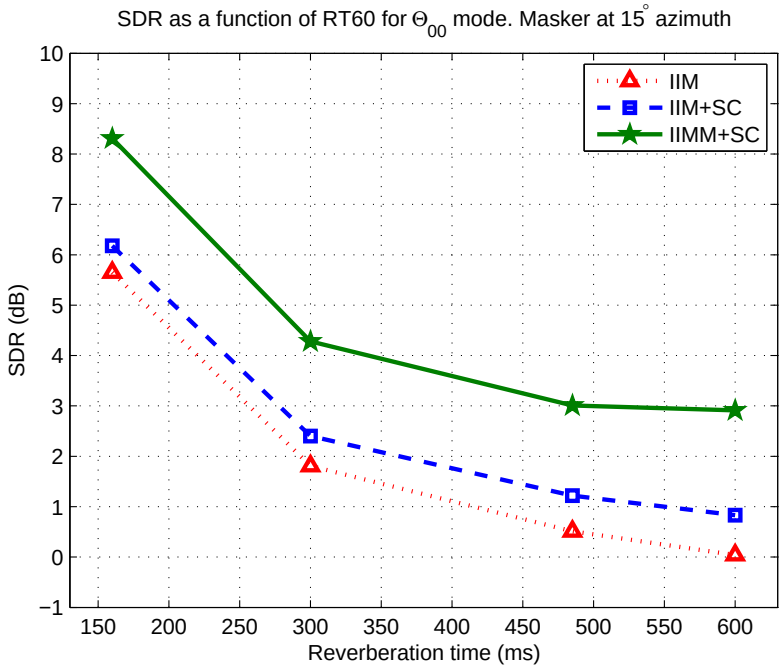


Figure 5.4. SDR (dB) for the Θ_{00} model with the interference placed at 15°

Fig. 5.5 shows results for the Θ_{11} complexity with the separation angle of 15°. Within this complexity, since there is contribution from the ILD and IPD cues (although the parameters of these models are not dependent on frequency), the contribution from the spatial covariance model is slightly reduced. For instance at 485 ms, the IIM+SC method is 0.5 dB while IIMM+SC is 2.26 dB better than IIM. When frequency dependency is introduced within the parameters of the ILD and IPD models, the $\Theta_{\Omega\Omega}$ complexity, results for which are shown in Fig. 5.6, the performance of the IIM technique further deteriorates. As stated previously, to the fact that the sources are too closely spaced, the interaural cues are almost identical. At 485 ms, the addition of the spatial covariance model to the ILD and IPD models, IIM+SC, improve the performance by 0.51 dB, while the IIMM+SC method by 1.94 dB.

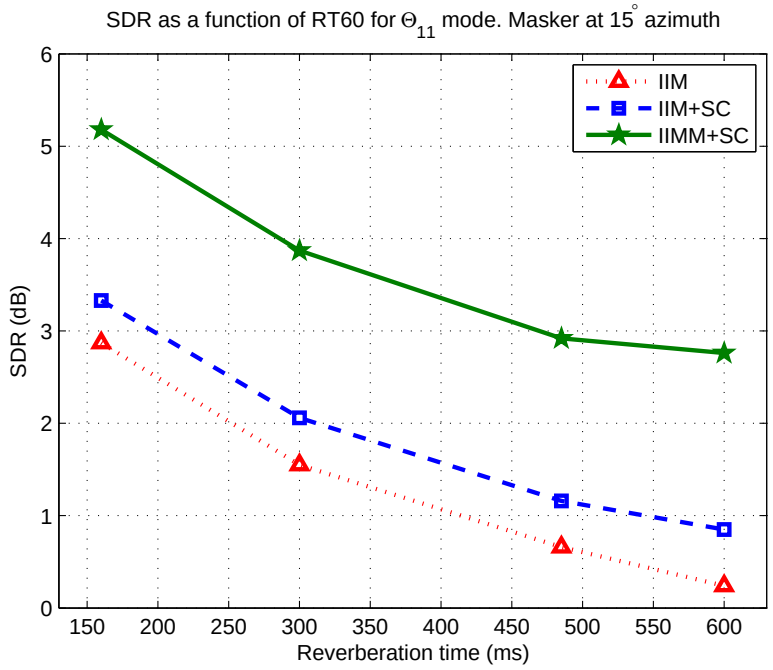


Figure 5.5. SDR (dB) for the Θ_{11} model with the interference placed at 15°

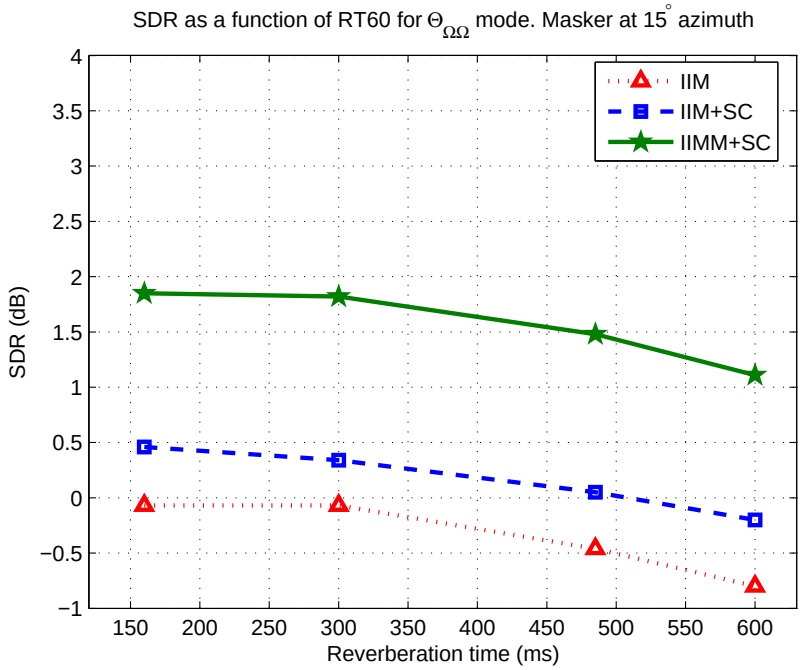


Figure 5.6. SDR (dB) for the $\Theta_{\Omega\Omega}$ model with the interference placed at 15°

5.5 Summary

This chapter presented the spatial covariance model that utilized the locations of the speakers, potentially estimated through a video process, and the attributes of the room such as its wall reflective properties, wall area and reverberation time of the room. The model was used in conjunction with the ILD and IPD models and ILD, IPD, and mixing vector models. The parameters for the models were obtained using the EM algorithm that produced improved TF masks for each source. The masks were used to extract the sources. Experimental results verified that the proposed algorithm can perform better, in general, than the algorithm that uses only the ILD and IPD models, over all considered levels of reverberation. The separation performance was specifically better when the separation angle between the sources was small and the mixture was highly reverberant. The inclusion of the spatial covariance model improves the separation, but the improvement is not very significant i.e. typically less than 1 dB in terms of SDR. Further refinements may be required to achieve additional improvement, possibly in terms of source variance estimation, or incorporating the model in way that it is refined at each EM iteration, but this is left for future research.

In a further step to tackle the room reverberation, the proceeding chapter explores binaural dereverberation schemes that suppress the late components of reverberation from the observed mixtures before source separation. Within this pre-processing, based on spectral subtraction, the late reverberant components are estimated, in the time-frequency domain, and are suppressed to dereverberate the mixture. A novel cascade structure is also investigated, within which three dereverberation stages are utilized provide an increased reverberation suppression. The source separation algorithm is then run on the dereverberated mixtures to give enhanced estimates of the sources.

Chapter 6

DEREVERBERATION BASED PRE-PROCESSING FOR THE SUPPRESSION OF LATE REVERBERATION BEFORE SOURCE SEPARATION

6.1 Introduction

Room reverberation, produced by multiple reflections of the sound on wall surfaces and objects in an enclosure, remains a challenge for many signal processing applications, such as automatic speech recognition (ASR), hearing aids and hands-free telephony. Specifically, the late reflections of the room impulse response (RIR) cause spreading of the speech spectra and degrade the quality of speech and the intelligibility [25]. The objective of dereverberation algorithms is to suppress the effects of reverberation while minimally distorting the speech structure.

Monaural dereverberation algorithms based on spectral subtraction, e.g. [25, 98], have been proposed to suppress the effects of late reflections. Effective extension of the monaural methods to the binaural context is important

as this would enable their utilization in multiple applications. Such extensions must produce minimal musical noise and also preserve the binaural cues i.e. interaural time difference (ITD) and the interaural level difference (ILD) [99, 100].

6.2 Monaural Dereverberation and Extension into the Binaural Context

In spectral subtraction based dereverberation techniques, given a reverberant signal in the TF domain, for instance, $S_{rev}(\omega, t)$, a dereverberated signal, $S_{cln}(\omega, t)$, can be obtained by subtracting the late reverberant component $S_{rev_{late}}(\omega, t)$ as,

$$S_{cln}(\omega, t) = S_{rev}(\omega, t) - S_{rev_{late}}(\omega, t) \tag{6.2.1}$$

where ω is the frequency index at the time frame t. Alternatively, the process can also be expressed as

$$S_{cln}(\omega, t) = G(\omega, t) S_{rev}(\omega, t) \tag{6.2.2}$$

where $G(\omega, t)$ is a gain function applied to the observed reverberant signal, and can be computed by estimating the late reverberant component as

$$G(\omega, t) = \frac{S_{cln}(\omega, t)}{S_{rev}(\omega, t)} = \frac{S_{rev}(\omega, t) - S_{rev_{late}}(\omega, t)}{S_{rev}(\omega, t)}. \tag{6.2.3}$$

In the monaural dereverberation method in [98], a statistical model of the room impulse response is proposed in order to subtract spectrally the late reverberant components, assuming that the direct-to-reverberant (DRR)

ratio is low. The gain function is computed as

$$G(\omega,t) = 1 - \frac{1}{\sqrt{SIR_{post}(\omega,t)+1}} \tag{6.2.4}$$

where $SIR_{post}(\omega,t)$ is the *a posteriori* signal-to-interference ratio (SIR) calculated as

$$SIR_{post}(\omega,t) = \frac{|S_{rev}(\omega,t)|^2}{\sigma^2_{S_{rev_{late}}}(\omega,t)} \tag{6.2.5}$$

where $\sigma^2_{S_{rev_{late}}}(\omega,t)$ is the variance of the late reverberant speech component and is estimated as

$$\sigma^2_{S_{rev_{late}}}(\omega,t) = \exp(-2\kappa T_l) \cdot \sigma^2_{S_{rev}}(\omega, t-n_{late}) \tag{6.2.6}$$

where $\kappa = \frac{3ln(10)}{RT60}$, T_l indicates the time from which the late reverberation starts, n_{late} is the number of samples related to T_l, $RT60$ indicates the reverberation time (assumed to be known), and $\sigma^2_{S_{rev}}$ is the variance of the reverberant mixture computed by recursive averaging [99]

$$\sigma^2_{S_{rev}}(\omega,t) = \delta \cdot \sigma^2_{S_{rev}}(\omega,t-1) + (1-\delta) \cdot |S_{rev}(\omega,t)|^2 \tag{6.2.7}$$

where $\delta \in [0,1]$ is the smoothing factor.

This monaural scheme is extended to the binaural form in [99] where a delay-and-sum beamformer is used to generate a reference signal by averaging the time-aligned left and right reverberant signals. The reference signal is then processed to generate the weighting gains using Eq. (6.2.4). In [100] the left and the right reverberant mixtures are separately processed to yield two gains. The two gains are then combined, e.g. by taking the minimum, maximum or average, and applied to both the channels. The procedure in [100] is adopted by independently processing the two channel signals and two gain functions are obtained. A single gain is then formed

using the following linear combination [101],

$$G_{new} = \alpha G_L + (1 - \alpha) G_R \tag{6.2.8}$$

where G_L and G_R are the left and right channel gains and α is a weighting factor chosen empirically, $0 \leq \alpha \leq 1$. The proposed scheme allows the suppression of late reverberation in a flexible way by selecting a suitable α. The processing is depicted in Fig. 6.1.

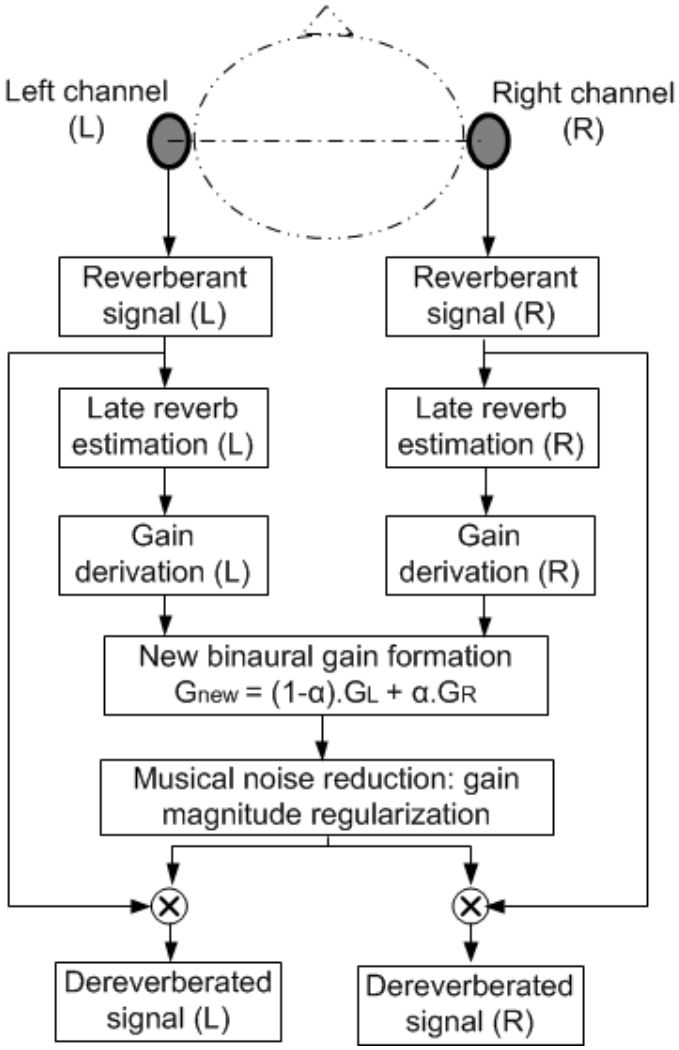


Figure 6.1. Processing overview with the bilateral signal processing and gain derivation.

Due to errors in the estimation of the weighting gains, musical noise is likely to have been introduced. Smoothing of the derived gain is performed as in [99], [102] where an estimation is performed to detect if a frame contains speech (has high SIR) or not, and thus attenuate the frames with low SIRs. The power ratio of the enhanced signal and the reverberant signal [102],

$$\zeta(t) = \frac{\sum_{\omega=1}^{W} G_{new}(\omega,t).|Y(\omega,t)|^2}{\sum_{\omega=1}^{W} |Y(\omega,t)|^2} \tag{6.2.9}$$

is computed to indicate whether the SIR of a time frame is low or high. If $\zeta(t)$ is approximately unity, the SIR of that frame is assumed to be high, and if $\zeta(t)$ is nearly zero, the SIR is supposed to be low. A moving average window is then applied to smooth the weighting gain magnitudes [99].

To verify the suitability of the above binaural dereverberation scheme, in the context of source separation, it is appended as a pre-processing stage to the source separation method described in Chapter 4. Since the source separation algorithm also utilizes models of the ILD and IPD, experimental evaluation is considered to be useful in that the results would indicate whether the binaural dereverberation based pre-processing preserves the binaural cues or not.

6.3 Experimental Evaluation

Experiments were conducted by pre-processing the observed reverberant mixture using the aforementioned binaural dereverberation method and then performing the source separation described in Chapter 4. Results were compared with only the source separation algorithm in order to highlight the gain that could be achieved by including the pre-processing. The speech files in these experiments also come from the TIMIT database [23]. Experiments

were performed for mixtures of two and three speech sources. Real RIRs used, that come from [103], were convolved with speech sources to generate the reverberant mixtures. These RIRs were measured in real rooms having different reverberation times.

For evaluation purposes, the signal-to-distortion ratio (SDR) [64] was used. Perceptual evaluation of speech quality (PESQ) [63] was also used as a performance measurement metric to reveal the quality of the processed speech. Results for the scheme with the pre-processing, referred to as derev+IIMM, are compared with the source separation method in Chapter 4, termed as IIMM.

Results in Fig. 6.2 depict the SDR (dB) at different RT60s when the interfering source is at a relatively smaller separation angle of 15°. The Θ_{11} model is under consideration here, where both the ILD and IPD models are frequency-independent. The graph clearly indicates the improvement achieved by incorporating the binaural dereverberation based pre-processing. The improvement is consistent over all the RT60s and generally increases when the RT60 grows. For instance, at RT60 of 320 ms, the derev+IIMM is 1.61 dB and at 890 ms it is 3.59 dB better than IIMM.

The PESQ results for the same scenario are shown in Fig. 6.3. These results indicate that the pre-processing, by suppressing the late reverberant components, improves the quality of the separated speech. At 320 ms, the derev-IIMM method improves the PESQ by 0.13 and at 890 ms by 0.15.

Fig. 6.4 provides results for the $\Theta_{\Omega\Omega}$ model with the masker at 15° azimuth. A similar general trend of improved performance over all RT60s is followed when the frequency-dependent ILD and IPD models are considered. The derev+IIMM method provides an improvement, in terms of SDR, of 1.02 dB at 320 ms and 1.71 dB at 890 ms. Fig. 6.5 shows the PESQ results for the similar experimental setting. Results over all RT60s show an improvement in quality, in terms of the PESQ measure, when the mixtures are first pre-

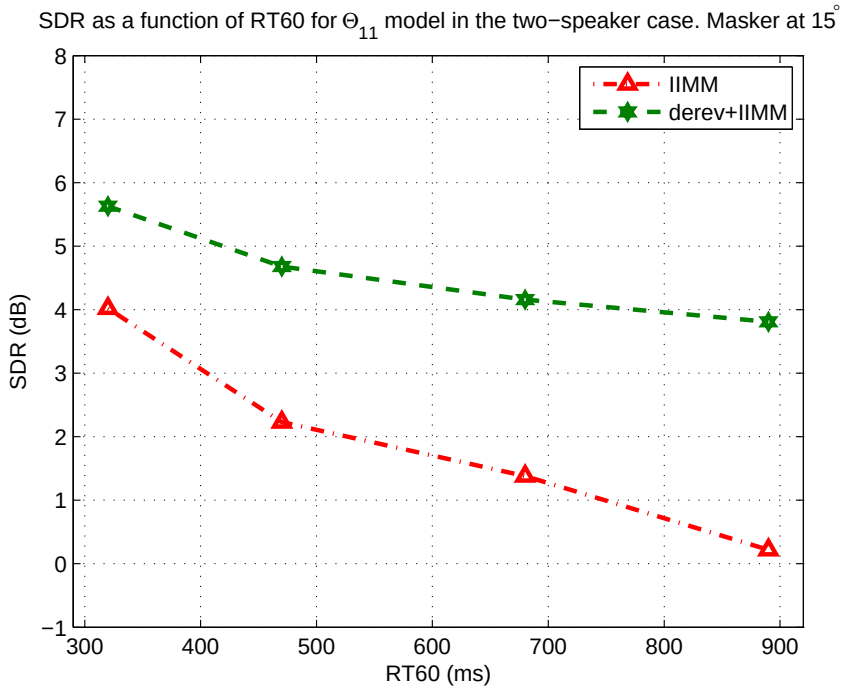


Figure 6.2. SDR (dB) for the Θ_{11} model with varying RT60s and the interference positioned at 15° azimuth.

processed before separation.

The following section explores a novel cascade structure for binaural dereverberation. The study is motivated by the fact that realistic environments are highly reverberant, and in these circumstances the performance of even the state-of-the-art methods degrades significantly, as such there is a need for additional processing to mitigate the distortions produced by reverberation and thus improve the separation performance.

6.4 Cascade Structure for Spectral Subtraction Based Binaural Dereverberation

A cascade structure for spectral subtraction based binaural dereverberation of audio signals is investigated. Three binaural dereverberation blocks are

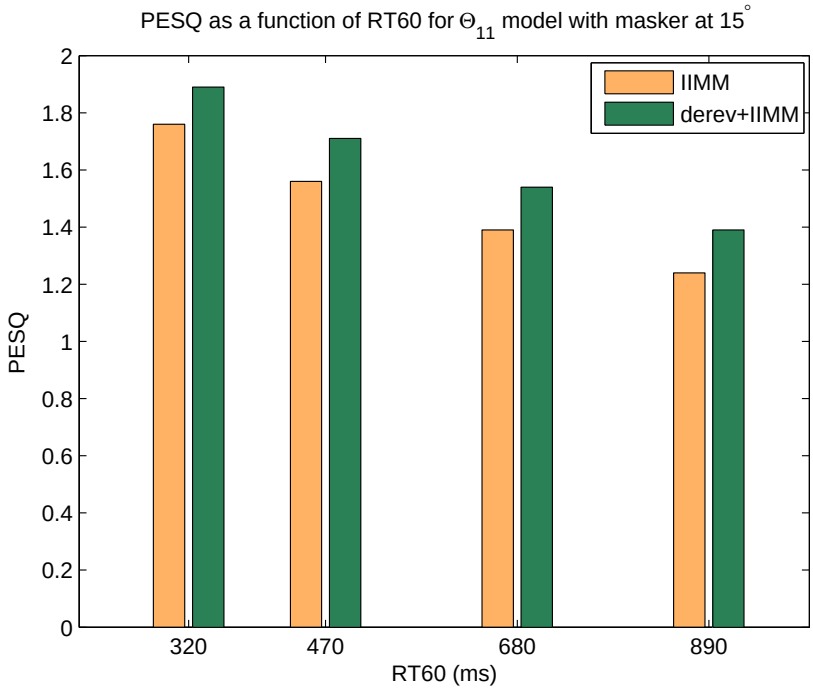


Figure 6.3. PESQ for the Θ_{11} model with varying RT60s with the interference located at 15° azimuth.

utilized. The first two stages exploit distinct observations to model and suppress the late reverberation by deriving a gain function. The musical noise artifacts generated due to the processing at the first two stages are mitigated by smoothing the spectral magnitudes of the weighting gains. The third stage linearly combines the gains obtained from the first two stages and further enhances the binaural signals. The binaural gains, obtained by independently processing the left and right channel signals are combined as a convex mixture.

The entire dereverberation process is a combination of three cascaded stages. Each stage takes in a binaural input and gives a binaural output in the time-domain. The algorithm diagram is given in Fig. 6.6. The enhancement of each stage is cumulative as the overall non-linearity in the processing is a form of nesting which relates to a fixed point iteration [104].

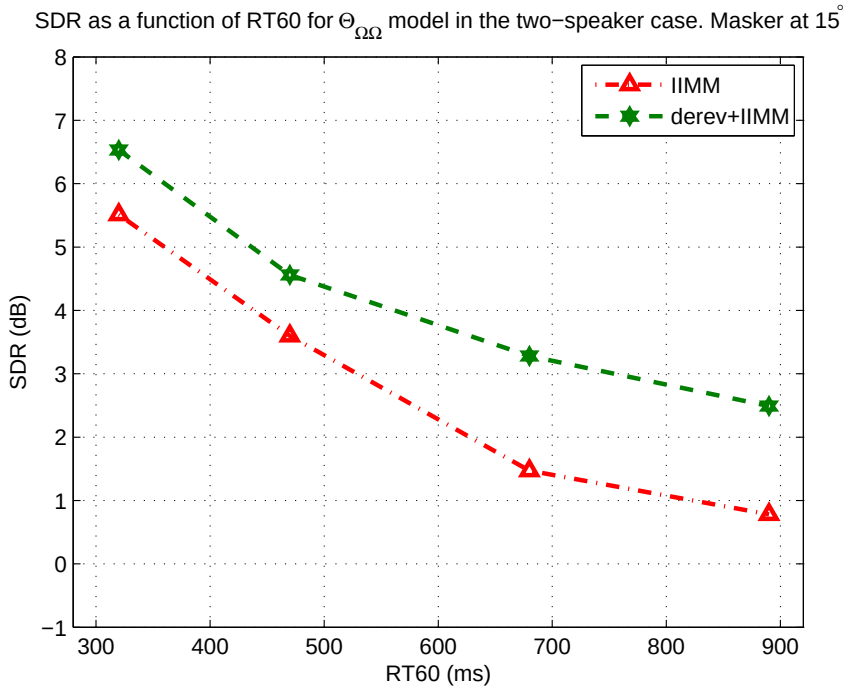


Figure 6.4. SDR (dB) for the $\Theta_{\Omega\Omega}$ model with varying RT60s and the interference positioned at 15° azimuth.

With a cascade of these non-linear processors, a higher overall enhancement is achievable which may not be possible by each stage individually, or by repeatedly cascading the same block.

The time-domain left and right channel reverberant signals are input to the first stage where they are independently processed using the monaural dereverberation method proposed in [98], described in Section 6.2. This method, which is referred to as LB-RIR (the acronym is derived from the authors' names, Lebart et al., and their technique which is based on RIR modeling), is extended into the binaural context using the proposed method to obtain a gain function which is then smoothed, as explained in Section 6.2.

Stage 2 makes use of the monaural scheme in [25], which is termed as

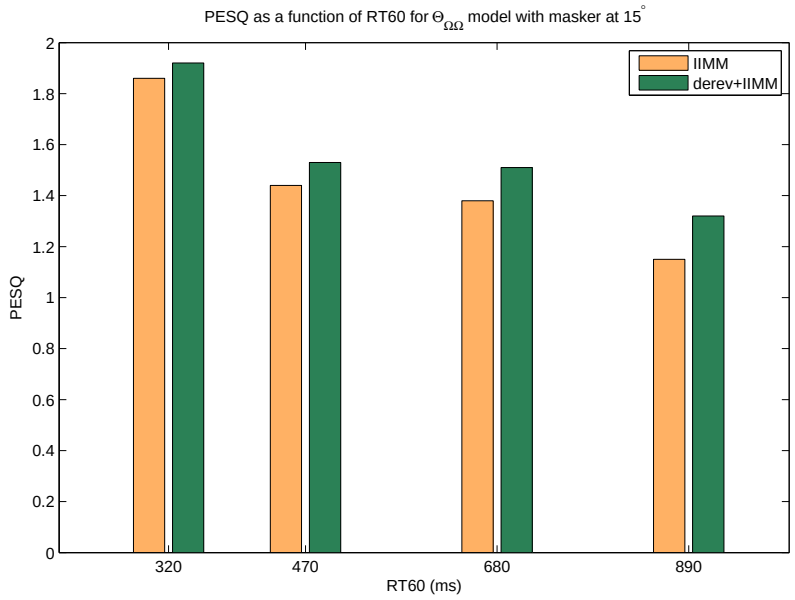


Figure 6.5. PESQ for the $\Theta_{\Omega\Omega}$ model with varying RT60s with the interference located at 15° azimuth.

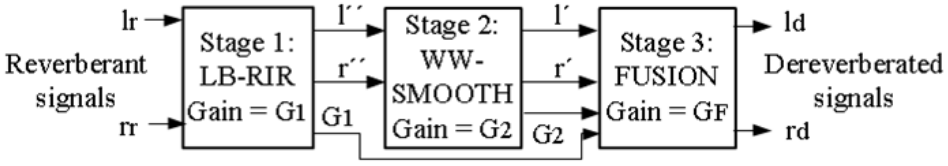


Figure 6.6. The proposed cascaded approach for binaural dereverberation.

WW-SMOOTH (the acronym is derived from the authors' names, Wu and Wang, and their method which is based on smoothing of the signal spectrum). This method is motivated by the observation that the spreading due to the late reverberation causes smoothing of the signal spectrum in the time domain. Thus, the power of the late reverberant component is estimated as the smoothed and shifted version of the power of the reverberant speech in

the TF domain

$$|X_{rev_{late}}(\omega,t)|^2 = \gamma\varpi(t-\rho) * |X_{rev}(\omega,t)|^2 \tag{6.4.1}$$

where $*$ indicates the convolution operation, γ is a scaling factor, and ρ is the shift delay. The term $\varpi(t)$ is a smoothing function given as the shifted Rayleigh distribution [25]

$$\varpi(t) = \begin{cases} \frac{t-a}{a^2}\exp(\frac{-(t-a)^2}{2a^2}), & if\ t > -a \\ 0, & otherwise \end{cases}$$

where a indicates the integer but non-zero number of frames and needs to be smaller than ρ. Here $a = 5$ while $\rho = 7$ as in [25].

The method in [25] is also extended to binaural in a similar manner as in stage 1, and the smoothing of the weighting gain follows accordingly. The enhanced signals from stage 2 are forwarded to stage 3. The weighting gains from stage 1 and stage 2 are linearly fused to form a combined gain. The fused gain is used to further suppress the late reverberant components from the left and right channel signals and give the final dereverberated signals. The advantage of the proposed approach is next experimentally verified.

6.5 Experimental Evaluation

The proposed cascade structure for binaural dereverberation is experimentally tested in two processing contexts: firstly, for the purpose of dereverberation only; secondly, using the proposed cascade as a pre-process to a source separation algorithm.

The anechoic speech utterances in all experiments come from the TIMIT database [23]. Real binaural RIRs (BRIRs) from the Aachen impulse response (AIR) database [105] were used in the dereverberation-only experiments while in the joint dereverberation and source separation experiments

RIRs are used from [103]. The frame length used was 512 and the frame overlap was 75 percent. The other parameter values were the same as in the original works [25, 98, 99].

6.5.1 Dereverberation-only

Speech files from TIMIT were chosen randomly containing both male and female speakers. In the AIR database [105], the first set of BRIRs used here were measured in an office room with source-to-microphone distance of 1 m and 3 m with an RT60 of 0.37 s and 0.48 s respectively. The BRIRs in the second set were measured in a lecture room with source-to-microphone distance of 2.25 m, 5.56 m and 10.2 m with an RT60 of 0.70 s, 0.79 s, and 0.83 s respectively. Both the LB-RIR and WW-SMOOTH schemes were applied to the observed reverberant signals without any inverse filtering.

For performance evaluation in the dereverberation-only case, three objective measures were used including the signal-to-noise ratio (SNR), segmental SNR (segSNR), and the perceptual evaluation of speech quality (PESQ) [63].

Table 6.1. Mean values of SNR (dB), segmental SNR (segSNR) (dB) and PESQ for three random signals from TIMIT convolved with BRIRs from the Aachen database. RT60s of 0.37, 0.48, 0.70, 0.79, and 0.83 seconds under consideration.

RT60	SNR (dB) Improv.		segSNR (dB) Improv.		PESQ Improv.	
(s)	LB-RIR	Cascade	LB-RIR	Cascade	LB-RIR	Cascade
0.37s	0.40	0.65	0.64	0.98	0.16	0.32
0.48s	0.86	1.37	1.21	2.06	0.09	0.20
0.70s	0.87	1.27	1.73	2.33	0.24	0.27
0.79s	0.75	1.22	1.14	1.87	0.16	0.30
0.83s	1.12	1.71	1.50	2.34	0.22	0.32

Table 6.1 summarizes the experimental results in the context of dereverberation-only processing. LB-RIR in the table means that the signal is enhanced using

the LB-RIR method and extended to binaural as in [99]. Each value in the table is an average of three randomly selected speech signals from the TIMIT database. It can be seen that the proposed approach provides an improvement in all the three evaluation metrics. Over all the RT60s, the proposed method gives a mean SNR gain of 1.13 dB, mean segSNR gain of 1.92 dB, and PESQ improvement of 0.28, compared to LB-RIR which gives an SNR gain of 0.8 dB, segSNR gain 1.24 dB, and a PESQ improvement of 0.17.

6.5.2 Dereverberation and Source Separation

In this set of experiments the proposed cascade structure is used as a preprocessing stage before the source separation (termed as Cascade+IIMM), as was also done in Section 6.3. BRIRs used here [103] were measured in four different rooms with RT60s of 0.32, 0.47, 0.68, and 0.89 seconds.

Fig. 6.7 provides a comparison, in terms of SDR (dB), between the source separation algorithm, IIMM, the derev+IIMM scheme, and the Cascade+IIMM approach. The scenario under consideration here is the same as in Fig. 6.4 so as to highlight the gain achieved with the proposed cascade. The Casacade+IIMM furthers the SDR consistently, and over all RT60s, provides an average improvement of around 1.5 dB over the derev+IIMM method.

Fig. 6.8 gives the PESQ comparisons for the different methods, with the experimental setting similar to Fig. 6.5. The cascaded approach performs well in terms of PESQ too, with notable improvement of 0.24 at the RT60 of 680 ms.

Figures 6.9 and 6.10 provide results in terms of SDR (dB) and PESQ respectively for mixtures of three sources. The three sources were mixed with varying levels of reverberation. The target was at 0° azimuth while the interfering sources were symmetrically located at a separation of 45° on its either sides. In terms of SDR, both the derev+IIMM and Cascade+IIMM

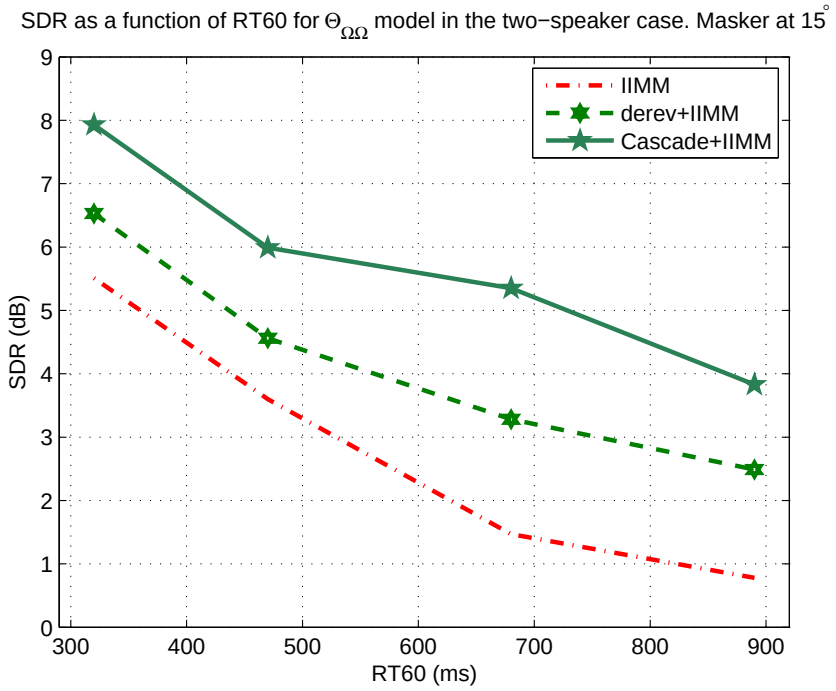


Figure 6.7. SDR (dB) for the $\Theta_{\Omega\Omega}$ model with varying RT60s and the interference positioned at 15° azimuth. Cascade+IIMM providing a superior performance.

improve performance over the IIMM method by 1.5 dB and 2.56 dB respectively on average over all the considered RT60s. Alongside SDR, the PESQ scores also show consistent improvement. On average over all RT60s, the Casacade+IIMM method provides 0.27 and the derev+IIMM scheme 0.15 improvement in terms of PESQ over the IIMM approach.

6.6 Summary

This chapter studied binaural dereverberation techniques based on amplitude spectral subtraction and their utilization in the context of source separation. Late reverberation is said to have deleterious effects on the fine signal spectrum and suppressing them can generally improve the performance of many signal processing applications. A monaural dereverberation scheme

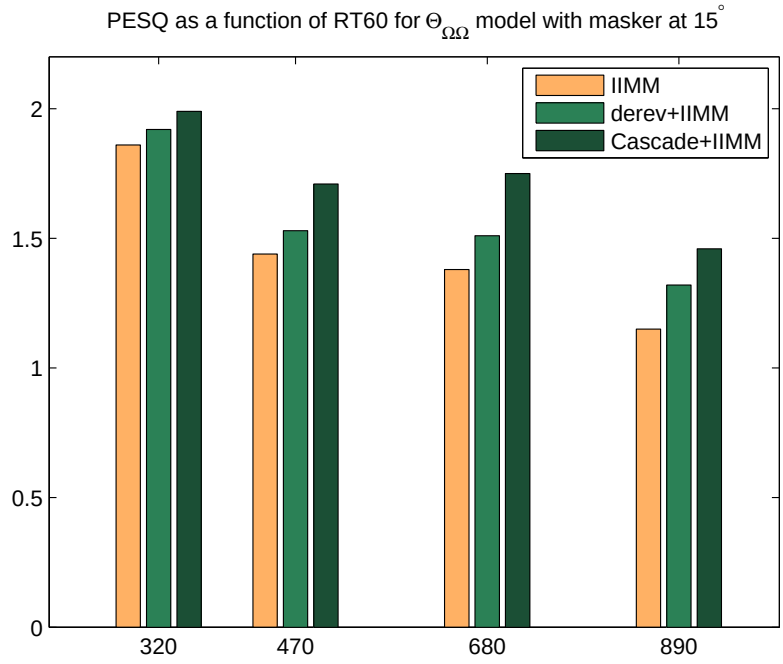


Figure 6.8. PESQ for the $\Theta_{\Omega\Omega}$ model with varying RT60s with the interference located at 15° azimuth. Cascade+IIMM showing an improved performance.

based on the model of room impulse response was first studied and then extended into the binaural context. It was then used as a pre-processing stage to the source separation algorithm and the performance was compared in different scenarios. The pre-processing proved to be useful in that it provided improvements both in terms of SDR and PESQ, when used with source separation.

Later, a cascade structure was explored to achieve further enhancement. The proposed cascade had three stages, with each stage providing signal enhancement. The cascade structure for binaural dereverberation was also used as a pre-process before source separation. The cascade was also evaluated for dereverberation purposes too. Detailed experiments were conducted and real data was utilized. The cascade structure was shown to provide im-

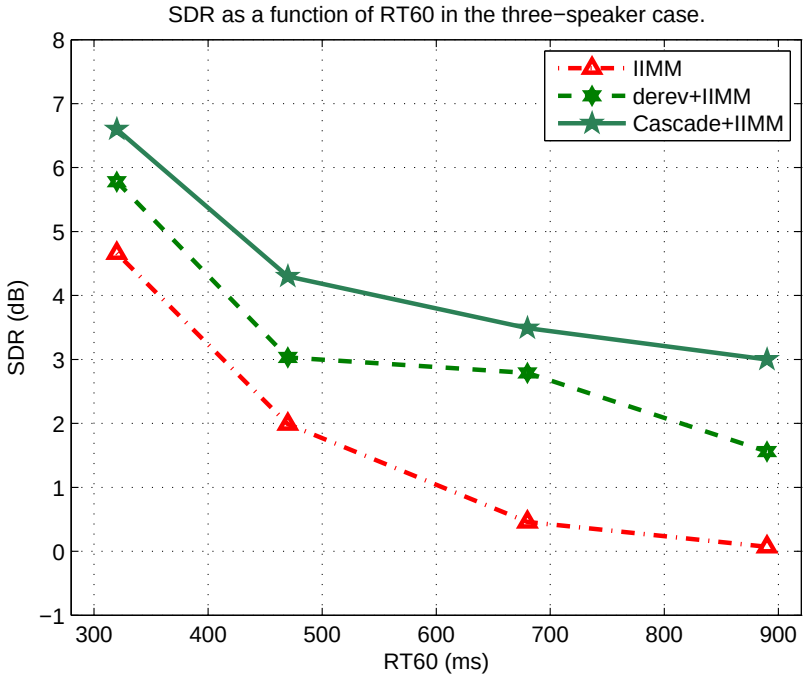


Figure 6.9. SDR (dB) for the $\Theta_{\Omega\Omega}$ model with varying RT60s for mixtures of three speakers. The interfering sources positioned at 45° symmetrically on both sides of the target source.

proved performance over its single-stage counterpart, both in the context of dereverberation only and joint dereverberation and source separation.

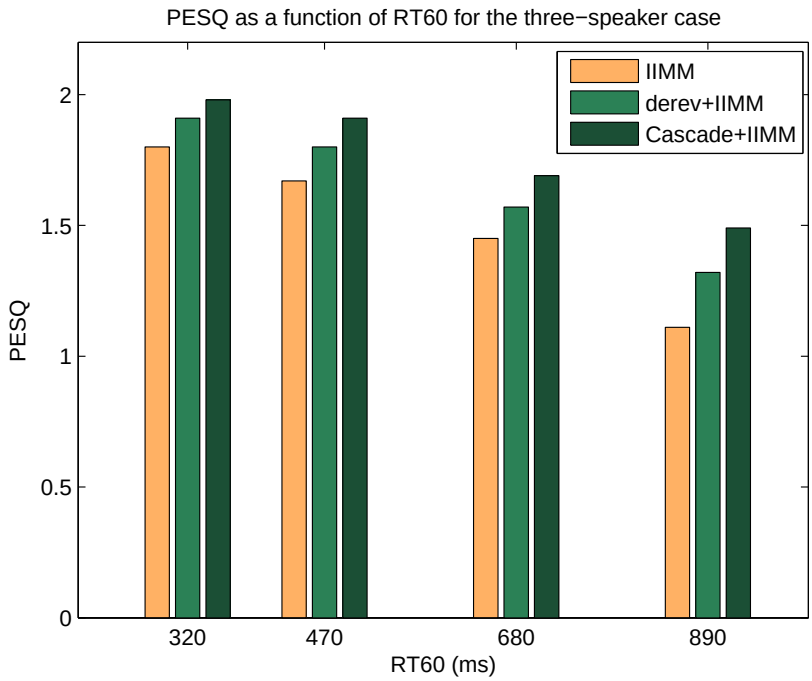


Figure 6.10. PESQ results for varying RT60s in the three-source case. Cascade+IIMM providing an improved performance.

Chapter 7

CONCLUSION

This thesis introduced new techniques for separating multiple sound sources from their reverberant mixtures. It assumed that the locations of the sound sources were known a priori or provided by independent video processing.

Humans are skilled at selectively attending to one sound of interest while many sounds are simultaneously present. Machines, in contrast, can only perform simple forms of these tasks i.e. in anechoic conditions or typically mixtures with only two sources. The performance of the current source separation systems in multi-source realistic reverberant environments is very limited. The work in this thesis aimed at improving the performance of such source separation algorithms in reverberant scenarios by exploiting the sound source locations.

In Chapter 3 a new multi-microphone array based method combined with binary time-frequency masking was presented. A robust least squares frequency invariant data independent beamformer was designed. The robust beamformer being aware of the source locations provided improved estimates of the sources. A white noise gain constraint was also added for further robustness. The beamformer weight vectors were estimated using convex optimization techniques. With the intention to further enhance the separated sources, binary time-frequency masking based post-processing was incorporated. The sources estimated by the beamformer were transformed into the time-frequency domain, and the amplitudes of the corresponding

time-frequency points were compared. Binary masks were thus obtained for each source. Since the ideal binary masks are likely to have introduced unwanted musical noise, smoothing was applied in the cepstral domain. The smoothed binary masks were applied to the mixture to give the final separated sources. Experimental results indicated that the binary time-frequency masking based processing significantly improved the separations, but introduced musical noise. For instance, at RT60 of 600 ms, the robust beamformer without post-processing provided an average advantage of 2.27 dB in terms of the signal-to-distortion ratio (SDR) over the independent vector analysis (IVA) based method. When the post-processing was introduced, for example at RT60 of 300 ms, the signal-to-interference-noise ratio improved from 11.25 dB to 12.18 dB, thus further enhancing the sources. The proposed method was applicable only in the over-determined setting. The next chapter thus pursued a two-microphone method inspired by human hearing.

In Chapter 4 a novel computational auditory scene analysis (CASA) based approach was proposed that utilized the combined probabilistic models of the interaural level and phase differences and mixing vectors, and exploited the information about the source locations. The method was based on the assumption that signals are sparse in the time-frequency domain and do not overlap. Using the source location estimates, direction vectors towards each source were calculated. The direction vectors were used as the mean parameter of the mixing vector model. The parameters of the probabilistic models were estimated by the iterative expectation-maximization (EM) algorithm. The source location estimates were also utilized in the overall algorithm initialization. After a fixed number of iterations, soft time-frequency masks were obtained using the posterior probabilities of the combined models. The probabilistic time-frequency masks were applied to the reverberant mixtures to estimate the individual sources. Extensive experiments were performed to test the advantage of the known source directions on the separation. This

was done through comparisons with other competing methods in varying scenarios. The proposed method was found to be more efficient than others, specifically in multi-source highly reverberant cases and when sources are in close proximity. For instance, the proposed scheme when compared with other audio-only methods, on average at different model complexities using real room impulse responses (RIRs) with RT60 around 565 ms, in terms of SDR, performed 2.39 dB and 1.53 dB better than the method in [59] and [86] respectively. When compared with methods that estimate the source locations through video (audio-visual methods), the proposed method in the two-source case, performed on average over 5 dB better than three methods, [87–89], when they also utilized two microphone mixtures. Increasing the number of microphones in the competing methods improves their performance. But even with eight microphones, the proposed method (utilizing only two microphones) is approximately 2.7 dB, 1.5 dB, and 2.6 dB better than the method in [87], [88] and [89] respectively. Furthermore, in the three-source case, the proposed method provided an average advantage of around 2.8 dB over [87], 2.1 dB over [88], and 0.2 dB over [89], when all the competing methods use eight microphones.

To investigate the usefulness of the knowledge of the spatial characteristics of the enclosure such as the reverberation time and the wall reflective properties, Chapter 5 introduced the spatial covariance model. The spatial covariance model was evaluated by combining it with the models described in Chapter 4. Results highlighted that the complementary information about the spatial properties of the sources and the room can be useful in furthering the separation performance and mitigating the effects of reverberation, specifically when sources are relatively closely spaced. For example, considering the frequency-dependent models, the spatial covariance model combined with the models described in Chpater 4, termed as IIMM+SC in Chapter 5, improves performance approximately 2 dB over the IIM model that utilizes

only the ILD and IPD models.

Finally, in Chapter 6 a pre-processing stage was presented for two-channel dereverberation based on amplitude spectral subtraction. The single-channel spectral subtraction methods were reviewed. The single-channel method was extended to the binaural context, and was incorporated in to the source separation algorithm proposed in Chapter 4. Experimental results indicated that the pre-processing was useful in suppressing the late reverberant components before source separation, and provided improvement in the separation. A novel cascade structure to further suppress the late reverberation was investigated. Three dereverberation blocks were concatenated where each stage provided signal enhancement. Two state-of-the-art monaural spectral subtraction schemes were utilized and were extended to the binaural context. The cascade structure was experimentally evaluated in two different processing contexts. Firstly, it was used for the purpose of two-channel dereverberation only. Secondly, the cascade was used in conjunction with the two-channel source separation algorithm. Results from extensive experiments in both processing contexts demonstrated that the cascade structure gives an increased late reverberation suppression. The method is also beneficial when used as a pre-processing stage to source separation. The two-channel dereverberation scheme also preserved binaural cues which were exploited within the source separation algorithm. The proposed cascade when used solely for dereverberation utilizing real RIRs, at RT60 of 790 ms, provided a signal-to-noise ratio (SNR) improvement of 1.22 dB compared with 0.75 dB by the single stage method, segmental SNR improvement of 1.87 dB as compared to 1.14 dB, and a perceptual evaluation of speech quality (PESQ) improvement of 0.30 compared to 0.16 by the single stage method. When the cascade structure was utilized as pre-processing stage to the separation algorithm in Chapter 4, termed Cascade+IIMM, it gave on average around 3 dB improvement in terms of SDR for RT60s over

400 ms when compared with the separation algorithm with no pre-processing (IIMM). The PESQ results were also consistent where the Cascade+IIMM provided an improvement of 0.27 over the IIMM method.

7.1 Future Work

The techniques proposed in this thesis could be extended in a number of ways and different directions could be explored. The robust beamformer proposed in Chapter 3 was based on a linear array with sixteen microphones. Circular microphone array [106] or other geometries could be investigated. Also, in the post-processing stage when the binary masks were applied to a mixture, not all the sixteen mixtures were utilized. Either by some means of combining masks from all mixtures or something as simple as selecting, at each time-frequency point, the microphone with the higher estimated signal-to-noise ratio might further improve the performance.

The two-channel model-based approach could potentially be improved in a number of ways. The models of the interaural level and phase differences (ILD and IPD) and mixing vectors are combined assuming they are conditionally independent since this assumption offers particular advantage in algorithm development. Although there is some dependence between the ILD, IPD and the source directions (the parameter that aids the mixing vector model), it was not modeled in this work. Modeling such dependence is a very interesting point to be investigated. This dependence modeling is likely to further improve the quality of the time-frequency masks and thus the estimated sources. Another possibility is including a model for reverberation. The model should be capable of better distinguishing the direct-path sounds than the later reflections. A possibility to do this might be including a model for the precedence effect [107]. The precedence effect is a perceptual mechanism that aids humans to localize sounds in reverberant environments. The

model is expected to give a greater weighting to the direct-sound compared to the later reflections.

Pitch cues are utilized to group sound components within the time-frequency domain [37] [108] in order to segregate them. Efficiently modeling this monaural cue, and combining it with the model of binaural cues exploiting source directions (as proposed in Chapter 4 of this thesis) is also a potential direction for future research.

The combined models also assumed the sound sources to be physically stationary. In practice, however, the sources are likely to change their positions. The case of the moving sources will be explored. A potential solution for this might be in the context of audio-visual source separation. Visual tracking could be utilized and the models be fed with the source locations. However, synchronization of the audio and visual measurements may be a challenge.

The spatial covariance model assumed the reverberation time was known. Estimating the reverberation time was not focussed upon in this thesis but could be pursued in future. The spectral subtraction based dereverberation exploited state-of-the-art monaural algorithms that were extended into the binaural context. When the bilateral gains were combined to form a single gain, the weighting factor was chosen empirically. An efficient mechanism could to be devised for the determination of the weighting factor.

The cascade structure proposed here was based on three concatenated stages and exploited two different monaural methods. Different combinations and number of stages could be investigated.

Finally, future work could focus on reducing the current algorithm complexity. This would allow real-time implementation and its utilization in multiple application fields.

References

[1] T. Kim, H. Attias, S. Lee, and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, 2007.

[2] M. S. Pedersen, D. Liang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures," *IEEE Transactions on Neural Networks*, vol. 19, no. 3, pp. 475–492, 2008.

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

[4] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3100–3115, 2005.

[5] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[6] H. Saruwatari, Y. Mori, T. Takatani, S. Ukai, K. Shikano, T. Heikata, and T. Morita, "Two-stage source separation based on ICA and binary masking for real-time robot audition system," *in Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2005, Edmont, Alberta, Canada.*, pp. 2303–2308, 2005.

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

[8] D. Wang and G. Brown, "Fundamentals of computational auditory scene analysis, in computational auditory scene analysis: Principles, algorithms and applications," *Hoboken, NJ: John Wiley and Sons, 144*, 2006.

[9] A. Hyvrinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[10] N. Roman and D. Wang, "Binaural sound separation for multisource reverberant environments," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing, Montreal, Quebec, Canada*, pp. 373–376, 2004.

[11] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[12] M. S. Khan, S. M. Naqvi, A.-Rehman, W. Wang, and J. A. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1900–1912, 2013.

[13] A. Rehman, S. M. Naqvi, W. R. Phan, and J. A. Chambers, "MCMC-PF based multiple head tracking in a room environment," *4th UK Computer Vision Student Workshop (BMVW)*, 2012.

[14] M. I. Mandel, "Binaural model-based source separation and localization," *PhD Thesis, Columbia Univeristy, USA*, 2010.

[15] H. Kuttruff, "Room acoustics," *Spon Press, Oxon*, 2009.

[16] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," *PhD Thesis, University of Surrey, UK*, 2011.

[17] P. A. Naylor and N. D. G. (Eds.), *Speech Dereverberation, Signals and Communication Technology*. Springer, 1st Edition, 2010.

[18] M. Jeub, "Joint dereverberation and noise reduction for binaural hearing aids and mobile phones," *PhD Thesis, RWTH Aachen University, Germany*, 2012.

[19] A. J. Watkins, "Perceptual compensation for effects of reverberation in speech identification," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 249–262, 2005.

[20] T. Kim, H. T. Attias, S. Lee, and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

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

[22] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech.," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1W.html,"

[24] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time frequency masking," *in Proc. IWAENC 2006, Paris*, 2006.

[25] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[26] T. Melia, "Underdetermined blind source separation in echoic environments using linear arrays and sparse representations," *PhD Thesis, University College Dublin, National University of Ireland, 2007.*

[27] M. A. Dmour, "Mixture of beamformers for speech separation and extraction," *PhD Thesis, University of Edinburgh, UK*, 2010.

[28] S. M. Naqvi, "Multimodal methods for blind source separation of audio sources," *PhD Thesis, Loughborough University, UK*, 2009.

[29] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2000.

[30] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 881–884, 2002.

[31] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Transactions on Speech and Audio Processing*, vol. 13(01), 2005.

[32] S. A. H. Sawada, R. Mukai and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source sepa-

ration," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[33] T. Kim, H. Attias, S. Lee, and T. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech and Language processing*, vol. 15, pp. 70–79, 2007.

[34] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4–24, 1988.

[35] S. U. Pillai, "Array signal processing," 1989.

[36] A. S. Bregman, "Auditory scene analysis: The perceptual organization of sound," *The MIT Press*, 1990.

[37] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[38] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Sigal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[39] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, no. 114, pp. 2236–2252, 2003.

[40] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 529–532, 2002.

[41] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," *in Proc. of ICA'2004*, pp. 832–839, 2004.

[42] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdeternined blind separation of speech in real environments with sparseness

and ICA," *in Proc. of the IEEE conference on acoustics, speech and signal processing, vol. 3.*, pp. 881–884, 2004.

[43] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdeternined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," *Independent component analysis and blind signal separation: Proc. of the fifth international congress, ICA 2004, Berlin: Springer.*, pp. 898–905, 2004.

[44] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. on Audio, Speech and Lang. Processing.*, vol. 14, no. 6, 2006.

[45] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdeternined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing 87.*, pp. 1833–1847, 2007.

[46] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Overcomplete blind source separation by combining ICA and binary time-frequency masking," *In proc. of IEEE International Workshop on machine learning for Signal Processing, Mystic,CT,USA.*, pp. 15–20, 2005.

[47] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixutres," *IEEE Transactions on Neural Networks*, vol. 19, no. 3, 2008.

[48] D. Kolossa, R. F. Astudillo, E. Hoffmann, and R. Orglmeister, "Independent component analysis and time-frequency masking for speech recognition in multitalker conditions," *EURASIP Journal on Audio, Speech and Music Proc.*, 2010.

[49] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation

alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 516–527, 2011.

[50] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Communication 53.*, pp. 524–539, 2011.

[51] N. Roman, S. Srinivasan, and D. Wang, "Binaural segregation in multisource reverberant envirinments," *Journal of the Acoustical Society of America*, no. 120, pp. 4040–4051, 2006.

[52] J. B. Boldt, U. Kjems, M. S. Pedersen, T. Lunner, and D. Wang, "Estimation of ideal binary mask using directional systems," *In Proc. of 11th Intern. Workshop on Acoustics Echo and Noise Control, Seattle, WA , USA*, 2008.

[53] J. Beh, T. Lee, D. Han, and H. Ko, "Sound source separation by using matched beamforming and time-frequency masking," *The 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan.*

[54] P. Aarabi, G. Shi, and O. Jahromi, "Robust speech separation using time-frequency masking," *ICME '03 Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2.*, pp. 741–744, 2004.

[55] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. on Sys., Man. and Cybernetics- part B: Cybernetics.*, vol. 34, no. 4, pp. 109–118, 2004.

[56] Y. Takenouchi and N. Hamada, "Time-frequency masking for bss problem using equilateral triangle microphone array," *In the Proc. of the 2005 International Symposium in Intel. Sig. Processing and Comm. Systems.pp.185-188, Hong Kong*, 2005.

[57] J. Cermak, S. Araki, H. Sawada, and S. Makino, "Blind source separation based on a beamforming array and time frequency binary masking," *In the proc. of the 10th international workshop acoustic echo and noise cancellation, Paris, France. September*, 2006.

[58] J. Mouba and S. Marchand, "A source localization/separation/respatialization system based on unsupervised classification of interaural cues," *in Proc. Digital Audio Effects (DAFx06) Conference, Montreal, Canada*, pp. 233–238, 2006.

[59] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[60] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[61] A.ur-Rehman, S. Naqvi, R. Phan, and J. Chambers, "Multispeaker direction of arrival tracking for multimodal source separation of moving sources," *in Proc. Sensor Signal Processing for Defence (SSPD 2011)*, pp. 1–5, 2011.

[62] K. Kondo, "Chapter 2, subjective quality measurement of speech: Its evlauation, estimation and applications," *Springer*, 2012.

[63] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

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

[65] M. I. Mandel, S. Bressler, B. S.-Cunningham, and D. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Transactions on Audio, Speech and Language processing*, vol. 18, no. 7, pp. 1872–1883, 2010.

[66] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," pp. 2819–2822, 1998.

[67] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications.* John Wiley, 2002.

[68] M. G. Jafari, "Novel sequential algorithms for blind source separation of instantaneous mixtures," *PhD thesis, King's College London*, 2002.

[69] P. Loizou, "Speech enhancement: Theory and practice," *Boca Raton, FL: CRC*, 2007.

[70] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing, Taipei, Taiwan*, 2009.

[71] L. C. Parra, "Steerable frequency-invarient beamforming for arbitrary arrays," *Journal of the Acoustical Society of America*, vol. 6, pp. 3839–3847, 2006.

[72] H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing.* John Wiley and Sons, Inc., 2002.

[73] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

[74] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Communication*, vol. 53, pp. 524–539, 2011.

[75] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing for spectral masks in the cepstral domain for speech separation," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 45–48, 2008.

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

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

[78] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measuremet in Blind Audio Source Separation," *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 1462–1469, 2006 /[http://sisec2010.wiki.irisa.fr/tiki-index.php].

[79] D. Ellis, "Chapter 1 evaluating speech separation systems," 2004.

[80] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Sawada, "The fundamental limitation of frequency domain blind source separtion for convolutive mixtures of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[81] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *Journal of the Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[82] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[83] M. I. Mandel, "Binaural model-based source separation and localization," *PhD thesis, Columbia University*, 2010.

[84] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," *in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY*, 2007.

[85] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," *in Proc. ICA 2004, ser. Lecture Notes in Computer Science, Springer-Verlag*, pp. 430–436, 2004.

[86] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209–212, 2011.

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

[88] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2257–2269, 2007.

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

[90] D. Sodoyer, J. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: a new approach exploiting the audio-

visual coherence of speech stimuli," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1165–1173, 2002.

[91] D. Sodoyer, L. Girin, C. Jutten, and J. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Communication*, vol. 44, no. 1-4, pp. 113–125, 2004.

[92] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 96–108, 2007.

[93] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.

[94] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 9–12, 2010.

[95] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 791–803, 2003.

[96] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial covariance models for under-determined reverberant audio source separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 129–132, 2009.

[97] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[98] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[99] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1732–1745, 2010.

[100] A. Tsilfidis, E. Georganti, and J. Mourjopoulos, "Binaural extension and performance of single-channel spectral subtraction dereverberation algorithms," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1737–1740, 2011.

[101] M. S. Khan, S. M. Naqvi, and J. A. Chambers, "A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation," *in Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2013.

[102] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4409–4412, 2009.

[103] C. Hummersone, "Binaural room impulse responses (BRIRs)," *University of Surrey, UK, 2010. Available: http://www.surrey.ac.uk/msr/people/chris-hummersone/BRIRs.*

[104] D. P. Mandic and J. A. Chambers, "Recurrent neural networks for prediction," *Wiley Series in Adaptive and Learning Systems for Signal Processing, Communication, and Control, chapter 7*, 2001.

[105] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Proc. 16th International Conference on Digital Signal Processing*, pp. 1–5, 2009.

[106] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *Journal of the Acoustical Society of America*, no. 120 (5), 2006.

[107] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1867–1871, 2010.

[108] J. Woodruff, "Integrating monaural and binaural cues for sound localization and segregation in reverberant environments," *PhD Thesis, Ohio State University, USA*, 2012.