Mixture of Beamformers for Speech Separation and Extraction

Mohammad A. Dmour

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
2010

Abstract

In many audio applications, the signal of interest is corrupted by acoustic background noise, interference, and reverberation. The presence of these contaminations can significantly degrade the quality and intelligibility of the audio signal. This makes it important to develop signal processing methods that can separate the competing sources and extract a source of interest. The estimated signals may then be either directly listened to, transmitted, or further processed, giving rise to a wide range of applications such as hearing aids, noise-cancelling headphones, human-computer interaction, surveillance, and hands-free telephony.

Many of the existing approaches to speech separation/extraction relied on beamforming techniques. These techniques approach the problem from a spatial point of view; a microphone array is used to form a spatial filter which can extract a signal from a specific direction and reduce the contamination of signals from other directions. However, when there are fewer microphones than sources (the underdetermined case), perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible.

In this thesis, we present a framework which extends the use of beamforming techniques to underdetermined speech mixtures. We describe frequency domain non-linear mixture of beamformers that can extract a speech source from a known direction. Our approach models the data in each frequency bin via Gaussian mixture distributions, which can be learned using the expectation maximization algorithm. The model learning is performed using the observed mixture signals only, and no prior training is required. The signal estimator comprises of a set of minimum mean square error (MMSE), minimum variance distortionless response (MVDR), or minimum power distortionless response (MPDR) beamformers. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the estimated posterior probabilities of the Gaussian mixture states. These weights are specific to each time-frequency point. The resulting non-linear beamformers do not need to know or estimate the number of sources, and can be applied to microphone arrays with two or more microphones with arbitrary array configuration. We test and evaluate the described methods on underdetermined speech mixtures. Experimental results for the non-linear beamformers in underdetermined mixtures with room reverberation confirm their capability to successfully extract speech sources.

To my parents

Declaration of Originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the School of Engineering at The University of Edinburgh.

Mohammad A. Dmour

Acknowledgements

At the very outset, I would like to thank Professor Mike Davies, my supervisor at the Institute for Digital Communications, University of Edinburgh, for his unflagging guidance and encouragement, without which this research work would not have reached the point of fruition. I would also like to thank Dr James Hopgood for the many interesting conversations and discussions, and all the help in the audio lab. I am also grateful to Massimiliano Tonelli for his advice on measuring reverberation time. I would also like to thank Peter Frith and Dr Anthony Magrath from Wolfson Microelectronics for the inspiring discussions I had with them.

I would like to thank my friends for their support and active encouragement, and also for sharing many unforgettable moments with me.

Moreover, I would like to acknowledge the financial support of the Wolfson Microelectronics Scholarship and the Institute for Digital Communications, University of Edinburgh.

No acknowledgement would be complete without expressing my deepest gratitude to my parents who have been a source of inspiration to me for years. They instilled in me a sense of morality, right and wrong, and encouraged me every step of the way. Without this, I would never have got to where I have reached on this day. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

ASA	auditory scene analysis
BSS	blind source separation
CASA	computational auditory scene analysis
DOA	direction of arrival
DUET	degenerate unmixing estimation technique (source separation algorithm)
EM	expectation maximisation
ENSIR	energy-normalised source to interference ratio
ERB	equivalent rectangular bandwidth
FIR	finite impulse response
FMV	frequency-domain minimum-variance (source extraction algorithm)
GMM	Gaussian mixture model
GSMM	Gaussian scaled mixture model
HMM	hidden Markov model
ICA	Independent component analysis
IID	inter-channel intensity difference
ILD	inter-channel level difference
IPD	inter-channel phase difference
ISTFT	inverse short-time Fourier transform
LCMV	linear constrained minimum variance
LMS	least mean squares
MAP	maximum a posteriori
MENUET	multiple sensor DUET (source separation algorithm)
MMSE	minimum mean square error
MoGs	mixture of Gaussians
MPDR	minimum power distortionless response
MVDR	minimum variance distortionless response
PSD	power spectrum density
RT	reverberation time
SAR	sources to artifacts ratio

SDR	signal to distortion ratio
SIR	source to interference ratio
STFT	short-time Fourier transform
WER	word error rate

Nomenclature

$\mathbf{a}$	array manifold
$\mathbf{A}$	mixing matrix
B	number of beamformers in the beamformer array method
c	speed of sound
c_{s,q_s}	prior probability of state q_s of a desired source GMM
c_{v,q_v}	prior probability of state q_v of an interference signal GMM model
c_{x,q_x}	prior probability of state q_x of a mixture signal GMM model
d	spacing between microphones
d_{max}	maximum separation between the reference microphone I and any other microphone
$\det$	matrix determinant
D	total array size
D_c	complete data (in EM algorithm)
exp	exponential function
E[.]	expectation
f	frequency
$G(.)$	Gaussian function
I	index of reference microphone
k_s	number of states in the desired source GMM
k_v	number of states in the interference GMM
k_x	number of states/clusters in the mixture model
l_c	log likelihood of the complete data
l	iteration number
ln	logarithm to the base e
$\log_{10}$	logarithm to the base 10
L	length of allowed filtering distortions (in performance evaluation)
M	number of sources
$\mathbf{M}$	time-frequency mask
n	frame number
N	number of microphones

$p(.)$	probability
P	number of paths between source and microphone
Q	number of PSD templates (states) in a spectral GMM
q	index for a state or a cluster
r_i	distance of a source to microphone i
r_{jqf}	PSD template for source j, state q, and frequency f
$\mathbf{R}_v$	covariance matrix of an interference signal
$\mathbf{R}_x$	covariance matrix of a mixture signal
$\mathbf{R}_{v,q_v}$	covariance matrix of state q_v of an interference signal GMM
$\mathbf{R}_{x,q_x}$	covariance matrix of state/cluster q_x of a mixture signal model
s	source signal
t	time index
Tr	Trace function
$\mathbf{u}$	input mixture signal (in beamformer array method)
v	interference signal
V	PSD
$\mathbf{w}$	beamformer (row vector)
$\mathbf{w}_1$	mixture of MMSE beamformers (row vector)
$\mathbf{w}_2$	mixture of MVDR beamformers (row vector)
$\mathbf{w}_3$	mixture of MPDR beamformers (row vector)
$\mathbf{w}_4$	mixture of MPDR beamformers using binary masks (row vector)
$\mathbf{w}_5$	mixture of MPDR beamformers using soft masks (row vector)
x	mixture signal
$\mathbf{y}$	vector signal representing each cluster (in $\mathbf{w}_4$ and $\mathbf{w}_5$)
Z	number of interferes selected in each input mixture (in beamformer array method)
α_i	ratio between attenuation of the source at microphone i and reference microphone I
$\delta(.)$	Dirac delta function
Δ_i	delay between microphone i and reference microphone I for a source signal
η	number of time frames in each frequency bin
θ	parameters of the GMM
ι	$\sqrt{-1}$
σ_s^2	variance of a source signal
σ_{s,q_s}^2	variance of state q_s of a source signal GMM

τ	posterior probability
ϕ	direction of arrival angles
$\Re$	real part of a complex number
$\hat{(.)}$	estimate
$(.)^*$	complex conjugate
$(.)^H$	matrix conjugate transpose
$(.)!$	factorial
$\lvert.\rvert$	magnitude (or modulus or absolute value) of a complex number
$\lVert.\rVert$	L_2 norm
$\angle(.)$	phase of a complex number
$\prod$	product
$\forall$	for all
$\star$	convolutive operator

Chapter 1
Introduction

1.1 The Cocktail Party Problem

Most audio signals result from the mixing of several sound sources. In many applications, there is a need to separate the multiple sources or extract a source of interest while reducing undesired interfering signals and noise. The estimated signals may then be either directly listened to or further processed, giving rise to a wide range of applications such as hearing aids, human-computer interaction, surveillance, and hands-free telephony.

Source mixing can occur in a wide variety of situations under different environments. The difficulty of source separation and extraction depends on the number of sources, the number of microphones and their arrangements, the noise level, the way the source signals are mixed within the environment, and on the prior information about the sources, microphones, and mixing parameters. Blind methods do not rely on specific characteristics of the sources, microphones, or mixing parameters. By contrast, informed methods exploit some prior information about the sources and microphones (for example, the location of a desired source, the identity of a musical instrument). In general, the problem is more difficult when the reverberation time of the acoustic environment is large, and when the number of sources is larger than the number of microphones.

The classical example of source separation and extraction is the "cocktail party problem", where different conversations occur simultaneously and independently of each other. The human auditory system exhibits a remarkable ability to follow only one conversation in a highly noisy environment, such as a cocktail party. This was first analysed and termed the cocktail party problem by Colin Cherry [1]. In 1953, Cherry reported on experiments performed on the recognition of speech received by one and two ears [1]. He proposed a few factors that can be used by a machine to recognise what a person is saying when multiple speakers are talking simultaneously: (a) Voices come from different directions. (b) Lip movements and gestures. (c) Different speakers have different voices, pitches, and speeds. (d) Different speakers have different accents. (e) Predicting word sequences.

Today, many different methods exist in order to enhance speech. There has been a lot of research on the speech enhancement problem, where the focus is on attenuating the background noise. Speech denoising algorithms are well established and have been used for many years [2, 3]. The extension of the speech enhancement problem to deal with mixtures of speech sources is a topic of intense research [4, 5].

One approach to address the source separation problem is to study and finally replicate the way humans perform audio source separation using a computer. Based on psychoacoustic experiments, it is believed that the human auditory system decomposes the acoustic signal into elements, where each element describes a significant acoustic event [6]. This decomposition is followed by a grouping process which combines elements that are likely to have originated from the same audio source [6]. This process is termed auditory scene analysis (ASA). Grouping processes may be data-driven (primitive), or schema-driven (knowledge-based) [7]. Data-driven grouping cues include energy appearing across different frequencies at the same time (common-onset), energy ceasing across different frequencies at the same time (common-offset), harmonics of the same fundamental frequency, similar inter-channel time and intensity differences (spatial cues), smooth spectral envelope (continuity), and correlated amplitude and frequency modulations. Schema-driven grouping rules implement knowledge acquired by learning, such as the voice of a known speaker or the syntax of a particular language.

Computational auditory scene analysis (CASA) refers to the set of algorithms developed with the aim of simulating auditory scene analysis processes [8–10]. The first stage of a CASA system is usually a filter bank that replicates the time-frequency analysis of the human ear [11, 12]. Most CASA systems further process the time-frequency representation in order to facilitate the extraction of certain grouping features. Grouping rules are then applied to group components that are likely to have originated from the same source. The desired source signal is eventually extracted by applying a weight to each time-frequency point, such that points believed to be dominated by the desired source receive a high weight [8]. Finally, the time-frequency representation is inverted in order to recover a time-domain estimate of the separated source. This allows the separated signals to be evaluated.

An alternative approach to speech source separation and extraction is to rely on adaptive beamforming techniques [13, 14]. These techniques approach the separation problem from a spatial point of view; a microphone array is used to form a spatial filter which can extract a desired signal from a specific direction and reduce interfering signals from other directions. This spatial filter can be expressed in terms of dependence upon angle and frequency. Beamforming is accomplished by filtering the microphone signals and combining the outputs to extract (by constructive combining) the desired signal and reject (by destructive combining) interfering signals according to their spatial location. Beamforming can separate sources with overlapping frequency content that originate at different spatial locations. The performance of beamforming techniques is often very good when the number of microphones is large. However, when the number of sources is larger than the number of microphones, perfect attenuation of all interferers becomes impossible using time-invariant beamforming techniques and only partial interference attenuation is possible. This is because in time-invariant beamforming, the number of directions of arrival that can be perfectly cancelled is limited by the number of microphones [13]. In general, beamforming techniques require information about the microphone array configuration and the sources (for example, the direction of the desired source).

Sometimes, the knowledge of the microphone array configuration and the sources positions is not available, and the only available data are the mixtures of the different sources recorded at the microphones. Methods which can obtain estimates of the sources using only the mixture signal, and with no information about the mixing process are termed blind source separation (BSS) methods. One approach is to use statistical modeling of source signals. Independent component analysis (ICA) is one of the major statistical tools used in BSS [15–18]. In ICA, separation is performed on the assumption that the source signals are statistically independent, and does not require information on microphone array configuration or the direction of arrival (DOA) of the source signals to be available. To perform source separation, we process the mixture channels by a set of linear time-invariant demixing filters. ICA implicitly estimates the source directions by maximising the independence of the sources, and acts as an adaptive null beamformer that reduces the undesired sources. However, some aspects limit the application of ICA to real-world environments. Most ICA methods assume the number of sources is given a priori. In general, classical ICA techniques cannot perform source separation when the number of sources is larger than the number of microphones.

Source separation methods initially only considered the case when the number of sources is equal to or less than the number of microphones. In this case, the problem is one of identifying the mixing matrix and the source estimates are simply obtained by inverting the mixing matrix. When the number of sources is larger than the number of microphones, linear source separation using the inverse of the mixing matrix is not possible. However, under certain assumptions, it is possible to extract a larger number of sources by using non-linear filtering methods. The assumption that the sources have a sparse representation under an appropriate transform is a very popular assumption. Sparseness of a signal means that only a small number of the source components differ significantly from zero. A sparse representation of a speech signal can be achieved by a short-time Fourier transform (STFT). One popular approach to sparsity-based separation is time-frequency masking [19–21]. This approach is a special case of non-linear time-varying filtering that estimates the desired source from a mixture signal by applying a time-frequency mask that attenuates time-frequency points associated with interfering signals while preserving time-frequency points where the signal of interest is dominant. If more than one mixture signal is available, the spatial information at each time-frequency point can sometimes be used to determine which time-frequency points belong to each source. A popular method to estimate the time-frequency masks using only two microphones is the degenerate unmixing estimation technique (DUET) [19, 20]. It is assumed that the time-frequency representation of speech signals are approximately disjoint (i.e., sources do not overlap too much). This assumption is not fully met in practice. In DUET, the sources are estimated by partitioning the mixture STFT coefficients based on the inter-channel level/phase difference. In general, time-frequency masking is capable of performing source separation when the number of sources is larger than the number of microphones. However, this method suffers from so-called musical noise or burbling artifacts due to masking of time-frequency points where the sources overlap. These distortions are introduced by the separation mask, and arise from the spectral components of the signal being turned on and off, which results in sinusoidal components that come and go in each short-time frame. Furthermore, separation methods based on time-frequency masking suffer from the fact that clustering becomes difficult in reverberation, as the inter-channel level/phase difference resulting from each sound source then tend to spread and overlap, and the disjoint assumption becomes less realistic.

When only one microphone is available, source separation/extraction becomes significantly more challenging, as spatial cues are absent in this case. In this situation, the assumptions of independence and time-frequency sparsity becomes insufficient, and more advanced source models relying on spectro-temporal models are needed. These models have been extensively studied and used in speech recognition. Different strategies have been employed using these models [22–24]. However, they require prior training and some knowledge about the identity of the speech or music sources in the mixture.

1.2 Applications of Audio Source Separation

There are many applications where speech source separation may be useful. We briefly discuss a selection of these applications:

Hearing Aids

The human auditory system exhibits an extraordinary ability to extract sounds of interest when multiple speakers are talking simultaneously. However, hearing impaired people loose the ability to extract desired sound sources and thus the ability to follow conversations. Hearing aids with source separation and extraction capabilities can assist the hearing impaired to select sounds and conversations in multi-talker scenarios. The hearing aid application presents particular challenges due to the requirements of real-time processing, low power consumption, small size and low weight.

Transmitter Noise Cancellation

A problem that occurs when we transmit audio signals via communication devices such as mobile phones or teleconferencing equipments, is that many visual and auditory spatial cues that are present on the transmitter side are not transmitted, making the extraction of the desired source more difficult. Therefore, it is desirable to be able to clean the audio signal before transmission. Traditional active noise cancellation methods work well for noises that are continuous, such as the noise generated by a car engine, but are less effective against non-stationary interference such as speech or other rapidly changing signals. Source separation and extraction methods can be used to suppress noise and interference before signal transmission.

Human-Computer Interaction

It is well known that automatic speech recognition accuracy can be severely degraded by background noise, reverberation and competing speakers. This problem is of particular concern if the distance between the talker and the microphone(s) is large. Source separation and extraction can be used as a pre-processing step for speech recognition in noisy and multi-talker environments. In this application, it would be advantageous to use video data captured by a camera to estimate the location of the desired sources [25, 26].

Surveillance

Source separation and extraction can be used in surveillance and forensic analysis of recordings that contain multiple simultaneous talkers to extract desired sources, carry out automatic keyword detection, or detecting whether a targeted speaker is present in a conversation.

Smart Meeting Rooms

Close-talk microphones have been traditionally used in teleconferencing rooms, as they provide a higher signal to noise ratio. However, they require each user to wear a microphone, or limit the movement of the speakers. In conjunction with microphone arrays, source separation and extraction can be used to meeting rooms to allow for hands free operation, automatic transcription of speech, and selective transmission of multiple speech sounds in teleconferences.

1.3 Thesis Overview

This work focuses on the separation of audio signals, and more specifically of speech signals. The main motivation of this work is to develop new solutions to the problem of source separation and extraction methods for mixtures where the number of sources is larger than the number of microphones. These mixtures are termed underdetermined mixtures. The problem is related to the problem of solving a system of linear equations when there are more unknowns than equations. This underdetermined problem is a challenging problem and it does not have a unique solution without additional constraints and prior information. As previously discussed in this chapter, traditional source separation techniques such as ICA and linear beamforming are not suited for underdetermined mixtures.

The main aim of this work is a framework which extends the use of beamforming techniques to underdetermined speech mixtures. We describe frequency domain non-linear mixture of beamformers that can extract a speech source from a known direction when there are fewer microphones than sources, and do not require knowledge of the number of speakers. In this framework, we introduce additional degrees of freedom to the beamformer by exploiting the super-Gaussianity and the sparsity of the speech signals in the time-frequency domain[1], and

[1] Due to the combination of the non-stationarity and harmonic content, speech signals have their energy concentrated in isolated regions in time and frequency, and most of the coefficients do not differ significantly from zero (sparse). Therefore, speech signals exhibit a zero mean super-Gaussian probability distribution, which has a sharper

dynamically finding suitable directivity patterns in order to reduce active interfering signals.

The work presented in this thesis is structured as follows:

Chapter 2 is an overview of source separation and extraction principles and methods. We describe the different possibilities of source separation and extraction environments and the potentially available prior knowledge of the sources, mixing process, and the microphones. We then introduce some of the properties of speech signals which are important in understanding why particular models and methods are used in speech separation and extraction. This is followed by a description of room acoustics. The different scenarios where audio mixtures can be obtained are discussed and we describe the instantaneous, anechoic, and echoic mixing models. We then discuss performance measures that can be used to evaluate source separation and extraction methods. Finally, the main existing source separation and extraction methods are discussed.

In **Chapter 3**, we exploit the speech's sparsity in the time-frequency domain in order to extend the use of beamforming techniques to underdetermined speech mixtures. We use Gaussian mixture models (GMMs) to model the speech non-Gaussianity and the spatial distribution of the sources. We present three frequency domain non-linear beamformers that can extract a desired source from a known direction. The first two non-linear beamformers are based on modeling the desired source signal and the interference separately. The desired source signal is modeled using a 1-dimensional GMM, and the observed interference is modeled using an N-dimensional GMM, where N is the number of microphones. The third non-linear beamformer is based on modeling the observed mixture signal (the desired source and interference together) using an N-dimensional GMM. In contrast to other speech enhancement and separation methods which use GMMs such as [24, 28, 29], our approach does not couple the Gaussian states across frequency, and the covariance matrices of each Gaussian state represent a spatial covariance matrix. To learn the GMMs parameters, we use the expectation maximisation algorithm [30]. The model learning is performed using the observed mixture signals only, and no prior training is required. Based on these models, we develop three non-linear beamformers. The signal estimator in these beamformers comprises of a set of minimum mean square error (MMSE), minimum variance distortionless response (MVDR), or minimum power distortionless response (MPDR) beamformers. In order to estimate the desired signal, all beamformers are concurrently applied to the observed signal, and a weighted sum of the beamformers' out-

peak and longer tails than the Gaussian probability distribution [27].

puts is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. These weights are specific to each time frequency point, and have a non-linear dependency on the observed data. The resulting non-linear beamformers combine the benefits of non-linear time-varying separation in time-frequency masking with the benefits of spatial filtering in the linear beamformers. We then present some simulation results that illustrate the performance of the proposed methods. More experiments and comparisons of the proposed methods with other source separation algorithms can be found in Chapter 5. The chapter is followed by two appendices where the detailed derivations of the GMM learning rules using the EM algorithm are presented.

In **Chapter 4**, we present a modification to the mixture of MPDR beamformers presented in the previous chapter. The algorithm presented in this chapter combines time-frequency masking techniques and mixture of beamformers. The proposed algorithm has two main stages. In the first stage, the mixture time-frequency points are partitioned into a sufficient number of clusters using time-frequency masking techniques. In the second stage, we use the clusters obtained in the first stage to calculate covariance matrices, one for each cluster in each frequency bin. These covariance matrices and the time-frequency masks are then used in a mixture of MPDR beamformers. The resulting non-linear beamformer has low computational complexity and removes the musical noise found in time-frequency masked outputs at the expense of lower interference attenuation. The mixture of MPDR beamformers stage can be regarded as a post-processing step for sources separated by time-frequency masking. Two variants of the proposed method are described and compared. The first one uses binary time-frequency masks, and the second one uses soft (real-valued) time-frequency masks. We then present some simulation results that illustrate the performance of proposed methods. More experiments and comparisons of the proposed methods with other source separation algorithms can be found in Chapter 5.

In **Chapter 5**, we investigate the effect of the mismatch between the assumed DOA of the desired source and the true one. We incorporate different methods developed to enhance the robustness of beamforming techniques against DOA mismatch in the mixture of beamformers framework, and study their effect. Finally, we compare the proposed non-linear beamformers with some other source separation and extraction methods in several reverberant rooms.

Finally, in **Chapter 6**, we summarise the results and contributions, and propose several directions for future work.

1.4 Publications

Associated with this work are the following publications:

Journal Article:

- M.A. Dmour, M.E. Davies; "A new framework for underdetermined speech extraction using mixture of beamformers", to appear in IEEE Transactions on Audio, Speech and Language Processing.

Conferences:

- M.A. Dmour, M.E. Davies; "An approach to under-determined speech separation based on a non-linear mixture of beamformers", European Conference on Signal Processing (EUSIPCO), 2009.
- M.A. Dmour, M.E. Davies; "Under-determined speech separation using GMM-based non-linear beamforming", European Conference on Signal Processing (EUSIPCO), 2008.
- M.E. Davies, M.A. Dmour; "A nonlinear frequency-domain beamformer for underdetermined speech mixtures", invited talk at Acoustics'08. The Journal of the Acoustical Society of America, vol. 123, issue 5, p. 3586, 2008.

Chapter 2
Audio Source Separation and Extraction: Overview and Principles

2.1 Audio Source Separation and Extraction

Source mixing problems emerge in a wide variety of signal processing applications. Most audio, radio, seismic, sonar, video, and biomedical signals are mixtures of several sources that are simultaneously active. In general, observations are obtained at the output of a set of sensors, each receiving different combinations of the source signals. In certain applications, we aim to decompose the mixture signal into the original source signals (source separation). In many practical applications, however, prior information about a desired source, such as source location or identity, might be available and exploited to extract only one source of interest while reducing undesired interfering signals and noise (source extraction).

In this research, we will concentrate on audio source separation and extraction for speech applications. There are many applications where speech source separation and extraction may be useful; such as in hearing aids, noise-cancelling headphones, human-computer interaction, teleconferencing, and surveillance.

Audio source mixing can occur in a wide variety of situations under different environments. The difficulty of source separation and extraction depends on the way the source signals are mixed within the environment and on the a priori knowledge of the sources, microphones, and mixing parameters. The number and type of the assumptions used to perform source separation and extraction varies depending on the model used. Blind methods do not use any training data and do not assume a priori knowledge of sources, microphones, or mixing parameters. By contrast, informed methods exploit some prior information about the sources and microphones (for example, their location). Most existing source separation and extraction techniques require prior information or assume probability models for the sources and/or the mixing process.

A very important factor affecting the separation difficulty is the length of the mixing channel impulse response. The simplest case is instantaneous mixing, where each source signal appears at all the mixture channels at the same time with differing intensity. The anechoic or delayed case is similar to the instantaneous case, but each source signal reach each microphone with a different delay. When we have multiple paths between each source and each microphone, the mixing is termed echoic or reverberant. Source separation and extraction is more difficult in the echoic case as each source signal arrives at each microphone from multiple directions at different times, and we need to remove multipath interference. Furthermore, source separation and extraction is more difficult if the sources are moving as this will lead to the mixing channel changing with time.

Another important factor influencing the complexity of source separation and extraction is the ratio of the number of sources to the number of microphones. A mixture is termed a determined mixture when the number of microphones is equal to the number of sources, overdetermined when the number of microphones is larger than the number of sources, and underdetermined when it is smaller. Source separation and extraction is more difficult when the number of microphones is smaller than the number of sources, as in this case, even if the mixing channel is identified, estimating the sources is not a trivial task and requires a priori knowledge on the sources. When only one microphone is available, source separation and extraction becomes much more challenging, as in this case spatial cues are absent. In this situation, more advanced source models relying on spectro-temporal cues are needed to make any separation feasible.

The amount of a priori knowledge can also influence the complexity of the source separation and extraction problem. Most methods assume the knowledge of certain information. Commonly assumed a priori knowledge include microphone geometry, source geometry, number of sources, type of sources (speech, music,...etc) and channel parameters.

Table 2.1 summarises the different possibilities of source separation and extraction environments and a priori knowledge.

2.2 Speech Signals

Before starting to describe speech separation and extraction methods, it is vital that we become familiarised with the speech production process and the speech signal. In this section, we introduce some of the properties of speech signals which are important in understanding why particular models and methods are used in speech separation.

mixing channel	• instantaneous/anechoic/echoic • known/unknown/partial knowledge • static/changing
sources/microphones ratio	• overdetermined • determined • underdetermined • single microphone
number of sources	• known/unknown • variable
source location	• static/moving • known/unknown
type of sources	• speech/music/other • point/spatially-spread source
microphone array geometry	• known/unknown • linear/planar/volumetric

Table 2.1: *Source separation and extraction categories.*

Speech is generated when the vocal tract is excited and is composed of two types: voiced and unvoiced. Voiced sounds, which includes the vowels and some consonants such as B, D, L, M, N, and R, are pronounced with an open vocal tract excited by a pulsating airflow resulting from the vibration of the vocal cords (or vocal folds). This excitation have spectral peaks at the harmonics of the speaker's fundamental frequency or pitch frequency. Males typically have a lower fundamental frequency than females because their vocal cords are longer and more massive [31]. In unvoiced sounds, which includes consonants such as F, S, and SH, the vocal cords do not produce a periodic output. Unvoiced sounds are pronounced when the air flow is constricted by the tongue, lips, and/or teeth, resulting in air turbulence.

Both voiced and unvoiced sound sources are modified by the acoustic cavities of the vocal tract formed from the tongue, lips, mouth, throat, and nose. The vocal tract acts as a resonant cavity, and the resonances are known as formants. Different sounds are generated by altering the size and shape of the vocal tract resulting in different formants.

2.2.1 Time-Frequency Representation

Time-frequency representation is an effective tool for processing speech signals whose frequency content changes in time. Time-frequency representation provides the distribution of signal energy versus time and frequency. There are many time-frequency representations available [32–34]. We discuss below some of the most popular ones in speech processing.

Short-Time Fourier Transform (STFT)

A popular method for time-frequency representation is the short-time Fourier transform (STFT). In the STFT, the time domain signal $s(t)$ sampled at frequency f_s is sectioned into short, windowed, and overlapping frames and the discrete Fourier transform is used to find the frequency spectrum of each frame. The STFT of the signal $s(t)$ in time frame n and at frequency f, with an F-sample frame and H-sample shift (hop size) is given by:

$$s(n, f) = \sum_t s(t)\text{win}_\text{a}(t - nH)e^{-\iota 2\pi f t} \tag{2.1}$$

where $\iota = \sqrt{-1}$, f is one of F frequencies $f = 0, (1/F)f_s, ..., ((F-1)/F)f_s$, and win_a is the analysis window that is non-zero only in an F sample interval $[0, F-1]$ and typically tapers smoothly to zero at each end of the interval.

To reconstruct the time domain signals $s(t)$ from $s(n, f)$, we use the inverse STFT (ISTFT):

$$s(t) = \sum_n \sum_f \mathrm{win}_\mathrm{s}(t - nH)s(n, f)e^{\iota 2\pi f t} \tag{2.2}$$

where win_s is a synthesis window that satisfies the condition [18]:

$$\sum_n \mathrm{win}_\mathrm{a}(t - nH)\mathrm{win}_\mathrm{s}(t - nH) = 1, \ \forall t \tag{2.3}$$

In the STFT (and any other time-frequency decomposition), there is a tradeoff between time and frequency resolution. We can improve the frequency resolution by increasing the STFT frame size, but this leads to a reduction in the temporal resolution.

Other Time-Frequency Representations

The STFT provides equal frequency resolution for all frequencies. However, speech signals concentrate most of their energy at low frequencies and overlapping of source signals is more probable to be present in this frequency region. The Constant-Q transform [35] imposes the condition that all subbands must have the same quality factor, therefore it has more frequency resolution in low frequencies, and less frequency resolution in the high frequencies than the STFT. Time-frequency representations can also be defined to simulate the non-uniform frequency resolution in the cochlea as in auditory filter banks [12, 36].

The Spectrogram

Figure 2.1 shows a speech waveform. A common way to display speech signals is the spectrogram, as in Figure 2.2. The time-frequency coefficients are stacked side-by-side, and converted into an image providing a way of observing the variation of the frequency spectrum of speech with time. In the image, the horizontal axis corresponds to time and the vertical axis to frequency, with colour or intensity indicating the strength of each time-frequency point (squared magnitude of the time-frequency coefficient).

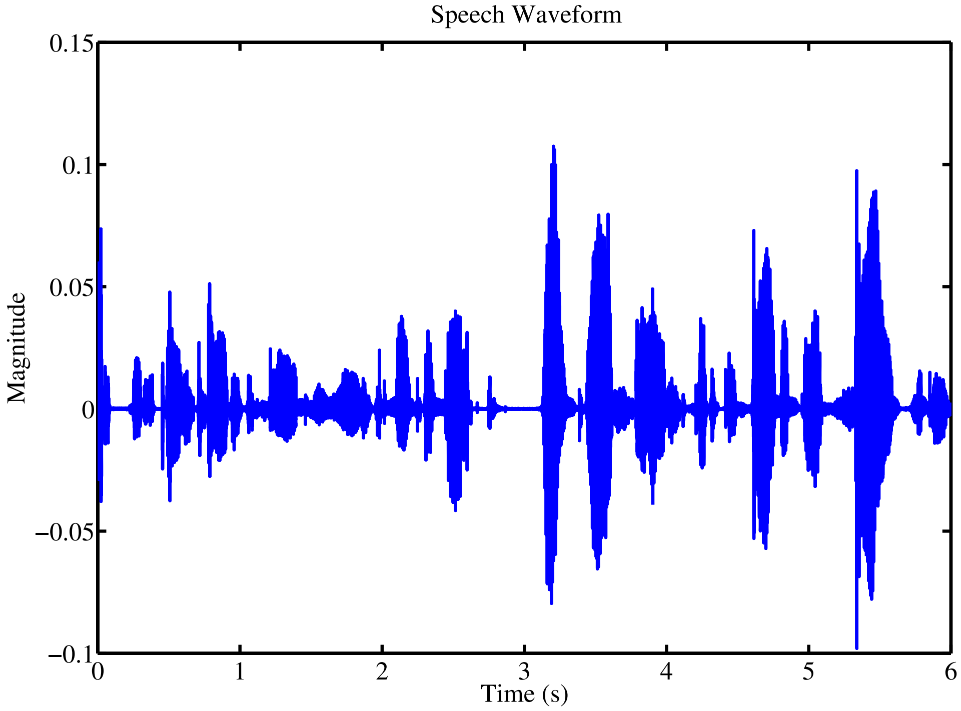

Figure 2.1: *Typical speech waveform.*

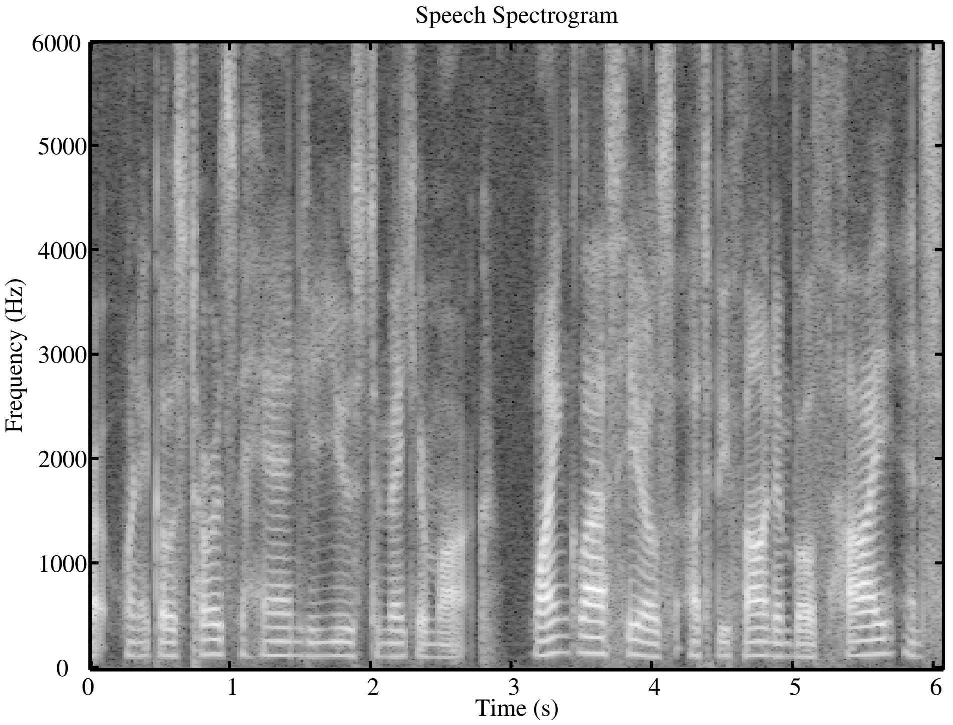

Figure 2.2: *Speech spectrogram calculated from the STFT coefficients.*

2.2.2 Speech Signal Characteristics and their Utilisation in Source Separation and Extraction

Many source separation methods exploit properties of speech signals to enable separation. Regularities in the sources such as common onset characteristics, (i.e., energy appearing at different frequencies at the same time), and harmonicity can be exploited in speech separation. Furthermore, statistical properties of speech such as non-gaussianity, non-stationarity, non-whiteness, and sparseness are popularly exploited in speech separation methods:

Non-stationarity

Speech signals are non-stationary signals in that their power spectra change over time, and amplitude modulations are largely responsible for the non-stationarity [31]. However, within short periods of time (approximately 20 ms), its spectral characteristics are fairly stationary [31, 37].

Non-Gaussianity

Speech signals typically exhibit a super-Gaussian probability distribution, which has a sharper peak and longer tails than the Gaussian probability distribution, though this is mainly due to the amplitude modulations of the signals. However, if we examine a speech signal over short term instances (approximately 20 ms), its distribution can appear super-Gaussian or sub-Gaussian [27, 38]. Speech signals are more super-Gaussian with an appropriate time-frequency representation even in short frames due to a combination of the non-stationarity and harmonic content of speech.

Non-whiteness

Speech signals are typically temporally correlated. Samples of each speech signal are not independent. However, samples from different talkers can be assumed to be statistically independent.

Sparseness

A signal is sparse when only a few instances have a value significantly different from zero. Due to the combination of the non-stationarity and harmonic content, speech signals have their energy concentrated in isolated regions in time and frequency, therefore, speech is sparse with an appropriate time-frequency representation. Moreover, the non-stationarity characteristics of individual talkers is not likely to be similar, which leads to the significant coefficients for one source to be often localised in different time-frequency points than for other sources. This leads to few overlaps of different sources in time and frequency.

2.3 Room Acoustics

This section introduces some basics of acoustic channels which are important for understanding why particular models are used in the audio source separation literature.

Sound waves spread out of sound sources and propagate at a speed which depends upon the pressure and density of the propagation medium. The speed of propagation in air is approximately $c = 340$ ms^{-1}. The amplitude of the wave decreases with the distance travelled. Generally, the radiation of point sources is modeled by spherical waves if the source size is small compared to the sound wavelength and if the positions of observation are close to the source. The area close to the point source is termed near field and in this case, the wave front is curved with respect to the distance between the positions of the observations. The sound waves may be approximated as plane waves at a sufficient distance from the source due to the decreasing curvature of the wave with respect to the distance between the positions of the observations. This is termed far field approximation. A wave source may be considered to come from the far field if [39, 40]:

$$r > \frac{2D^2}{\lambda} \tag{2.4}$$

where r is the distance to the source, λ is the wavelength, and D is the aperture width of the array (the geometric extent of the array). When a source is close to the array, the differences in distance to different parts of the array can be significant, which results in amplitude differences across the array. This must be taken into account with array processing methods especially methods that are sensitive to errors.

In real-life audio recordings, a recorded signal may contain multiple delayed and attenuated versions of the sources due to reflections of the sound waves by the room surfaces and other objects present in the room. These reflections are known as reverberation. The length of the room impulse response can be described by its reverberation time (RT). Reverberation time is the time interval it takes for the reverberation level to decay by 60 dB. Room reverberations typically last longer at low frequencies than they do at high frequencies. In general, the reverberation time is proportional to the room dimensions and inversely proportional to the absorption factor of the wall materials. The reverberation time is on the order of 150 to 500 ms in office rooms, and 1 to 2 s in concert halls [41, 42].

Measuring the Reverberation Time

Reverberation time is a local property of a room. Different positions of the source/microphone will give different reverberation times. Generally speaking, a reverberation time of a room is usually done on many source/microphone positions.

The reverberation time can be determined experimentally. The first step is to measure the impulse response $a(t)$. Impulsive excitations are usually avoided because they can only be approximations to true impulses and because it is difficult to attain a high signal to noise ratio. Another way to measure the room impulse response is to run a logarithmic sweep with instantaneous frequency varying exponentially with time through the loudspeaker, measure it at the point of interest, and then convolve the acquired signal with a time-reversed version of the excitation signal [43]. Logarithmic sweeps have a pink spectrum and have the property that its spectral distribution is often quite well adapted to ambient noise, resulting in a good signal to noise ratio at lower frequencies [44]. Figure 2.3 shows a room impulse response measured using this method.

To determine the reverberation time from the impulse response, the Schroeder method can be used [45]. First the squared impulse response is summed over time to generate the Schroeder energy decay curve:

$$\mathrm{E}_{\mathrm{decay}}(t) = \sum_{\tau=t}^{\infty} a^2(\tau) \tag{2.5}$$

The presence of noise causes a reduction in the slope of the late part of the Schroeder curve. The slope change is due to noise being integrated along with reverberation. The standardised methods [46] recommend extrapolating the segment between -5 dB and -35 dB of the measured decay curve to 60 dB by linear least-squared regression. If this 30 dB linear decay cannot be measured, then a shorter range can be used. Figure 2.4 shows the result of applying the Schroeder method to the impulse response of Figure 2.3.

2.4 Audio Mixtures

There are many scenarios where audio mixtures can be obtained. This results in different characterstics of the sources and the mixing process that can be exploited by the separation methods.

The observed spatial properties of audio signals depend on the spatial distribution of a sound source, the sound scene acoustics, the distance between the source and the microphones, and the directivity of the microphones. In general, speakers and small sound sources can be modeled as point sources if the sound source has a "negligible" extent and can be considered to be producing sound from a single point in space. Larger sound sources such as the piano or large loudspeakers emit sound at different spatial positions at the same time and are called extended sources. In many cases, the sources get filtered before being mixed. In natural recordings in a reverberant environment, the filters correspond to the acoustic impulse response between the source and the microphones. In hearing aid applications, the acoustic impulse response includes the effects of filtering by the human head. Sound effects such as panning, artificial reverberation, and equalising can also be introduced artificially using a mixing console (also called sound board or audio mixer) or dedicated software. Artificial sound effects are popular in music and movies applications.

In this research we focus on natural audio mixtures in which the mixing parameters are determined by the relative positions of sound sources and the microphones. Researchers who focus on this problem can acquire audio mixtures for the purpose of developing and testing source separation methods using a wide variety of methods.

A popular method to acquire audio mixtures for algorithm testing and development purposes is to use synthetic mixing. Synthetic mixtures can be acquired by filtering audio sources by a measured impulse response [47] or a synthetic impulse response obtained from room simulation software. While synthetic impulse responses are not as realistic as those recorded in real environments, they allow more control of the experiment parameters and conditions such as reverberation time, location of sources, array geometry and noise level.

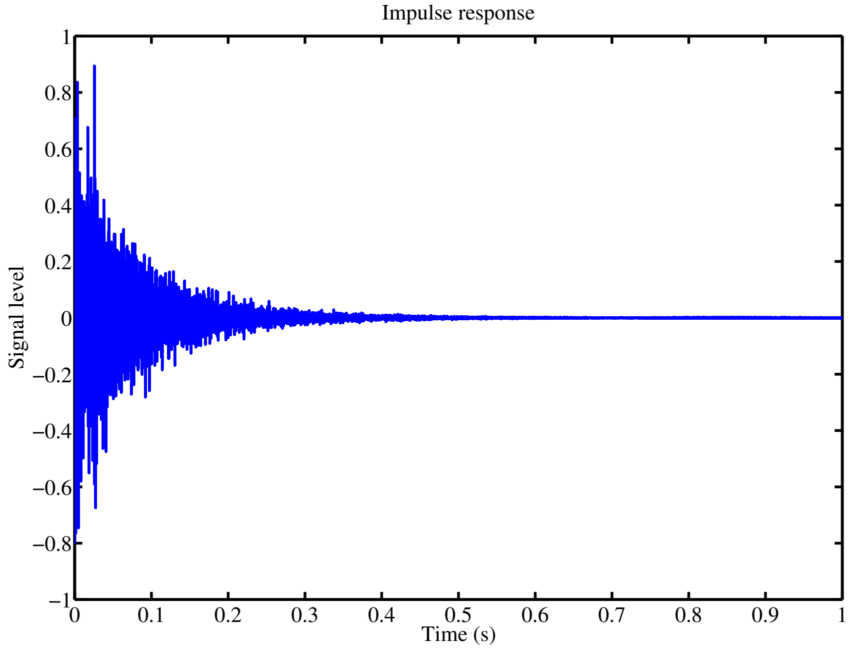

Figure 2.3: *Impulse response of a room.*

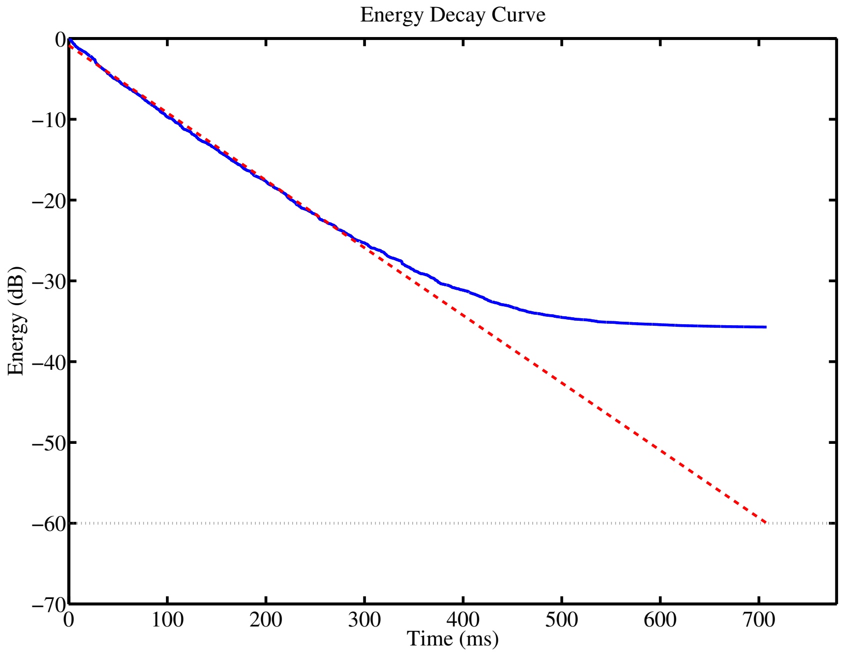

Figure 2.4: *Energy decay curve of the room impulse response (solid line) and the extrapolation of the linear segment of the measured decay curve (dashed line) yielding RT = 708 ms.*

Simulated live recordings are another popular method to acquire audio mixtures for algorithm testing and development purposes. In simulated live recordings, audio sources are played through loudspeakers, recorded one at a time by a microphone array, and subsequently added together. The sources are recorded one at a time in order to acquire the contribution of each source to the mixture individually as this helps at the performance computation stage (see Section 2.7). Compared to using synthetic impulse responses, live recordings include the effects of non-ideal microphones and any properties of the acoustic environment that the software did not simulate. Live recording can also be acquired using real human speakers. This approach includes effects such as of the speakers moving their head, but requires considerable organisational effort. Results from the stereo audio source separation evaluation campaign [48] suggested that the difficulty of separating live recordings and synthetic mixtures was similar, provided they featured the same reverberation time.

2.5 Simulating Room Acoustics

As discussed in the previous section, using room simulation software allows us to quickly explore the behaviour of source separation methods in various scenarios. There are many different methods for modeling room acoustics, but no single method can model the entire audible frequency range [49]. At low frequencies, the wavelength of sound can be comparable to the dimensions of the room, and acoustics of a room can be analysed using solutions of the wave equation [42]. In a standard room, the dimensions of the room are large compared with the wavelength of sound for a wide range of frequencies, and specular reflections and the sound ray approach can be used model the room acoustics [42].

The image method is the most well known technique for simulating the impulse response of a rectangular rooms [50]. Instead of tracing all reflections, mirror images of the source with respect to the room surfaces are created.

When a sound wave strikes a wall, the reflection angle is equal to the incident angle. Therefore the reflection of sound from the source can be replaced by placing an image source symmetrically on the other side of the wall. In the image method, the walls of the room are replaced by point sources of varying strength and location. The image source will have less power to account for the signal energy lost due to absorption by the walls and the decay with distance. As the image source gets further from the microphone, its contribution gets weaker.

Figure 2.5 shows the construction of an image source from the reflection of a wall.

This method can then be extended to include the reflections from all the room walls. The process is then repeated as each image is itself imaged in order to simulate multiple reflections. Figure 2.6 shows the construction of multiple images.

In [51], the image method was extended to arbitrary polyhedra rooms with any number of sides.

2.6 Mixing Models

Generally, the problem of source separation is stated to be the process of estimating the signals from M unobserved sources,

$$\mathbf{S} = [\mathbf{s}(1), \mathbf{s}(2), ..., \mathbf{s}(T)] = \begin{bmatrix} s_1(1) & s_1(2) & \dots & s_1(T) \\ s_2(1) & s_2(2) & \cdots & s_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ s_M(1) & s_M(2) & \cdots & s_M(T) \end{bmatrix} \tag{2.6}$$

given only a set of T samples from N microphones,

$$\mathbf{X} = [\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(T)] = \begin{bmatrix} x_1(1) & x_1(2) & \dots & x_1(T) \\ x_2(1) & x_2(2) & \cdots & x_2(T) \\ \vdots & \vdots & \ddots & \vdots \\ x_N(1) & x_N(2) & \cdots & x_N(T) \end{bmatrix} \tag{2.7}$$

which arises when the signals from the M unobserved sources are linearly mixed together. The signal recorded at the i^{th} microphone at time t can be modeled as:

$$x_i(t) = \sum_{j=1}^{M} \sum_{p=0}^{P-1} \alpha_{ij}^{p} s_j(t - \Delta_{ij}^{p}) \tag{2.8}$$

where P is the number of paths between each source-microphone pair, α_{ij}^{p} represents the attenuation of the p^{th} acoustic path from source j to microphone i, and Δ_{ij}^{p} is the delay of the p^{th} path from source j to microphone i. This model assumes a fixed number of sources and a static channel. The mixing process can be written in a vector form as follows:

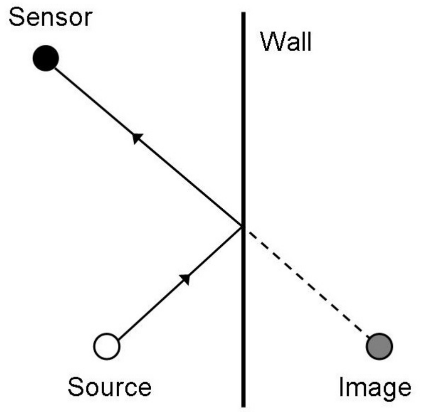

Figure 2.5: *Creating the image source. The solid line represents the actual path, while the dashed line represents the perceived path.*

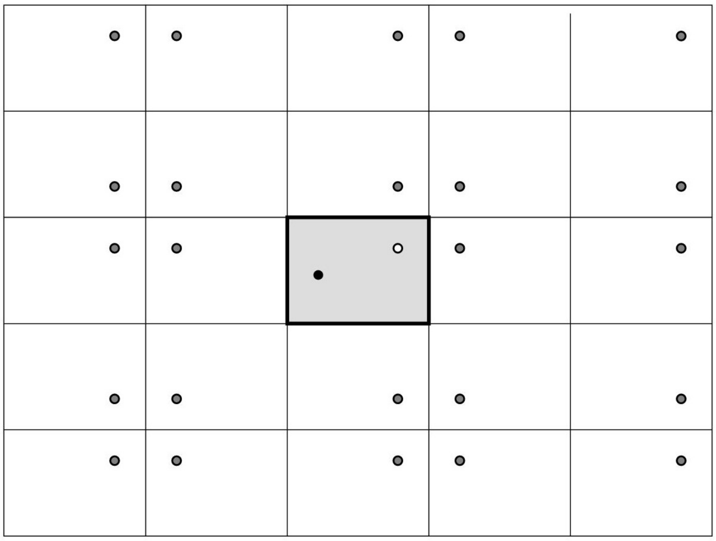

Figure 2.6: *The original room and the images. The microphone is in black, the source is in white, and the images are in grey. The actual image space is three dimensional.*

$$\mathbf{x}(t) = \mathbf{A}(t) \star \mathbf{s}(t) \tag{2.9}$$

where $\star$ denotes the convolution operator and $\mathbf{A}$ is a $N \times M$ matrix of mixing filters:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix} \tag{2.10}$$

with elements $a_{ij} = \sum_{p=1}^{P} \alpha_{ij}^{p} \delta(t - \Delta_{ij}^{p})$. For mathematical convenience, it is common to normalise the magnitude and delay of each column in the mixing matrix (having a reference microphone with no attenuation and zero delay).

These equations can be mapped to the time-frequency domain using the STFT. Denoting the STFT coefficients of $x_i(t)$ and $s_j(t)$ as $x_i(n, f)$ and $s_j(n, f)$, in the time frame n and frequency bin f, and assuming that the STFT frame length is larger than the length (the largest delay) of the mixing filters a_{ij}, and that the narrowband assumption holds, [1] we can approximate the mixing filters by complex mixing scalars $a_{ij}(f)$, and the convolution in the time domain can be written in the frequency domain as separate multiplications for each frequency:

$$x_i(n, f) \approx \sum_{j=1}^{M} a_{ij}(f) s_j(n, f) \tag{2.11}$$

This mixing process can be written in a vector form as follows::

$$\mathbf{x}(n, f) = \mathbf{A}(f)\mathbf{s}(n, f) \tag{2.12}$$

Assuming we are only interested in extracting source j', $j' \in \{1, 2, ..., M\}$, the mixing model

[1] We decompose the speech broadband signal into narrowband bandpass signals using the STFT. We assume that the array size is small enough relative to the STFT frame length and the bandwidth of each bandpass signal such that relative delay between the microphones can be expressed as a phase shift. This is known as the narrowband assumption [13, 52]. This condition is easily met in non-echoic rooms with moderately sized STFT windows. For example, the maximum delay in a 5 cm linear array is 0.05/340 = 0.15 ms, and a 1024 point window corresponds to 64 ms at 16 kHz sampling rate.

in (2.11) can be reformulated as:

$$\begin{aligned} x_i(n,f) &= a_{ij'}(f)s_{j'}(n,f) + \sum_{j \neq j'} a_{ij}(f)s_j(n,f) \\ &= a_{ij'}(f)s_{j'}(n,f) + v_i(n,f) \end{aligned} \tag{2.13}$$

where v_i represents the contribution of the interferers to the mixture signal x_i. In vector form, the mixing model can be written as:

$$\mathbf{x}(n,f) = \mathbf{a}(f)s(n,f) + \mathbf{v}(n,f) \tag{2.14}$$

where $\mathbf{x}(n,f) = [x_1(n,f), ..., x_N(n,f)]^T$ is the observed multichannel mixture signal, $\mathbf{a}(f)$ is the $N \times 1$ array response vector in the direction of the desired source signal s (also called the propagation vector or steering vector), and $\mathbf{v}(n,f) = [v_1(n,f), ..., v_N(n,f)]^T$ is the $N \times 1$ vector of the contribution of the interferers and any possible noise to the mixture signal. In this model, no assumptions are made about the interferers. The interferers can be be of any nature such as point sources, spatially extended sources, diffuse sources, or a combination of them. The array response vector $\mathbf{a}(f)$ is the representation of the delays and the attenuation in the frequency domain, and depends on the array geometry and the location of the desired source signal. If we consider near field conditions, we have:

$$\mathbf{a}(f) \;=\; [\alpha_1 e^{-\iota 2\pi f \Delta_1}, ..., \alpha_N e^{-\iota 2\pi f \Delta_N}]^T \tag{2.15}$$

where $\iota = \sqrt{-1}$, $\alpha_i = \frac{r_i}{r_I}$ is equal to the ratio of the distance from the source to the i^{th} microphone (r_i) to the distance from the source to the reference microphone (r_I), and:

$$\Delta_i = (r_i - r_I)/c \tag{2.16}$$

When the source distance to the microphone array is much larger than the microphone spacing, the value of α_i is very close to 1, and we have:

$$\mathbf{a}(f) \;\approx\; [e^{-\iota 2\pi f \Delta_1}, ..., e^{-\iota 2\pi f \Delta_N}]^T \tag{2.17}$$

In the special case of uniform linear array, and far-field sources, we can approximate the delay between the microphones as $\Delta_i = (i - I)(d/c)\sin\phi$, where d represents the microphone spacing for the uniform linear array, c the sound velocity, and ϕ the DOA relative to broadside.

In this case (far field), the delay between the microphones is independent of the source distance.

In general, the mixing process can be classified as either instantaneous, anechoic, and echoic. We now describe these mixing processes in more detail.

2.6.1 Instantaneous Mixtures

In the case of instantaneous mixing, where the samples of each source arrive at the microphones at the same time and with differing attenuations, each elements of the mixing matrix a_{ij} is a scalar that represents the amplitude scaling between source j and microphone i. Instantaneous mixing can be encountered in synthetic audio mixtures.

2.6.2 Anechoic Mixtures

The anechoic mixing model is an extension of the instantaneous mixing model where delays between microphones are considered. In anechoic mixing, it is assumed that the samples of each source signal can arrive at the microphones with different delays and a_{ij} takes the form $\alpha_{ij}\delta(t-\Delta_{ij})$. Anechoic mixtures can be encountered in reverberation free environments (such as acoustic anechoic chambers), where the samples of each source signal can arrive at the microphones only from the line of sight path, and the attenuation and delay of source j would be determined by the physical position of the source relative to the microphones. Anechoic mixtures can also be encountered in synthetic mixtures.

2.6.3 Echoic Mixtures

In echoic (also referred to as reverberant) mixing, there exist multiple paths between each source-microphone pair, and a_{ij} takes the form $\sum_{p=1}^{P}\alpha_{ij}^{p}\delta(t-\Delta_{ij}^{p})$, where P is the number of paths between each source-microphone pair. In this case, each microphone signal at time t depends on source signals from not just the same instant of time, but also from previous instants. The echoic mixing model is the most natural mixing model, and can be encountered in live recordings or synthetic mixtures.

2.7 Evaluation Measures

Many performance measures for source separation and extraction have been proposed. If the goal of source separation is the creation of audio for a human listener, the opinion of human listeners in subjective tests is the gold standard to evaluate the quality of the audio signal. Subjective evaluation involves a group of listeners being asked to rate the quality of speech along a predetermined scale. In order for the listening tests to be accurate, they are conducted with a large number of listeners and are therefore expensive, slow to conduct, and require considerable organisational effort. On the other hand, assuming the true sources are known, it is feasible to compute numerical performance measures for the separation quality by measuring the numerical "distance" between the true sources and their estimates . For an objective measure to be valid, it needs to correlate well with subjective listening tests.

An obvious objective measure to evaluate the quality of the estimated sources is the signal to distortion ratio (SDR). The SDR measures the ratio between the energy of the true source and estimation error:

$$\text{SDR} = 10 \log_{10} \frac{||s||^2}{||\hat{s} - s||^2} \tag{2.18}$$

This SDR is an overall measure of all distortions, including residual interference and any artifacts and changes to the desired signal introduced by the separation method.

However, depending on the application, different distortions can be allowed between a source estimate and the true source [53]. In most cases, it is perfectly acceptable to recover the source signals with an arbitrary gain factor, and in some cases, a linear filtering distortion can be allowed. Moreover, it is useful to measure separately the amount of residual interference and the amount of other artifacts, such as musical noise.

In [53], an evaluation method has been proposed that allows linear filtering distortions and provides a separate measure for residual interference, termed the source to interference ratio (SIR) and forbidden artifacts, termed the sources to artifacts ratio (SAR). The computation of the evaluation measures involves two steps. In the first step, each estimated signal $\hat{s_j}$ is decomposed as:

$$\hat{s_j} = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}} \tag{2.19}$$

where s_{target} is a version of the desired source s_j modified by an allowed distortion, and where e_{interf} and e_{artif} are respectively the interferences and artifacts error terms. The decomposition is based on orthogonal projections. In [53], it is shown how to decompose $\hat{s}$ when the allowed distortions are: time-invariant gain, time-invariant filtering, time-varying gain, and time-varying filtering.

In our work, we will allow for time-invariant filtering. This is achieved as follows. s_{target} is computed by projecting $\hat{s_j}$ on the subspace spanned by delayed versions of s_j [53]:

$$P_{s_j} = \prod\{(s_j(t-\delta)_{0\leq\delta\leq L-1}\} \tag{2.20}$$

$$s_{\text{target}} = P_{s_j}\hat{s_j} \tag{2.21}$$

where $L-1$ is the maximum delay allowed and $\prod\{y_1,...,y_k\}$ denotes the orthogonal projector onto the subspace spanned by the vectors $y_1,...,y_k$. e_{interf} is obtained via [53]:

$$P_{\mathbf{s}} = \prod\{(s_{j'}(t-\delta)_{0\leq j'\leq M, 0\leq\delta\leq L-1}\} \tag{2.22}$$

$$e_{\text{interf}} = P_{\mathbf{s}}\hat{s_j} - P_{s_j}\hat{s_j} \tag{2.23}$$

and finally e_{artif} is obtained via:

$$e_{\text{artif}} = \hat{s_j} - P_{\mathbf{s}}\hat{s_j} \tag{2.24}$$

The computation of the above projections are detailed in [53].

In a second step, energy ratios are computed to evaluate the relative amount of each of these terms as follows:

$$\text{SDR} = 10\,\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}+e_{\text{artif}}||^2} \tag{2.25}$$

$$\text{SIR} = 10\,\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2} \tag{2.26}$$

$$\text{SAR} = 10\,\log_{10}\frac{||s_{\text{target}}+e_{\text{interf}}||^2}{||e_{\text{artif}}||^2} \tag{2.27}$$

These performance measures are implemented within a MATLAB toolbox named BSS_EVAL distributed online under the GNU public license [54]. These performance measures are the most popular in the audio source separation community.

Perceptual measures have also been proposed to evaluate audio source separation algorithms [55, 56]. However, most perceptual measures have been developed for audio codec evaluation, and have been optimised for signals where the difference between the output and true signals is caused by quantisation and compression. Therefore, it has been argued that they are not well suited for source separation applications [57].

On a final note, when source separation is used as a pre-processing stage to some subsequent application, the effectiveness of source separation can be judged by the performance of the final application. Metrics can be devised to measure the effectiveness of the source separation stage on the overall application. For example, the word error rate (WER) of a speech recognition device [58] or the transcription accuracy of a music transcription system [59] have been proposed.

2.8 Existing Techniques for Speech Source Separation and Extraction

2.8.1 Beamforming

Many approaches to speech source separation and extraction rely on beamforming techniques. These techniques approach the separation problem from a spatial point of view; the microphone array is used to form a spatial filter which can extract a signal from a specific direction and reduce signals from other directions. This spatial filter can be expressed in terms of dependence upon angle and frequency. Beamforming is accomplished by filtering the microphone signals and combining the outputs to extract (by constructive combining) or reject (by destructive combining) signals according to their spatial location. Beamforming can separate signals with overlapping frequency content that originate at different spatial locations [13, 14].

Beamforming for broadband signals like speech can, in general, be performed in the time domain or frequency domain. In time domain beamforming, a finite impulse response (FIR) filter is applied to each microphone signal, and the filter outputs combined to form the beamformer output. Beamforming can be performed by computing multichannel filters whose output is $\hat{s}(t)$

an estimate of the desired source signal. The output can be expressed as:

$$\hat{s}(t) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} w_{i,p} x_i(t-p) \tag{2.28}$$

where $P-1$ is the number of delays in each of the N filters. In frequency domain beamforming, the microphone signal is separated into narrowband frequency bins (for example using a STFT), and the data in each frequency bin is processed separately.

Beamformers can be classified as either deterministic (also termed data-independent) or statistically optimum. The filters in a deterministic beamformer do not depend on the microphone signals and are chosen to approximate a desired response. For example, we may wish to receive any signal arriving from a certain direction, in which case the desired response is unity over at that direction. As another example, we may know that there is interference operating at a certain frequency and arriving from a certain direction, in which case the desired response at that frequency and direction is zero. The simplest deterministic beamforming technique is delay-sum beamforming, where the signals at the microphones are delayed and then summed in order to combine the signal arriving from the direction of the desired source coherently, expecting that the interference components arriving from off the desired direction cancel to a certain extent by destructive combining. Assuming that the broadband signal can be decomposed into narrowband frequency bins, the delays can be approximated by phase shifts in each frequency band.

In statistically optimum beamforming, the filters are designed based on the statistics of the arriving data to optimise some function that makes the beamformer optimum in some sense. For example, a beamformer can be designed to minimise the expectation of the squared difference between the beamformer output and the desired source. This beamformer can be viewed as a multichannel Wiener filter and is called the Minimum Mean Square Error (MMSE) Beamformer. As in deterministic beamforming, constraints can be used to shape the directivity pattern of the beamformer in order to emphasise or attenuate specific directions and frequencies. For example, in minimum-variance distortionless response (MVDR) beamforming, the beamformer response is constrained so that signals from the direction of interest are passed with no distortion, while minimising the power of the beamformer's output. This leads to a reduction in noise and interference power while preserving the integrity of the desired source signal.

In practice, the statistics needed to design the optimum beamformers are usually not known and/or are changing, and we must estimate them from the incoming data and follow their changes over time. The statistics can be estimated from a temporal block of data (block adaptation), or adjusted as the data arrives (continuous adaptation).

In [60, 61], beamforming weights were calculated using time-domain recursive algorithms. In the Frost beamformer [60], beamforming weights are recursively adjusted using a constrained least mean squares (LMS) algorithm to minimise the output power of the beamformer while maintaining a constant response in the look-direction. In the generalised sidelobe canceller [61], the constrained minimisation problem is converted to an unconstrained minimisation problem using a two path structure: a deterministic path and an adaptive path. The deterministic path is a deterministic beamformer with constraints on the desired signal. The adaptive path employs a spatial blocking structure that blocks the desired signal and adaptively minimise the non-desired components using an unconstrained LMS.

Recently, it was shown in [62] that in the case of two microphones, a frequency-domain minimum-variance (FMV) beamformer which performs sample matrix inversion using statistics estimated from a short sample support (sliding block adaption) gives better performance than time-domain recursive algorithms [60, 61] in multi-talker acoustic environments. In the FMV algorithm [62], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could vary over longer time spans. In the FMV algorithm, STFT values of the mixture signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the most recent 32 values. MVDR weights are then calculated using the correlation matrix. Only statistics gathered over a very short period of time are used in the calculation of weights. The quick adaptation of the beamformer weights can substantially reduce a large number of nonstationary interferences while utilising few microphones [62]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. Thus the FMV beamformer is forced to make a compromise between long intervals (good statistics) and short intervals (rapid response).

In general, beamforming techniques require information about the microphone array configuration and the sources (for example, the direction of the desired source). However, beamforming techniques can attenuate spatially spread and reverberant interferers, and there is no need to determine the number of interferers. In general, linear adaptive beamforming can attain excellent separation performance in determined or over-determined mixtures. However, in under-determined mixtures, perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible.

2.8.1.1 Statistically Optimum Beamformers

In this subsection, we discuss some of the statistically optimum beamformers that are relevant to this thesis. We consider the mixing model in (2.14). Note that $\mathbf{x}$, $\mathbf{a}$, s, and $\mathbf{v}$ are complex valued, and depend on frequency f, but for readability and simplicity, we will omit this variable in the rest of this subsection. From now on, we implicitly work in a given frequency band under the narrowband assumption.

We will begin our discussion by modeling the desired source signal as an unknown nonrandom signal arriving at the array from a known direction and present a derivation for the optimum filter that minimises the output power and passes any signal arriving to the array from the specified direction through the filter undistorted. For this reason, this filter is termed the minimum variance distortionless response (MVDR) filter. We then present a derivation for the optimum linear beamformer to estimate the signal waveform using a MMSE criterion. We demonstrate that this linear MMSE estimator consists of an optimum distortionless beamformer followed by a scalar filter. Finally, we consider a beamformer that assumes that we know or can measure $\mathbf{R}_x = \mathrm{E}[\mathbf{x}\mathbf{x}^H]$, the correlation matrix of the observed mixture signal $\mathbf{x}$, but do not know $\mathbf{R}_v = \mathrm{E}[\mathbf{v}\mathbf{v}^H]$, the correlation matrix of the interference component. We choose a look direction and find the optimum distortionless filter for that direction that minimises the mean square output power. We refer to this filter as the minimum power distortionless response (MPDR) beamformer. A more detailed derivation and discussion can be found in many array processing books, such as [52].

Linear Minimum Variance Distortionless Response (MVDR) Beamformer

Consider the mixing model:

$$\mathbf{x} = \mathbf{a}s + \mathbf{v} \tag{2.29}$$

We assume that the interference signal is a sample function of a random process with known second order statistics, but the desired source signal is an unknown nonrandom signal arriving at the array from a known direction. We process the observed mixture signal with a linear vector filter $\mathbf{w}$ (row vector) to get an estimate of the desired source signal:

$$\hat{s} = \mathbf{w}\mathbf{x} = \mathbf{w}\mathbf{a}s + \mathbf{w}\mathbf{v} \tag{2.30}$$

We require that in the absence of interference $\hat{s} = s$. This distortionless constraint implies that:

$$\mathbf{w}\mathbf{a} = 1 \tag{2.31}$$

We wish to minimise the variance of $\hat{s}$ in the presence of interference under the distortionless constraint. This is equivalent to minimising the power of the interference at the output. The power of the output interference can be written as:

$$\mathrm{E}[|\mathbf{w}\mathbf{v}|^2] = \mathbf{w}\mathbf{R}_v\mathbf{w}^H \tag{2.32}$$

where $\mathbf{R}_v$ is the correlation matrix of observed interference. The minimisation of the output interference power can be done by taking derivatives with respect to $\mathbf{w}$ and setting them to be equal to zero, while also including a Lagrangian term to account for the constraint in (2.31). This gives the minimum variance distortionless response beamformer [52, 63]:

$$\mathbf{w}^{\text{MVDR}} = \frac{\mathbf{a}^H\mathbf{R}_v^{-1}}{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}} \tag{2.33}$$

It can be shown [52] that if the interference is modeled as a circular complex Gaussian random vector, then the output of the MVDR filter is the maximum likelihood estimate of the signal s.

Linear Minimum Mean Square Error (MMSE) Beamformer

We now consider the optimum linear filter whose output is the MMSE estimate of the desired signal s. We assume that the desired source signal is a scalar zero mean random variable with known variance σ_s^2. We also assume that the interference $\mathbf{v}$ have a spatial covariance matrix $\mathbf{R}_v$. Additionally, it is assumed that the signal and interference snapshots are uncorrelated.

Therefore the spatial covariance matrix of $\mathbf{x}$ is:

$$\mathbf{R}_x = \mathbf{R}_v + \sigma_s^2 \mathbf{a}\mathbf{a}^H \tag{2.34}$$

The linear MMSE estimator of the desired signal s minimises the mean-square error:

$$\zeta = \mathrm{E}\left[\left|s - \mathbf{w}^{\mathrm{MMSE}}\mathbf{x}\right|^2\right] \tag{2.35}$$

Taking the derivative with respect to $\mathbf{w}^{\mathrm{MMSE}}$ and setting the result to zero gives:

$$\mathrm{E}\left[s\mathbf{x}^H\right] - \mathbf{w}^{\mathrm{MMSE}}\mathrm{E}\left[\mathbf{x}\mathbf{x}^H\right] = 0 \tag{2.36}$$

Thus:

$$\begin{aligned} \mathbf{w}^{\mathrm{MMSE}} &= \mathbf{R}_{sx^H}\mathbf{R}_x^{-1} \\ &= \sigma_s^2 \mathbf{a}^H \mathbf{R}_x^{-1} \end{aligned} \tag{2.37}$$

$\mathbf{R}_x^{-1}$ can be simplified using the matrix inversion lemma:

$$\mathbf{R}_x^{-1} = \mathbf{R}_v^{-1} - \sigma_s^2 \mathbf{R}_v^{-1}\mathbf{a}(\sigma_s^{-2} + \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a})^{-1}\mathbf{a}^H\mathbf{R}_v^{-1} \tag{2.38}$$

Therefore, the linear MMSE estimator can be expressed as:

$$\begin{aligned} \mathbf{w}^{\mathrm{MMSE}} &= \sigma_s^2 \mathbf{a}^H(\mathbf{R}_v^{-1} - \sigma_s^2\mathbf{R}_v^{-1}\mathbf{a}(\sigma_s^{-2} + \mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a})^{-1}\mathbf{a}^H\mathbf{R}_v^{-1}) \\ &= \frac{\mathbf{a}^H\mathbf{R}_v^{-1}}{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a} + \sigma_s^{-2}} \end{aligned} \tag{2.39}$$

This can alternatively be expressed as [52]:

$$\mathbf{w}^{\mathrm{MMSE}} = \underbrace{\frac{\mathbf{a}^H\mathbf{R}_v^{-1}}{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}}}_{\text{MVDR}} \underbrace{\frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}\right)^{-1}}}_{\text{post filter}} \tag{2.40}$$

The first term is an MVDR spatial filter, which suppresses the interfering signals without distorting the signal propagating along the desired source direction. The second term is a single channel post-filter. We see that the linear MMSE estimator is just a shrinkage of the MVDR beamformer.

In general, the MMSE estimator is not linear in the data. The MMSE estimator is linear if either the estimator is constrained to be linear, or all the signals are Gaussian.

Linear Minimum Power Distortionless Response (MPDR) Beamformer

The MVDR and MMSE beamformers assume that $\mathbf{R}_v$ and $\mathbf{a}$ are known, and the MMSE beamformer also assumes that σ_s^2 is known . We now assume that we know $\mathbf{R}_x$, the correlation matrix of the observed mixture signal $\mathbf{x}$, but do not know $\mathbf{R}_v$, σ_s^2, or $\mathbf{a}$. We will steer the distortionless constraint to the direction $\widetilde{\mathbf{a}}$ that we believe the signal is arriving from. We want to minimise the mean square output power subject to a distortionless constraint in the direction of the steering vector. This gives the so called minimum power distortionless response (MPDR) beamformer [52]:

$$\mathbf{w}^{\text{MPDR}} = \frac{\widetilde{\mathbf{a}}^H \mathbf{R}_x^{-1}}{\widetilde{\mathbf{a}}^H \mathbf{R}_x^{-1} \widetilde{\mathbf{a}}} \tag{2.41}$$

It can be shown that when the steering vector is equal to the array response vector in the direction of the desired source signal, the MPDR and the MVDR beamformers are identical [52].

2.8.2 Independent Component Analysis

Another approach to source separation is to exploit statistical properties of source signals. One popular assumption is that the different sources are statistically independent, and is termed Independent Component Analysis (ICA) [15]. In ICA, separation is performed on the assumption that the source signals are statistically independent, and does not require information on microphone array configuration or the direction of arrival (DOA) of the source signals to be available. In the high-SNR instantaneous and determined mixtures case, the source separation problem can be performed by estimating the mixing matrix $\mathbf{A}$, and this allows one to compute a separating matrix $\mathbf{W} = \mathbf{A}^{-1}$ whose output:

$$\hat{\mathbf{s}}(t) = \mathbf{A}^{-1}\mathbf{x}(t) = \mathbf{W}\mathbf{x}(t) \tag{2.42}$$

is an estimate of the source signals. The mixing matrix or the separating matrix $\mathbf{W}$ is determined so that the estimated source signals are as independent as possible. The separating matrix acts as a linear spatial filter or beamformer that attenuates the interfering signals.

ICA can be approached with the assumption of non-Gaussianity of the sources, called independent components, using many methods such as the infomax principle [64], maximum likelihood estimation [65], minimising Kullback-Leibler (KL) divergence [66], or using kurtosis or negentropy as non-Gaussianity measures [67, 68]. Other ICA approaches use the time structure of the source signals, and assume source signals have different power spectra (i.e., different autocovariance functions) [69] or have non-stationary variances [70]. In the ICA model, we cannot determine the variances of the independent components (scale ambiguity), and we cannot determine their order (ordering ambiguity). The identifiability and separability of ICA models was investigated in [71, 72]. For some excellent reviews of ICA, see [16, 73, 74].

2.8.3 Convolutive Independent Component Analysis

In real room recordings, the mixing of speech signals is convolutive. Estimating the unmixing filters for long acoustic channels in the time domain is generally computationally expensive. Therefore, researchers transformed the task into the frequency domain. Transformation into the frequency domain is usually done via the STFT. If a sufficiently long frame for the STFT is used, the convolutive mixture can be approximated with an instantaneous mixture in each frequency bin. Another good side-effect of the STFT is that speech signals are much sparser (and hence much more non-Gaussian) in the time-frequency domain. ICA algorithms display improved performance when sources are sparse [16]. In [75], unmixing was performed in the frequency domain. However, the sources were modeled in the time domain in order to estimate the unmixing matrix $\mathbf{W}(f)$. This requires the algorithm to transform between the frequency and time domains for every update, thereby resulting in an increase in computational complexity. It was proposed in [76] to work solely in the frequency domain; the source probability model was applied in the frequency domain in order to avoid the extra complexity of transforming between the frequency and time domains for every update. In this case, however, the update is performed independently for each frequency bin, and the estimated sources are arbitrarily ordered in each frequency bin (permutation ambiguity). In order to synthesise the separated signals in the time domain, frequency domain separated signals originating from the same source should be grouped together. Therefore, it was proposed in [76] that some frequency coupling be applied between neighboring frequency bins. This in turn imposes some smoothness constraints across frequencies. However, it had limited effect on solving the permutation problem. In [38], a time-frequency probabilistic source model was proposed to solve the permutation problem; it is used to couple the frequency bins by measuring the signal envelope along the frequencies.

In [77], geometric constraints were used to resolve some of the permutation ambiguities by performing careful initialisation of the filter parameters, and using spatial penalty terms. Beamforming methods were also combined with the frequency domain framework in order to resolve the permutation problem; in [78–80], the sources are separated using frequency domain ICA, following which the sources are sorted according to the estimated DOA. The DOA approach can be used to effectively align permutations. But, at low frequencies, we cannot estimate the DOA accurately enough, and we also have the spatial aliasing problem at high frequencies when the distance between the microphones is smaller than half the wavelength at these frequencies. In [81], inter-frequency correlation of the speech signal envelope at neighboring frequencies was used to align permutations, after fixing permutations at frequency bins when the confidence of the DOA approach is sufficiently high. In [82], the harmonic structure of speech signals was exploited and used with the DOA and neighboring frequencies correlation approaches to provide a robust permutation alignment method.

In general, convolutive ICA can attain good separation performance in determined or overdetermined time-invariant mixtures. However, some aspects limit the application of ICA to real-world environments. Long demixing filters are required to handle long reverberation in order to reduce the reverberant components of interferers. On the other hand, it is desired that demixing filters can be estimated using short data in order to cope with changing channels in real environments. Furthermore, the ability to measure independence might fail with too few samples [83]. In general, most ICA methods assume the number of sources is given a priori, and classical ICA techniques cannot perform source separation in the underdetermined mixtures case. For some excellent reviews of convolutive ICA methods for speech separation, see [17, 18].

2.8.4 More Sources than Mixtures

Many source separation applications are limited by the number of available microphones. It is not always guaranteed that the number of microphones is more than or equal to the number of sources. Source separation techniques initially only dealt with determined or overdetermined mixtures where the number of sources was equal to or less than the number microphones ($M \leq N$). In this case, the problem is one of identifying the mixing matrix $\mathbf{A}$ and the source estimates are simply $\mathbf{s}(t) = \mathbf{A}^{-1}\mathbf{x}(t)$. In the underdetermined mixing case ($M > N$), linear source separation using the inverse of the mixing matrix is not possible (see (2.42)) as, in this case, $\mathbf{A}$ is not square. Therefore, in underdetermined mixtures, there are two interrelated problems: (1) the estimation of the mixing matrix, and (2) the estimation of the sources. In general, linear methods can not completely remove more than $N - 1$ sources from the mixture. However, under certain assumptions, it is possible to extract a larger number of sources by using non-linear methods. The assumption that the sources have a sparse representation in a given basis is a very popular assumption in underdetermined mixtures. Sparseness of a signal means that only a few instances have a value significantly different from zero.

The mixing matrix estimation may be performed using either clustering or Bayesian approaches. In [84], it is assumed that only one source can be non-zero at each sample. In this case, the collection of points representing the mixture signal vectors are more or less aligned along straight lines passing through the origin and the direction of these lines is given by the columns of the mixing matrix. This alignment can be observed on the scatter plot of mixture signal. Figure 2.7 shows an example of the scatter plot of a two channel instantaneous mixture of three super-Gaussian distributed sources. The essence of the clustering approach is the identification of line orientation vectors from the mixtures. In [85], it was proposed to use a linear sparse transform to enhance the performance of clustering and source separation. In [86, 87], the mixing matrix was learned by maximising the probability of the data given the model. The likelihood computation requires performing marginalisation over all sources:

$$p(\mathbf{x} \mid \mathbf{A}) = \int p(\mathbf{x} \mid \mathbf{A}, \mathbf{s}) p(\mathbf{s}) d\mathbf{s} \tag{2.43}$$

This integral is intractable for under-determined mixing. In [86], it was proposed to fit a multivariate Gaussian around the posterior mode, and the mixing matrix was learned by performing gradient ascent. In [88], it was proposed to learn the source densities $p(s_j)$ from the observed

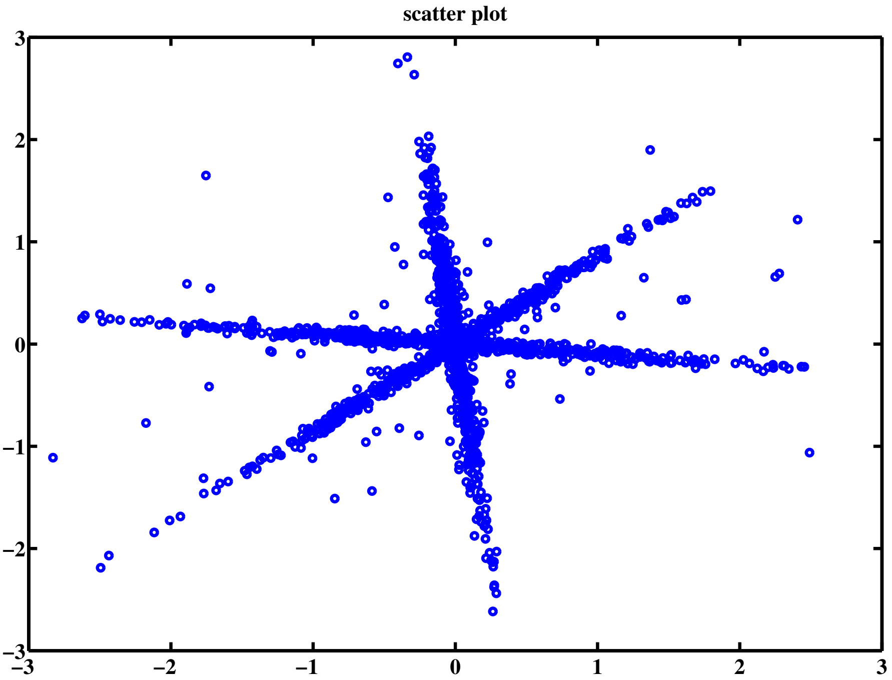

Figure 2.7: *Scatter plot of an instantaneous mixture of three super-Gaussian distributed sources.*

data. The sources were modeled as independent random variables with mixture of Gaussians (MoGs) distributions; a parametric model, but sufficiently general to model arbitrary source densities. An expectation maximisation (EM) algorithm [30] was used to learn the parameters of the model, namely the mixing matrix, noise covariance, and source density parameters. In [89], approximations were used to overcome the problem that the number of mixtures in the observation density in [88] grows exponentially with the number of sources. The observation density is written as a summation of Gaussians with decaying weights, and then the number of Gaussians is truncated in order to retain only those with reasonable sized weights.

Given the mixing matrix, the sources can be estimated from the mixture using the LMS or maximum a posteriori (MAP) estimation [87, 88, 90]. Imposing a source model, and given the mixing matrix, the sources can be estimated by a gradient-based algorithm. However, this approach is usually slow. In [86], linear programming was used to maximise the log posterior likelihood under a Laplacian prior. The Laplacian prior reduces the problem to one of minimising the $L1$-norm of the estimated sources. In [91], an efficient implementation of the $L1$-norm minimisation for the two mixture case was presented; where a reduced matrix includes only the columns of $\mathbf{A}$ whose directions (on the scatter plot) are closest from below and from above to the direction of the mixture vector is used.

Another very popular approach to underdetermined source separation is time-frequency masking [19–21, 92–95]. This approach estimates the time-frequency representation of the desired source $s_j(n, f)$ from the time-frequency representation of the mixture signal $x_i(n, f)$ by:

$$\hat{s}_j(n, f) = \mathbf{M}_j(n, f)x_i(n, f) \tag{2.44}$$

where $\mathbf{M}_j$ is a time-frequency mask containing positive gains which must be adapted to extract source j from the observed mixture. If the sources do not overlap in the time-frequency domain it is possible to separate them with a binary mask. A mask can be applied in the time-frequency domain to attenuate interfering signals while preserving time-frequency points where the signal of interest is dominant. Figures 2.9 and 2.8 show the application of time-frequency masking to extract a source from a mixture of four sources (the mask is computed using the DUET algorithm, described below).

If more than one mixture signal is available, the spatial information at each time-frequency point can sometimes be used to determine which time-frequency points belong to each source.

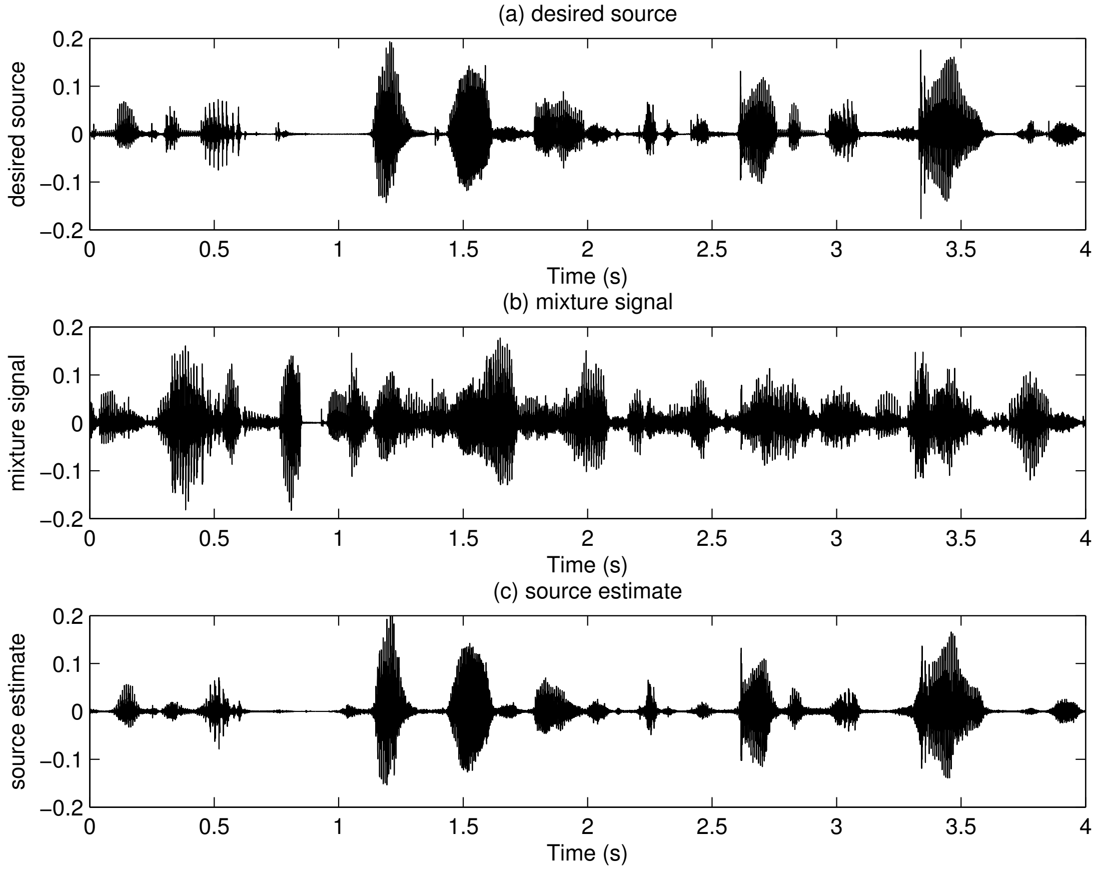

Figure 2.8: *Time-frequency masking. Plot (a) shows the desired source signal. Plot (b) shows the mixture signal. Plot (c) shows an estimate of the desired source extracted by applying time-frequency binary masks to the mixture signal.*

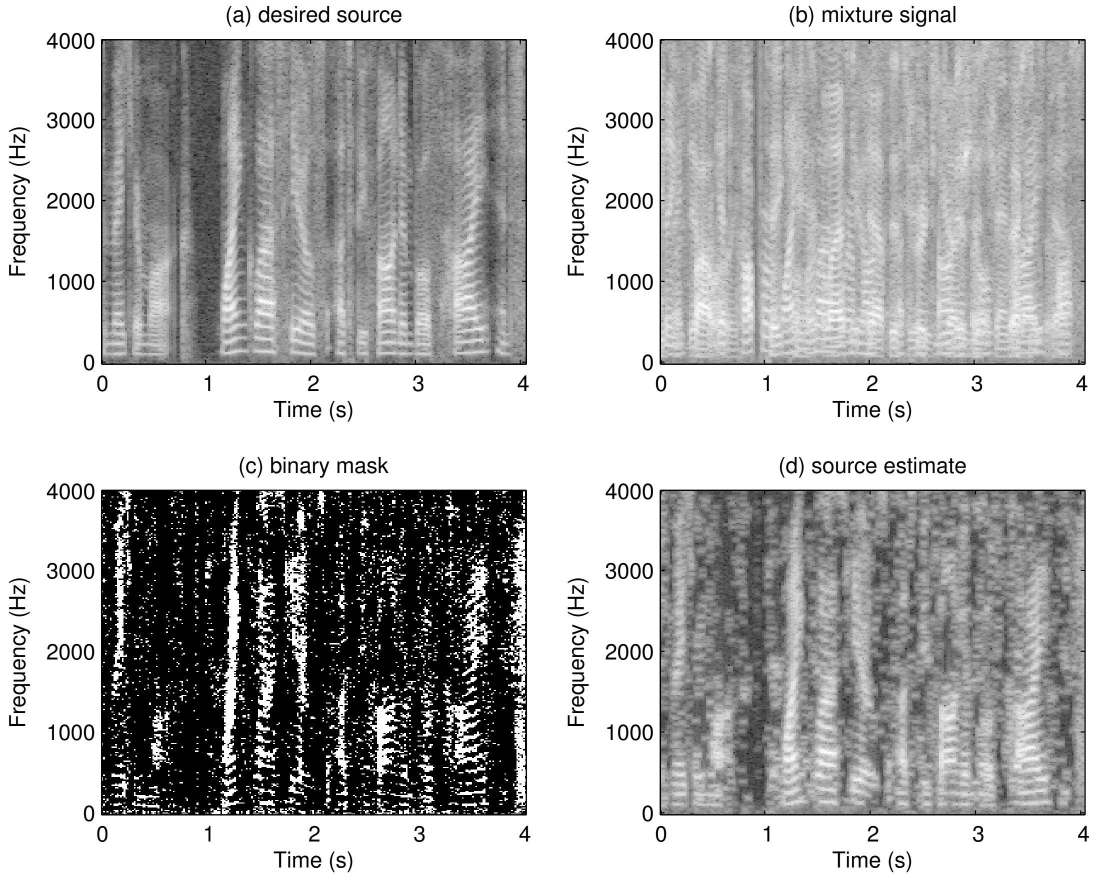

Figure 2.9: *Time-frequency masking. Plot (a) shows the spectrogram of the desired source. The brighter the time-frequency point, the higher energy it has. Plot (b) shows the spectrogram of the mixture signal. Plot (c) shows the time-frequency mask used to extract the desired source. White parts of the mask indicate selected region. Plot (d) shows the spectrogram of the desired source estimate.*

A popular method to estimate the time-frequency masks using only two microphones is the degenerate unmixing estimation technique (DUET) [19, 20, 96]. It is assumed that the time-frequency representation of speech signals are approximately disjoint (i.e., sources do not overlap too much):

$$s_i(n,f)s_j(n,f) \simeq 0, \qquad \forall i \neq j, \forall f, \forall n \tag{2.45}$$

This assumption is not fully met in practice, but speech signals exhibit a level of approximate windowed orthogonality [97]. If we consider the anechoic model, and assume that at any time-frequency point only one source is active, we get:

$$\begin{bmatrix} x_1(n,f) \\ x_2(n,f) \end{bmatrix} = \begin{bmatrix} \alpha_{1j} e^{-j2\pi f \Delta_{1j}} \\ \alpha_{2j} e^{-j2\pi f \Delta_{2j}} \end{bmatrix} s_j, \text{ for active source } j \tag{2.46}$$

We can see that when only one source is active, the ratio of the STFT of the two mixtures at any time-frequency point depends only on the spatial information embedded in the channel parameters associated with the source active at that time-frequency point. This spatial information is given by the inter-channel level difference (ILD):

$$\text{ILD}(n,f) = \left| \frac{x_2(n,f)}{x_1(n,f)} \right| \tag{2.47}$$

which is also termed inter-channel intensity difference (IID), and the inter-channel phase difference (IPD):

$$\text{IPD}(n,f) = \angle \left(\frac{x_2(n,f)}{x_1(n,f)} \right) \tag{2.48}$$

where $\angle(.)$ denotes the phase of a complex number in $(-\pi, \pi]$. There is a phase ambiguity in the IPD that follows from the periodicity of the complex exponential. A given value of IPD(n, f) leads to several possible values of relative delay between the two microphones $\Delta_{2j} - \Delta_{1j}$. This ambiguity is a problem at frequencies above $f = \frac{1}{2\Delta_{\text{max}}}$, where Δ_{max} is the maximum possible delay between the two microphones. To avoid this ambiguity, the microphones should be separated by less than $c/(2f_{\text{max}})$ where f_{max} is the maximum frequency present in the sources and c is the speed of sound. For example, for a maximum frequency of 8 kHz the microphones spacing must be no more than 2.1 cm. The ILD has the property $1 > \text{ILD}(n, f) > 0$ if the signal on microphone 1 is louder, and $\infty > \text{ILD}(n, f) > 1$ if the signal on microphone 2 is louder. Therefore, a symmetric attenuation estimate SA(n, f) is estimated from ILD(n, f) as follows [19]:

$$\text{SA}(n, f) = \text{ILD}(n, f) - \text{ILD}(n, f)^{-1} \tag{2.49}$$

The symmetric attenuation has the property that if the microphone signals are swapped, the attenuation is reflected symmetrically about a centre point (SA $= 0$).

The DUET algorithm calculates the relative attenuation and delay values between the two observations and constructs a two dimensional histogram $(\text{SA}(n, f), \text{IPD}(n, f)/(2\pi f))$. The histogram is weighted (by the power of each time-frequency point) and smoothed [19, 20]. In DUET, each time-frequency point in proximity of a peak centre in the histogram is assigned to the source corresponding to that peak, and a time-frequency binary mask is constructed to perform separation of the corresponding source from the time-frequency representation of the mixtures. The original method works only for closely arranged microphones in order to avoid the phase ambiguity problem. Two extensions to DUET that allow for arbitrary microphone spacing were presented in [20]. DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. Despite its simplicity, DUET remains one of the state of the art techniques for underdetermined speech separation. However, this method suffers from so-called musical noise or burbling artifacts due to binary masking of time-frequency points where the sources overlap. These distortions are introduced by the separation mask, and arise from the spectral components of the signal being turned on and off, which results in sinusoidal components that come and go in each short-time frame.

In [21, 92–95], probabilistic models are used to model the IPD/ILD, and after estimating its parameters with an EM algorithm [30], soft masks are derived. All of these methods require the number of sources to be given a priori, and it is difficult to expand these methods to more than two microphones. Furthermore, separation methods based on time-frequency masking suffer from the fact that clustering becomes difficult in reverberation, as the ILD/IPD resulting from each sound source then tend to spread and overlap, and the disjoint assumption becomes less realistic.

In order to utilise the information provided by more than two microphones in designing the time-frequency mask, the Multiple sENsor dUET (MENUET) method was proposed in [98, 99]. It was proposed to cluster the normalised observation vectors. The normalisation is performed by selecting a reference microphone I and evaluating for each channel

$$\bar{x}_i(n,f) = |x_i(n,f)| \exp\left[j\frac{\angle\left(x_i(n,f)/x_I(n,f)\right)}{4fc^{-1}d_{\text{max}}}\right] \tag{2.50}$$

where c is the speed of sound, and d_{max} is the maximum separation between the reference microphone I and any other microphone. Then a unit-norm normalisation is applied to prevent outliers in the level ratio affecting the clustering performance [99]:

$$\bar{\mathbf{x}}(n,f) \leftarrow \frac{\bar{\mathbf{x}}(n,f)}{\|\bar{\mathbf{x}}(n,f)\|} \tag{2.51}$$

By this normalisation, the vectors $\bar{\mathbf{x}}(n,f)$ are N-dimensional complex vectors and dependent only on the source geometry. Clustering can then be performed in an N-dimensional complex space with the k-means algorithm [100]. This method can be applied to non-linear microphone arrangements with 2- or 3-dimensional arrays.

2.8.5 Computational Auditory Scene Analysis (CASA)

The previously presented techniques address the source separation problem from a purely mathematical point of view. Another possible approach to address the source separation problem is to study and finally mimic the way humans perform audio source separation using a computer. In fact, separation methods like time-frequency masking existed in the computational auditory scene analysis (CASA) community for quite some time before it attracted attention in the signal processing community.

The human auditory system has a remarkable ability to perform source separation in real-time using only the sounds acquired from our ears. In the human auditory system, it is believed that the basilar membrane in the cochlea performs a time-frequency analysis of the sound [12]. This segmentation of an auditory signal into small components in time and frequency is followed by a grouping where each component is assigned a certain auditory stream that combines segments that are likely to have originated from the same audio source [6]. This process is termed auditory scene analysis. The human auditory system is very good at paying attention to a single auditory stream at a time. Psychological and psychoacoustic research uncovered a number of cues or grouping rules which describe how to group different parts of an audio signal into a single stream [6]. The following auditory cues are examples of these cues: i) spatial cues such as

ITD and ILD ii) common onset characteristics, (i.e., energy appearing at different frequencies at the same time), iii) amplitude or frequency modulations, iv) harmonicity or periodicity, v) proximity in time and frequency, vi) continuity (i.e. temporal coherence). Humans also rely on the knowledge of language to restore words or sentences interrupted by noise bursts.

CASA refers to the set of algorithms developed with the aim of simulating auditory scene analysis processes [9, 10]. CASA methods perform source separation in several stages. First, the acoustic mixture is split into several subbands using a perceptually motivated filter bank that simulates the basilar membrane, and the vibrating hair cells [11, 12]. The bandwidth of each filter varies proportionately to its centre frequency. This results in a time-frequency representation called the cochleagram. Sometimes, an autocorrelation function of the absolute value of each subband signal on short time frames is computed, resulting in a three-dimensional representation known as the correlogram [33]. Time-frequency points are then grouped into small clusters, each associated with one source, by applying ASA grouping rules. The following auditory cues are examples of these grouping rules applied in CASA methods: common periodicity (Pitch) [101], spectral proximity and common onsets and offsets [102], common modulations [103], common spatial origin [104], and continuity [105]. Further processing may be implemented using knowledge of language or speaker's timbre. The source signals are eventually extracted by binary time-frequency masks.

2.8.6 Single Microphone

When only one microphone is available, source separation becomes more challenging, as in this case spatial cues are absent. In this situation, the assumptions of independence and time-frequency sparsity becomes insufficient and more advanced source models relying on spectro-temporal cues are needed. The models typically represent the magnitude of the source STFTs, and the spectro-temporal cues considered may include structures such as phonemes and temporal continuity characteristics.

One popular approach to model speech and monophonic music signals is to assume that the local power spectrum density (PSD) $V_{jn}(f)$ of a source j at a given time frame n is one of Q local PSD templates $r_{jq}(f)$ indexed by state q, which may represent a certain sound event. The states underlying different sources in a mixture are generally modeled as independent.

Denoting by c_{jq} the prior probability of state q of source j, this yields the mixture model:

$$V_{jn}(f) = r_{jq}(f), \quad \text{with probability } c_{jq} \tag{2.52}$$

This model is called spectral Gaussian mixture model (GMM) when $V_{jn}(f)$ is a parameter of a Gaussian distribution [22]. If the each state of source depends on the previous state via a set of transition probabilities, the model is called hidden Markov model (HMM) [106]. The Gaussian scaled mixture model (GSMM) generalises the GMM model by multiplying each PSD template by a time-varying scale factor to account for recurring PSD shapes but with variable intensities [22].

GMMs, HMMs, and GSMM have been applied to the separation of single microphone mixtures [22–24]. In general, they require prior training and some knowledge about the identity of the speech or music sources in the mixture. In [106], an individual HMM is trained for each source beforehand. Then the most probable mixture state sequence is inferred from the mixture signal, and the spectrum of each source is derived from the source state sequence. Finally, the corresponding signals are computed by attributing each time-frequency point in the mixture to the loudest source. In [23, 24], the problem of singing voice extraction from mono audio recordings was studied. The recording is segmented in a succession of vocal and non-vocal parts. Then an adapted music GMM model is learned on the non-vocal parts. Finally, using the adapted music model as a prior, an adapted voice GMM model is learned from the vocal parts. Separation is then performed in the STFT domain with MMSE estimators (soft masks). In [22], GMMs are learned for samples of each source separately (prior training). Then the GSMM scale factors are estimated in a ML scheme under positivity constraints, and finally MMSE estimates of the sources are derived given the observed audio mixture and the scale factors.

2.8.7 Post-Processing

In order to improve the performance of source separation and extraction algorithms, it was suggested to post-process their output. The application of a post-filter can improve the quality of the output signal by suppression of residual interference and noise. However, in many cases, the suppression of residual interference and noise can lead to musical noise.

In [107], it was proposed to compare the separated outputs at each time-frequency point. Based on the assumption of disjoint sources (see (2.45)), only one of the outputs should have a non-zero value at any time-frequency point. Therefore, at each time-frequency point, only the output with the greatest magnitude is retained, and the other outputs are set to zero. This method has negligible computational complexity, and offers improvements in SIR at the expense of artifacts. However, this method can only be applied in source separation algorithms where the mixture signal is decomposed into the original source signals, and can not be applied when one desired source is extracted.

In [108], it was proposed to perform post-processing based on an energy-normalised SIR (EN-SIR). This ratio compares the energy of each extracted signal and the difference signal between the observed mixture and the extracted source (an estimate of the interference). The extracted source and the interference energies are then compared at each time-frequency point, and a time-frequency point is nulled or attenuated if the ratio is lower than a certain threshold. The choice of the threshold is a tradeoff between interference attenuation and artifacts. As the value of the threshold increases, the number of nulled or attenuated time-frequency points increases. This technique can be applied to sources which have been extracted using any source separation or extraction algorithm.

2.9 Chapter Summary

This review was a non-exhaustive study of some of the main concepts and methods used for audio source separation and extraction, particularly those which are relevant to this thesis. We have described the different possibilities of source separation environments and prior knowledge of the sources, mixing process, and the microphones. Properties of speech signals were presented, followed by a description of room acoustics. We then described the instantaneous, anechoic, and echoic models. In Section 2.7, we described source separation evaluation measures. This was followed by a study of the various methods used in source separation and extraction.

In summary, spatial diversity is the main assumption used in determined and overdetermined mixtures. Most classical beamforming techniques require knowledge of the array geometry and the desired source location, but they do not make assumptions about the number of interfering sources and their location. The separating filters can be adapted to minimise the output power subject to a response constraint (such as unity gain in the look-direction). On the other hand, ICA methods use unsupervised adaptive filtering in order to blindly form an adaptive null beamformer that reduces the undesired sources by forming a spatial null towards them. ICA methods cannot perform separation of under-determined mixtures as there are not enough degrees of freedom to null interferers. Both methods are multichannel linear filtering methods, and the number of degrees of freedom is limited by the number of mixture channels. This explains the poor performance of linear techniques in under-determined mixtures. In under-determined mixtures, non-linear techniques which exploit the sparseness of speech sources and time-frequency diversity play a vital role. However, these methods typically suffer from musical noise. Furthermore, in reverberant environments, the approximation of time-frequency disjoint sources does not hold at the microphone array, and the separation performance deteriorates.

Chapter 3

Mixtures of Beamformers

3.1 Introduction

In this chapter, we present a framework which extends the use of beamforming techniques to underdetermined speech mixtures. We describe frequency domain non-linear mixtures of beamformers that can extract a speech source from a known direction when there are fewer microphones than sources (the underdetermined case), and do not require knowledge of the number of speakers. These beamformers utilise GMMs to model the data in each frequency bin. This, in turn, can be learned using the EM algorithm. The model learning is performed using the observed mixture signals only, and no prior training is required. The signal estimator comprises of a set of MMSE, MVDR, or MPDR beamformers. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. These weights are specific to each time-frequency point. This approach results in a soft decision filter for the observed signal. The resulting non-linear beamformer combines the benefits of non-linear time-varying separation in time-frequency masking with the benefits of spatial filtering in the linear beamformers.

The remainder of this chapter is structured as follows. Section 3.2 presents the signal mixing model used in this chapter. Section 3.3 reviews the MMSE estimator when the signals are assumed to be Gaussian. Then, in Section 3.4, the proposed GMM-based non-linear beamformers are described. We present some simulation results that illustrate the performance of proposed methods in Section 3.5. More experiments and comparisons of the proposed methods with other source separation algorithms can also be found in Chapter 5. In Section 3.6, we give the conclusions.

3.2 Mixing Model

In this chapter, we consider the convolutive mixing model in (2.14), repeated here for convenience:

$$\mathbf{x}(n,f) = \mathbf{a}(f)s(n,f) + \mathbf{v}(n,f) \tag{3.1}$$

where $\mathbf{x}(n,f) = [x_1(n,f),...,x_N(n,f)]^T$ is the observed multichannel mixture signal, $\mathbf{a}(f)$ is the $N \times 1$ array response vector in the direction of the desired source signal s (also called the propagation vector or steering vector), and $\mathbf{v}(n,f) = [v_1(n,f),...,v_N(n,f)]^T$ is the $N \times 1$ vector of the contribution of the interferers and any possible noise to the mixture signal. This model is suitable when we are only interested in extracting one source of interest. In this model, no assumptions are made about the interferers or their number. The interferers can be of any nature such as point sources, spatially extended sources, diffuse sources, or a combination of them.

Note that $\mathbf{x}$, $\mathbf{a}$, s, and $\mathbf{v}$ are complex valued, and depend on frequency f, but for readability and simplicity, we will omit this variable in the rest of the chapter. From now on, we implicitly work in a given frequency band, and we will use this notation:

$$\mathbf{x} = \mathbf{a}s + \mathbf{v} \tag{3.2}$$

3.3 Optimum Beamformers

In the previous chapter (Section 2.8.1.1), we discussed some of the well known statistically optimum beamformers. In particular, we presented a derivation of the MVDR beamformer by modeling the desired source signal as an unknown nonrandom signal arriving to the array from a known direction and designing the optimum filter that minimises the output power and passes any signal arriving to the array from the specified direction through the filter undistorted. We then presented a derivation for the optimum linear beamformer to estimate the signal waveform using a MMSE criterion. And finally, we considered the beamformer MPDR beamformer, which assumes that we know or can measure the statistics of the observed mixture signal $\mathbf{x}$, but do not know the statistics of the interference component.

In this section, we consider the case in which the desired source and interference processes are Gaussian and design the MMSE estimate of the signal [52]. In contrast to the derivation of the linear MMSE estimator in the previous chapter, the MMSE estimator is not constrained to be linear, however the signals are assumed to be Gaussian. This will serve as a good starting point for our mixture of beamformers method.

3.3.1 Minimum Mean Square Error (MMSE) Beamformer

We consider the optimum filter whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference, assuming a known desired signal direction. We assume that the desired source signal is a sample function from a zero mean complex valued Gaussian random process, $s \sim G(0, \sigma_s^2)$, where σ_s^2 is the known signal variance. We also assume a zero mean complex valued Gaussian interference, $\mathbf{v} \sim G(0, \mathbf{R}_v)$, where $\mathbf{R}_v$ is the signal covariance matrix. Additionally, it is assumed that the signal and interference snapshots are uncorrelated. Hence, $\mathbf{x} \sim G(0, \mathbf{R}_v + \sigma_s^2 \mathbf{a}\mathbf{a}^H)$, and $\mathbf{x}|s \sim G(\mathbf{a}s, \mathbf{R}_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}^{\text{MMSE}} = \mathrm{E}\left[s|\mathbf{x}\right] = \int s \; p(s|\mathbf{x}) \; ds \tag{3.3}$$

This mean is referred to as the conditional mean. Using Bayes' theorem, the a posteriori density can be expressed as:

$$\begin{aligned} p(s|\mathbf{x}) &= \frac{p(\mathbf{x}|s).p(s)}{p(\mathbf{x})} \\ &\propto p(\mathbf{x}|s).p(s) \\ &\propto \exp\left(-(\mathbf{x} - \mathbf{a}s)^H \mathbf{R}_v^{-1} (\mathbf{x} - \mathbf{a}s) - s^* \sigma_s^{-2} s\right) \end{aligned} \tag{3.4}$$

We can see that $p(s|\mathbf{x})$ is a Gaussian in the general form. The MMSE estimator is the expectation of $p(s|\mathbf{x})$, which is the mean μ_s of the a posteriori density:

$$p(s|\mathbf{x}) \propto \exp\left(-(s - \mu_s)^* \widetilde{\sigma}^{-2} (s - \mu_s)\right) \tag{3.5}$$

The variance $\widetilde{\sigma}^2$ of the a posteriori density can be found by equating the quadratic terms in (3.4) and (3.5) together:

$$-s^* \widetilde{\sigma}^{-2} s = -s^* s \left(\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a} + \sigma_s^{-2}\right) \tag{3.6}$$

$$\widetilde{\sigma}^2 = \left(\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a} + \sigma_s^{-2}\right)^{-1} \tag{3.7}$$

The conditional mean μ_s can be found by equating the linear terms of (3.4) and (3.5) together:

$$\widetilde{\sigma}^{-2} \mu_s s^* + \widetilde{\sigma}^{-2} \mu_s^* s = \mathbf{x}^H \mathbf{R}_v^{-1} \mathbf{a} s + \mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{x} s^* \tag{3.8}$$

Thus, the conditional mean is:

$$\mu_s = \widetilde{\sigma}^2 \mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{x} = \frac{\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{x}}{\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a} + \sigma_s^{-2}} \tag{3.9}$$

The conditional mean can alternatively be expressed as [52]:

$$\hat{s}^{\text{MMSE}} = \underbrace{\frac{\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{x}}{\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a}}}_{\text{MVDR}} \; \underbrace{\frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{a}^H \mathbf{R}_v^{-1} \mathbf{a}\right)^{-1}}}_{\text{Wiener post filter}} \tag{3.10}$$

We can see that when Gaussian signals are assumed, the optimum MMSE processor will be the linear processor derived in Section 2.8.1.1 (Equation (2.40)). The first term is an MVDR spatial filter, which suppresses the interfering signals without distorting the signal propagating along the desired source direction. The second term is a single channel Wiener post-filter.

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear (as in Section 2.8.1.1), or all the signals are Gaussian. However, speech sources are generally non-Gaussian. This suggests extending the optimum beamformers to exploit the non-Gaussianity of speech signals.

3.4 Mixture of Beamformers

In the time-frequency domain, speech signals typically have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some time-frequency points, not all speech sources in a mixture are equally active at the same time-frequency points. It is therefore advantageous to exploit the sparsity property of speech signals in the time-frequency domain in order to perform separation in underdetermined environments.

In the previous section, we considered the MMSE estimator when the desired source and the interference are Gaussian. The MMSE estimator in this case is linear. However, speech sources are generally non-Gaussian. In this section we extend the MMSE estimator to deal with the non-Gaussianity of speech signals. In this section, we use GMMs to model the speech non-Gaussianity and the spatial distribution of the sources. GMMs are widely used to model complex densities in terms of simpler Gaussian densities, and are used because they are sufficiently general to model arbitrary distributions. Another advantage of GMMs is that they can be mathematically convenient because the individual Gaussian components can be easily studied.

Figure 3.1 shows the histogram of the real value of the STFT coefficients of a speech signal at one frequency bin. It can be seen that only few time frames have a value significantly different from zero (super-Gaussian probability distribution).

Figure 3.2 shows an example of a 2-dimensional GMM of four components fitted on the scatter plot of a two channel instantaneous mixture of three super-Gaussian distributed sources. In this Figure, three components (red, green, and blue) represents the three sources when they are active, and one component (yellow) represents the case when the three sources are inactive.

In this section, we present three non-linear beamformers that can perform underdetermined speech separation. The first two non-linear beamformers are based on modeling the desired source signal s and the interference $\mathbf{v}$ separately. The desired source signal in each frequency bin is modeled using a 1-dimensional GMM, and the observed interference in each frequency bin is modeled using an N-dimensional GMM. The third non-linear beamformer is based on modeling the observed mixture signal $\mathbf{x}$ in each frequency bin using an N-dimensional GMM.

3.4.1 Mixture of MMSE and MVDR Beamformers

We shall describe the density of the interference signal $\mathbf{v}$ in each frequency bin as a mixture of k_v zero mean, complex valued, N-dimensional Gaussians with indices $q_v = 1, ..., k_v$, covariances $\mathbf{R}_{v,q_v}$, and mixing proportions c_{v,q_v}:

$$p(\mathbf{v}|\theta_v) = \sum_{q_v=1}^{k_v} c_{v,q_v} \frac{1}{\pi^N \det \mathbf{R}_{v,q_v}} \exp\left(-\mathbf{v}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{v}\right) \tag{3.11}$$

where $\det$ denotes a matrix determinant, $\theta_v = \{c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_v \leq k_v\}$, and the mixing proportions $c_{v,q_v} = p(q_v)$ (prior probabilities of the Gaussian states) are constrained to sum to

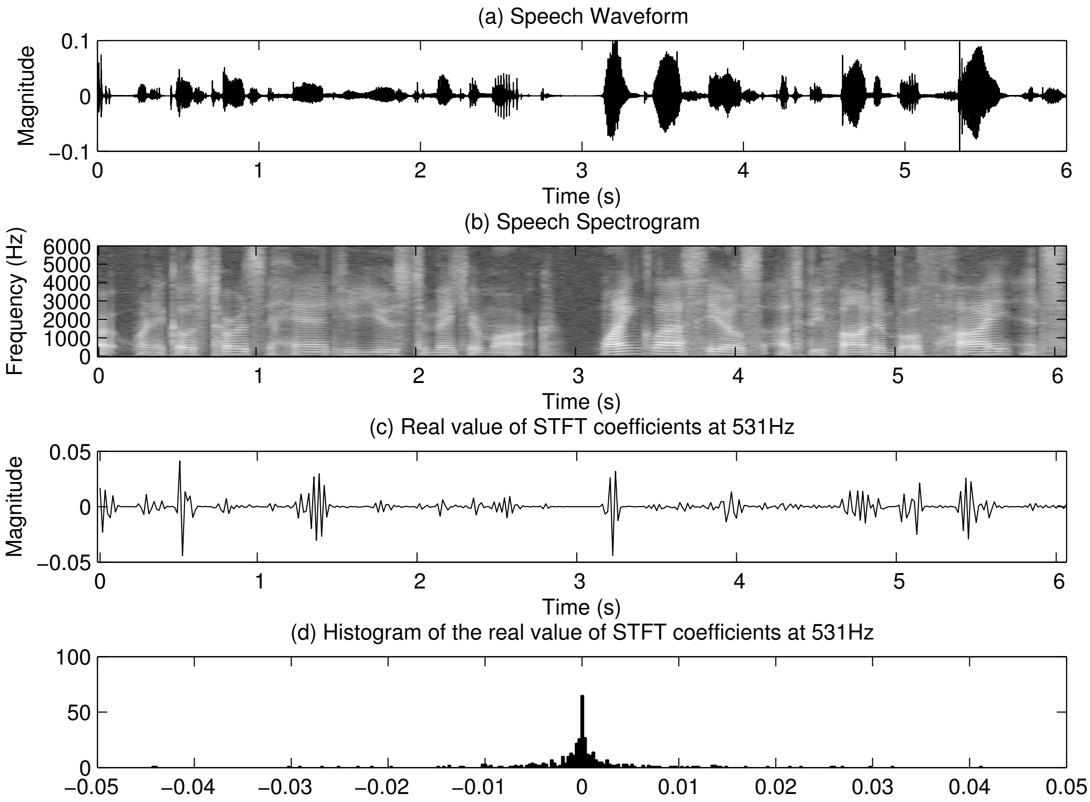

Figure 3.1: *(a) speech waveform. (b) speech spectrogram calculated from the STFT coefficients. (c) real value of the STFT coefficients at one frequency bin. (d) histogram of the real value of the STFT coefficients at one frequency bin.*

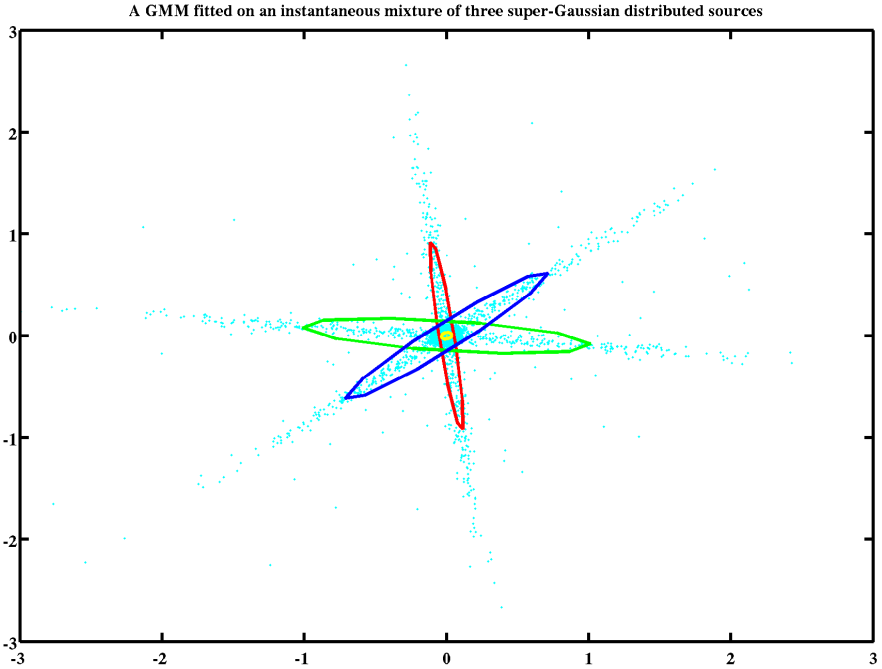

Figure 3.2: *Scatter plot of an instantaneous mixture of three super-Gaussian distributed sources.*

one. In addition, we shall describe the density of the desired source signal s in each frequency bin as a mixture of k_s zero mean complex valued 1-dimensional Gaussians with indices $q_s = 1, ..., k_s$, variances σ^2_{s,q_s}, and mixing proportions c_{s,q_s}:

$$p(s|\theta_s) = \sum_{q_s=1}^{k_s} c_{s,q_s} \frac{1}{\pi\sigma^2_{s,q_s}} \exp\left(\frac{-|s|^2}{\sigma^2_{s,q_s}}\right) \tag{3.12}$$

where $\theta_s = \{c_{s,q_s}, \sigma^2_{s,q_s} : 1 \leq q_s \leq k_s\}$, and the mixing proportions $c_{s,q_s} = p(q_s)$ (prior probabilities of the Gaussian states) are constrained to sum to one. The number of components k_s and k_v control the flexibility of the model. In our model, the Gaussian states are not coupled across frequency, and the parameters $\{\theta_s, \theta_v\}$ are frequency dependent.

The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{\text{MMSE}} &= \mathrm{E}\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(s, q_s, q_v|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x})\ p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}) \int p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \mathrm{E}\left[s|\mathbf{x}, q_s, q_v\right]
\end{aligned} \tag{3.13}$$

where

$$\begin{aligned}
\tau_{q_s,q_v} &= p(q_s, q_v|\mathbf{x}) \\
&= \frac{p(\mathbf{x}|q_s, q_v)\ p(q_s)\ p(q_v)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p(\mathbf{x}|q_s', q_v')\ p(q_s')\ p(q_v')}
\end{aligned} \tag{3.14}$$

is the a posteriori probability that - the desired source GMM model being in state q_s and the interference GMM model being in state q_v - when observing $\mathbf{x}$, with $\sum_{q_s} \sum_{q_v} \tau_{q_s,q_v} = 1$. The posteriori probability is specific to each time frequency point, and has a non-linear dependency on the observed data.

We can see that the conditional mean $\mathrm{E}\left[s|\mathbf{x}, q_s, q_v\right]$ is the linear MMSE beamformer estimator in (2.40), with $\mathbf{R}_v = \mathbf{R}_{v,q_v}$ and $\sigma_s^2 = \sigma_{s,q_s}^2$. The desired signal estimator in (3.13) is a non-linear weighted sum of linear MMSE beamformers over all the GMM components, and the weighting coefficients are the a posteriori probabilities of the GMM components τ_{q_s,q_v} (specific to each time-frequency point). This mixture of MMSE beamformers will be denoted by $\mathbf{w}_1$ and is given by [109]:

$$\mathbf{w}_1 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\sigma_{s,q_s}^2}{\sigma_{s,q_s}^2 + \left(\mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}\right)^{-1}} \frac{\mathbf{a}^H \mathbf{R}_{v,q_v}^{-1}}{\mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}} \tag{3.15}$$

In comparison to independent factor analysis [88], where sources were also modeled with GMMs, the mixture of MMSE beamformers models all the interfering sources using one N-dimensional mixture of Gaussians in the observation (microphones) domain. Consequently, the number of interferers in the mixture is not required to be known or have a unique mixing structure. This also avoids the exponential growth of the number of Gaussian components in the observation density with the number of sources.

If a distortionless response in the direction of the desired source is required, a distortionless response mixture of MVDR beamformers can be used [109]:

$$\mathbf{w}_2 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\mathbf{a}^H \mathbf{R}_{v,q_v}^{-1}}{\mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}} \tag{3.16}$$

This mixture of MVDR beamformers is a non-linear weighted sum of linear distortionless MVDR beamformers, where the weights sum to unity. As a result, it is constrained to a distortionless response in the look-direction. By distortionless we mean it has a unity gain in the look-direction at all time-frequency points. Therefore, the signal arriving from the look-direction will pass through the filter without any distortion.

In practice, the computation of $\mathbf{R}_{v,q_v}$ might give an ill-conditioned or singular matrix. A matrix is ill-conditioned if the ratio of the largest to smallest singular value is too large, and singular if it is infinite. Various regularisation techniques can be used to help avoid matrix singularity and to improve robustness to steering vector offsets (such as errors in the assumed direction of arrival of the desired source) [62, 110, 111]. In our algorithm implementations, we use the multiplicative diagonal loading regularisation applied to the diagonal terms of the correlation matrix as suggested in [62]. In this regularisation method, we multiply each diagonal element

Algorithm 1 Separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$

1. Compute the STFT of the mixture $\mathbf{x}$.

2. Apply the EM algorithm (see Section 3.4.2) separately in each frequency bin to compute $\{\tau_{q_s,q_v}(n,f), \sigma^2_{s,q_s}(f), \mathbf{R}_{v,q_v}(f) : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$.

3. For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:

$$\hat{s}(n,f) = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}(n,f)\, \mathbf{w}_{q_s,q_v}(f)\, \mathbf{x}(n,f) \tag{3.17}$$

   where $\mathbf{w}_{q_s,q_v}(f)$ can be either a linear MMSE or a linear MVDR beamformer:

$$\mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) = \frac{\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f)}{\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f) \mathbf{a}(f)} \tag{3.18}$$

$$\mathbf{w}^{\text{MMSE}}_{q_s,q_v}(f) = H^{\text{Wiener}}_{q_s,q_v}(f)\, \mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) \tag{3.19}$$

   where the scalar, single channel Wiener post filter is given by:

$$H^{\text{Wiener}}_{q_s,q_v}(f) = \frac{\sigma^2_{s,q_s}(f)}{\sigma^2_{s,q_s}(f) + \left(\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f) \mathbf{a}(f)\right)^{-1}} \tag{3.20}$$

4. The corresponding time domain signal $\hat{s}$ is derived by an STFT inversion.

in the correlation matrix with $1+\beta$, where β is very small number (we use $\beta = 1e-3$). The effect of various values of β and other regularisation methods is studied in Section 5.2.6.

In Section 3.4.2, we develop an EM algorithm to learn the model density parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$.

We briefly summarise the main steps in the separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$ in Algorithm 1. Note that the model learning step is applied separately in each frequency bin, and that the Gaussian states' posterior probabilities are specific to each time-frequency point (no coupling across all frequencies).

3.4.2 Learning Interference and Desired Source Parameters

Using the EM algorithm, we can estimate the model density parameters from a set of observations $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$, where η is the number of time frames in each frequency bin (we remind the reader that we are working in each frequency bin independently). The EM algorithm is used to find a maximum likelihood estimate of parameters in probabilistic models with latent variables (incomplete data problems). In our case, $\mathbf{x}$ is the observed (or incomplete) data, and the latent variables are the state sequence of the Gaussian mixtures that indicate which Gaussian components are responsible for $\mathbf{x}(n)$. In EM terminology, the complete data is composed of both the observed data and the latent variables. The EM algorithm is an iterative algorithm, consisting of a linked pair of steps: (1) an expectation step (E-step), and (2) a maximisation step (M-step). In the E-step, we calculate the conditional expectation of the complete data log likelihood. The expectation is taken with respect to the conditional probability of the hidden data, given the observed data and the parameter values obtained in the previous iteration. In the M-step, the new estimates of the parameters are calculated to maximise the conditional expectation of the complete data log likelihood. As shown in [30], each EM iteration increases the incomplete (observed) data log likelihood, unless a local maximum has already been reached. Depending on initial parameter values, the EM algorithm may converge to a local maximum of the incomplete data log likelihood.

In this section, the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \le q_s \le k_s, 1 \le q_v \le k_v\}$ of the interference $\mathbf{v}$ and desired source s are estimated using the EM algorithm. These parameters are required for the non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ of (3.15) and (3.16). A more detailed derivation can be found in Appendix 3.A.

Let us define a complete data set $D_c = \{\mathbf{x}, s, q_s, q_v\}$ composed of both the observed and the latent data. If we were to actually have such a complete data set, we could define its log likelihood as:

$$l_c(\theta|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), s, q_s, q_v|\theta) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), s, q_s, q_v|\theta) \tag{3.21}$$

Given an initial value θ^0, the EM algorithm performs the following steps at each iteration l:

E-step: In the E-step, we compute the expectation of the complete data log likelihood:

$$Q(\theta, \theta^{l-1}) = \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}(n), \theta^{l-1}\right) \ln p(\mathbf{x}(n), s, q_s, q_v|\theta) \tag{3.22}$$

This reduces to computing the posterior probability of the GMM states $p\left(q_s, q_v|\mathbf{x}(n), \theta^{l-1}\right)$, and the conditional mean and variance of the desired source given both the observed mixture and the GMM states.

The posterior probability of the GMM states can be evaluated as follows:

$$\begin{aligned}\tau_{q_s,q_v}^{(l)}(n) &= p\left(q_s, q_v|\mathbf{x}(n), \theta^{(l-1)}\right) \\ &= \frac{p\left(q_s, q_v|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s, q_v, \theta^{(l-1)}\right)}{\sum_{q_s'=1}^{k_s}\sum_{q_v'=1}^{k_v} p\left(q_s', q_v'|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s', q_v', \theta^{(l-1)}\right)} \qquad (3.23)\end{aligned}$$

where

$$p(\mathbf{x}|q_s, q_v) = G\left(0, \mathbf{R}_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{a}\mathbf{a}^H\right) \qquad (3.24)$$

The conditional mean and variance of the desired source given both the observed mixture and the GMM states, which are denoted by $\langle s|\mathbf{x}(n), q_s, q_v\rangle$ and $\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle$ respectively, can be evaluated from the following density function:

$$\begin{aligned}p(s|\mathbf{x}, q_s, q_v) &= \frac{p(s|q_s)\, p(\mathbf{x}|s, q_v)\, p(q_s)\, p(q_v)}{p(\mathbf{x}|q_s, q_v)\, p(q_s)\, p(q_v)} \\ &= G\left(\alpha_{q_s,q_v},\ \beta_{q_s,q_v}\right) \qquad (3.25)\end{aligned}$$

where

$$\begin{aligned}\alpha_{q_s,q_v} &= \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}\right)^{-1} \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{x} \qquad (3.26) \\ \beta_{q_s,q_v} &= \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}\right)^{-1} \qquad (3.27)\end{aligned}$$

M-step: In the M-step, we maximise the expected complete log likelihood with respect to the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma_{s,q_s}^2, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. This can be done by taking derivatives with respect to θ and setting them to be equal to zero (under the constraints $\sum_{q_s=1}^{k_s} c_{s,q_s} = 1$ and $\sum_{q_v=1}^{k_v} c_{v,q_v} = 1$). This results in the following update rules:

$$c_{v,q_v}^{(l)} = \frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n) \qquad (3.28)$$

$$c_{s,q_s}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n) \tag{3.29}$$

$$\sigma_{s,q_s}^{2(l)} = \frac{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n) \; \langle ss^* | \mathbf{x}(n), q_s, q_v \rangle}{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n)} \tag{3.30}$$

$$\mathbf{R}_{v,q_v}^{(l)} = \frac{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n) \, \Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n)} \tag{3.31}$$

where

$$\begin{aligned} \Lambda_{q_s,q_v}(n) &= \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n) \langle s^* | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{a}^H \\ &\quad - \mathbf{a} \langle s | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{x}(n)^H \\ &\quad + \mathbf{a} \langle ss^* | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{a}^H \end{aligned} \tag{3.32}$$

In this model, there is an ambiguity in associating variance between the desired source and the interference. It is possible to incorporate some of the source signal into the interference. To deal with this, updating the desired source component variances is not performed in the first few iterations. This prevents the source components shrinking to zero variance.

3.4.3 Mixture of MPDR Beamformers

The model learning for non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ is dependent on the location of the desired source (look-direction). In some applications, scanning for the source direction is needed, and in this case, a learning algorithm which is independent of the look-direction is desired in order to reduce the computational complexity. By modeling the observed mixture directly instead of the desired source and interference signal separately, the model learning will be independent of the look-direction.

In this secion, we use a mixture of k_x zero mean, complex valued, N-dimensional Gaussians with indices $q_x = 1, ..., k_x$, covariances $\mathbf{R}_{x,q_x}$, and mixing proportions c_{x,q_x} to model the observed mixture $\mathbf{x}$ (the desired source and interference together) in each frequency bin:

$$p(\mathbf{x}|\theta_x) = \sum_{q_x=1}^{k_x} c_{x,q_x} \frac{1}{\pi^N \det \mathbf{R}_{v,q_v}} \exp\left(-\mathbf{x}^H \mathbf{R}_{x,q_x}^{-1} \mathbf{x}\right) \tag{3.33}$$

where $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x\}$, and the mixing proportions $c_{x,q_x} = p(q_x)$ (prior probabilities of the Gaussian states) are constrained to sum to one. This leads to a simple learning algorithm, and the learning of model parameters is independent on the desired source direction. The desired signal can be estimated using this mixture of MPDR beamformers [112]:

$$\mathbf{w}_3 = \sum_{q_x=1}^{k_x} \tau_{q_x} \frac{\mathbf{a}^H \mathbf{R}_{x,q_x}^{-1}}{\mathbf{a}^H \mathbf{R}_{x,q_x}^{-1} \mathbf{a}} \tag{3.34}$$

where

$$\begin{aligned} \tau_{q_x} &= p(q_x|\mathbf{x}) \\ &= \frac{p(\mathbf{x}|q_x)\ p(q_x)}{\sum_{q_x'=1}^{k_x} p(\mathbf{x}|q_x')\ p(q_x')} \end{aligned} \tag{3.35}$$

is the relative contribution for each linear MPDR beamformer, and is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. The resulting beamformer has a unity gain in the look-direction at all time-frequency points.

In practice, the computation of $\mathbf{R}_{x,q_x}$ might give an ill-conditioned or singular matrix. In our algorithm implementations, we use the multiplicative diagonal loading regularisation applied to the diagonal terms of the correlation matrix as suggested in [62].

In Section 3.4.4, we develop an EM algorithm to learn the observation model density parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$.

The main steps in the separation procedure using $\mathbf{w}_3$ are summarised in Algorithm 2.

3.4.4 Learning Observed Mixture Parameters

In this section, the parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$ of the observed mixture $\mathbf{x}$ are estimated using the EM algorithm. These parameters are required for the non-linear beamformer $\mathbf{w}_3$ defined in (3.34). A more detailed derivation can be found in Appendix 3.B.

Algorithm 2 Separation procedure using $\mathbf{w}_3$

1. Compute the STFT of the mixture $\mathbf{x}$.

2. Apply the EM algorithm (see Section 3.4.4) separately in each frequency bin to compute $\{\tau_{q_x}(n,f), \mathbf{R}_{x,q_x}(f) : 1 \leq q_x \leq k_x\}$.

3. For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:
$$\hat{s}(n,f) = \sum_{q_x=1}^{k_x} \tau_{q_x}(n,f)\,\mathbf{w}_{q_x}^{\text{MPDR}}(f)\,\mathbf{x}(n,f) \tag{3.36}$$
where:
$$\mathbf{w}_{q_x}^{\text{MPDR}}(f) = \frac{\mathbf{a}(f)^H\,\mathbf{R}_{x,q_x}^{-1}(f)}{\mathbf{a}(f)^H\,\mathbf{R}_{x,q_x}^{-1}(f)\,\mathbf{a}(f)} \tag{3.37}$$

4. The corresponding time domain signal $\hat{s}$ is derived by an STFT inversion.

Let us define a complete data set $D_c = \{\mathbf{x}, q_x\}$ composed of both the observed data $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$ and the latent data. If we were to actually have such a complete data set, we define its log likelihood as:

$$l_c(\theta_x|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), q_x|\theta_x) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), q_x|\theta_x) \tag{3.38}$$

The EM algorithm may be executed as follows:

E-step: In the E-step, we compute the expectation of the complete data log likelihood:

$$Q(\theta_x, \theta_x^{l-1}) = \sum_{n=1}^{\eta} \sum_{q_x=1}^{k_x} p\left(q_x|\mathbf{x}(n), \theta_x^{l-1}\right) \ln p(\mathbf{x}(n), q_x|\theta_x) \tag{3.39}$$

This reduces to calculating $p\left(q_x|\mathbf{x}(n), \theta_x^{l-1}\right)$, the posterior probability of the latent variables given the observed data and the current estimates of the parameters:

$$\begin{aligned} \tau_{q_x}^{(l)}(n) &= p\left(q_x|\mathbf{x}(n), \theta_x^{(l-1)}\right) \\ &= \frac{c_{x,q_x}^{(l-1)}\, G\left(0, \mathbf{R}_{x,q_x}^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} c_{x,q_x'}^{(l-1)}\, G\left(0, \mathbf{R}_{x,q_x'}^{(l-1)}\right)} \end{aligned} \tag{3.40}$$

M-step: In the M-step, we maximise the expected complete log likelihood with respect to the parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$. This can be done by taking derivatives with respect to θ_x and setting them to be equal to zero, while also including a Lagrangian term to account for the constraint that $\sum_{q_x=1}^{k_x} c_{q_x} = 1$. This results in the following update rules:

$$\mathbf{R}_{x,q_x}^{(l)} = \frac{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)\, \mathbf{x}(n)\, \mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)} \tag{3.41}$$

$$c_{x,q_x}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \tag{3.42}$$

3.5 Experimental Evaluation

3.5.1 Setup

In order to illustrate the performance of the non-linear beamformers, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [50]. The positions of the microphones and the sources were as illustrated in Figure 3.3. Two microphone arrays were used. The first has three microphones with a spacing $d = 2.5$ cm, and the second has two microphones with a spacing $d = 5$ cm. Both microphone arrays have a total length of $D = 5$ cm. We used speech files taken from the TIMIT speech corpus [113] to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of the sources in each mixture was four. The sources were placed in a semicircle of radius 1 m around the microphone arrays at angles $\phi = \{-45,\ -15,\ 10,\ 50\}°$.

3.5.2 Evaluation Measures

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we used the source to distortion ratio (SDR), source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [53]. The computation of the evaluation measures is described in Section 2.7.

In this chapter, the SDR, SIR and SAR values were averaged over all the sources and mixtures.

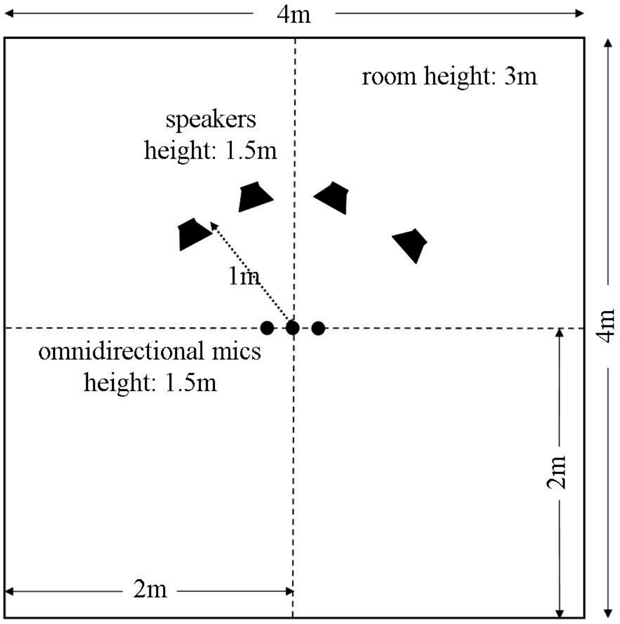

Figure 3.3: *Layout of room used in simulations.*

		$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
STFT frame		1024	1024	1024
STFT step		256	256	256
GMM components	(2 mics)	$k_s = 2$, $k_v = 15$	$k_s = 2$, $k_v = 15$	$k_x = 15$
	(3 mics)	$k_s = 2$, $k_v = 5$	$k_s = 2$, $k_v = 5$	$k_x = 5$
EM Iterations	(2 mics)	100	100	50
	(3 mics)	100	100	20
Learning block duration		10 s	10 s	10 s

Table 3.1: *Algorithm parameters for* $\mathbf{w}_1$, $\mathbf{w}_2$, *and* $\mathbf{w}_3$.

3.5.3 Algorithm Parameters

Unless mentioned otherwise, we use the values listed in Table 3.1 for the STFT frame size, STFT step size, number of GMM components, number of iterations, and learning block duration. It will become clear later in this chapter why these values where chosen. In particular, increasing the number of GMM components or the number of EM iterations, or reducing the STFT step size beyond the values listed in the table requires more computational time with an insignificant performance improvement.

3.5.4 Effect of Design Parameters

We first investigate the effect of various parameters on the performance of the non-linear beamformers. We study the effect of the number of Gaussian components in the GMM model, the required number of EM iterations, and the effect of the learning block size. In these experiments, we study the performance of the three non-linear beamformers when four sources are operating in an anechoic environment (RT=0), and the microphone array used has two or three microphones with a total array size $D = 5$ cm. We assume that the location of a desired source is known.

3.5.4.1 Effect of the Number of Gaussian Components

Figure 3.4 shows the average performance at the output of the mixture of MPDR beamformers $\mathbf{w}_3$ defined in (3.34) as a function of the number of Gaussian components k_x in the GMM model. The case of $k_x = 1$ is equivalent to a time-invariant MPDR beamformer. The SIR increases with k_x, but the improvement is insignificant at $k_x > 10$. The increase in the SIR is more pronounced in the two microphone case, where the separation using a time-invariant

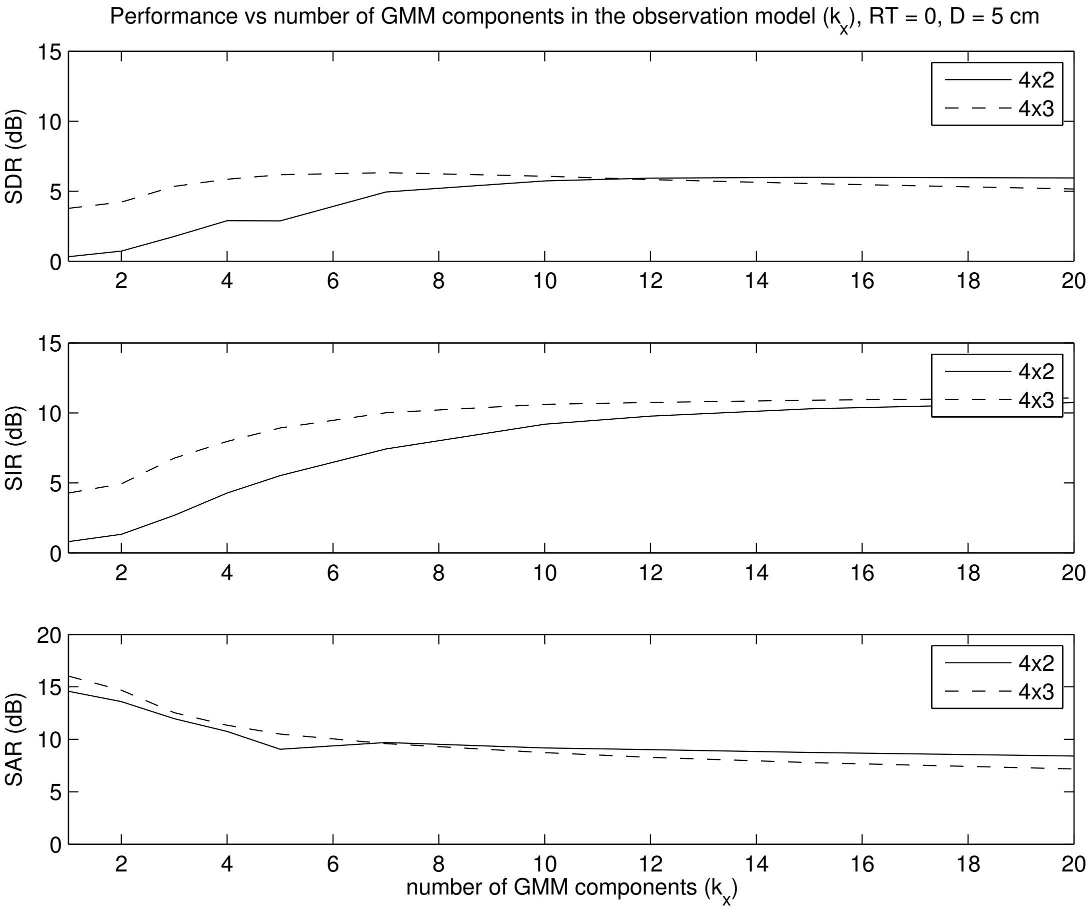

Figure 3.4: *Average performance of* $\mathbf{w}_3$ *in the anechoic case as a function of the number of Gaussian components* k_x *in the GMM model.*

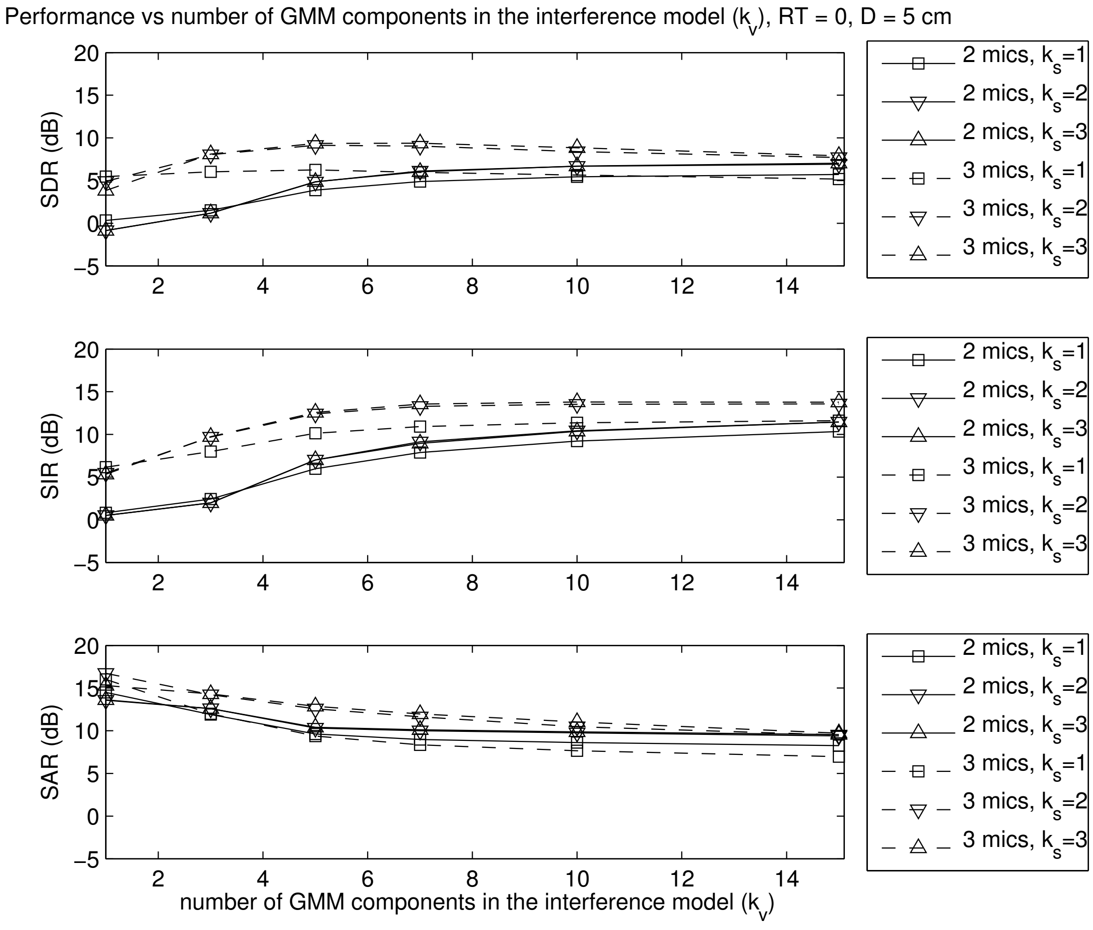

Figure 3.5: *Average performance of $\mathbf{w}_2$ in the anechoic case as a function of the number of Gaussian components k_v and k_s in the GMM model.*

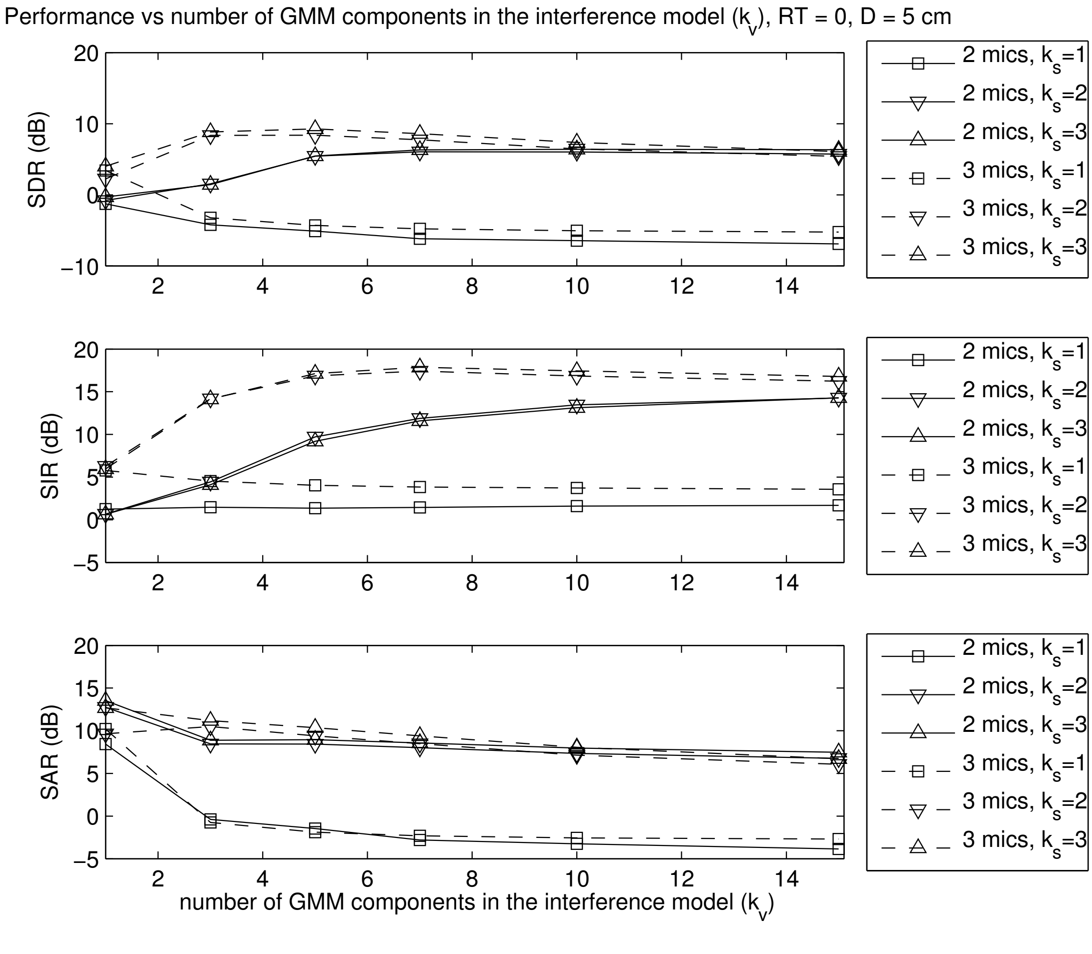

Figure 3.6: *Average performance of* $\mathbf{w}_1$ *in the anechoic case as a function of the number of Gaussian components* k_v *and* k_s *in the GMM model.*

beamformer $k_x = 1$ gives bad results. Although there is a unity gain response in the direction of the desired source signal, the SAR decreases with k_x. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals, and are not as harmful as distortions in the desired source signal. We stress that the mixture of MPDR beamformers is by definition distortionless in the look direction, and this is reflected in the output audio signal. The non-linear beamformer can attain a SIR of 10.7 dB in the two microphones case, and 11.1 dB using three microphones. The number of components in the GMM model is not directly related to the number of sources in the mixture and can be used to trade-off complexity with performance.

Figure 3.5 shows the average performance at the output of the mixture of MVDR beamformers $\mathbf{w}_2$ defined in (3.16) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. We can see that there is little to be gained in increasing the number of source Gaussian components k_s to more than two. In the two microphones case, The SIR increases with k_v, but the improvement is insignificant at $k_v > 10$. In the three microphones case, the SIR peaks around $k_v = 7$, and then levels off at higher k_v. The non-linear beamformer can attain a SIR of 11.5 dB in the two microphones case, and 14 dB using three microphones.

Figure 3.6 shows the average performance at the output of the mixture of MMSE beamformers $\mathbf{w}_1$ defined in (3.15) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. The non-linear beamformer can attain a SIR of 14.3 dB in the two microphones case, and 16.8 dB using three microphones. However, the SAR was reduced in comparison to Figure 3.5 because the distortionless constraint is no longer held.

3.5.4.2 Effect of the Number of Iterations

Figures 3.7 and 3.8 show the average performance at the output of the non-linear beamformers in the anechoic case as a function of the number of EM iterations. The non-linear beamformer $\mathbf{w}_3$ defined in (3.34) require less than 20 iterations to converge, whereas the other two non-linear beamformers require more iterations to converge (about 50 iterations).

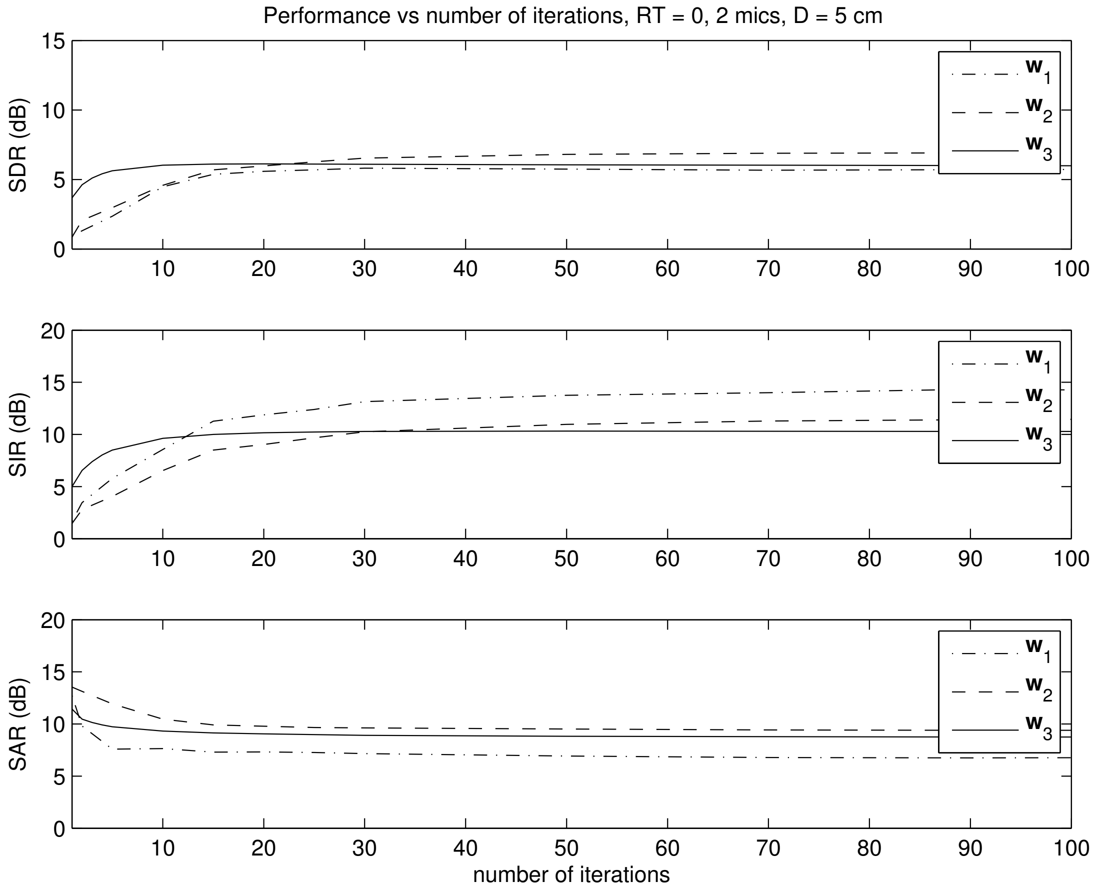

Figure 3.7: *Separation using two microphones: average performance of* $\mathbf{w}_1$*,* $\mathbf{w}_2$*, and* $\mathbf{w}_3$ *in the anechoic case as a function of the number of EM iterations. Used parameters:* $k_s = 2$ *and* $k_v = 15$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$*,* $k_x = 15$ *in* $\mathbf{w}_3$*.*

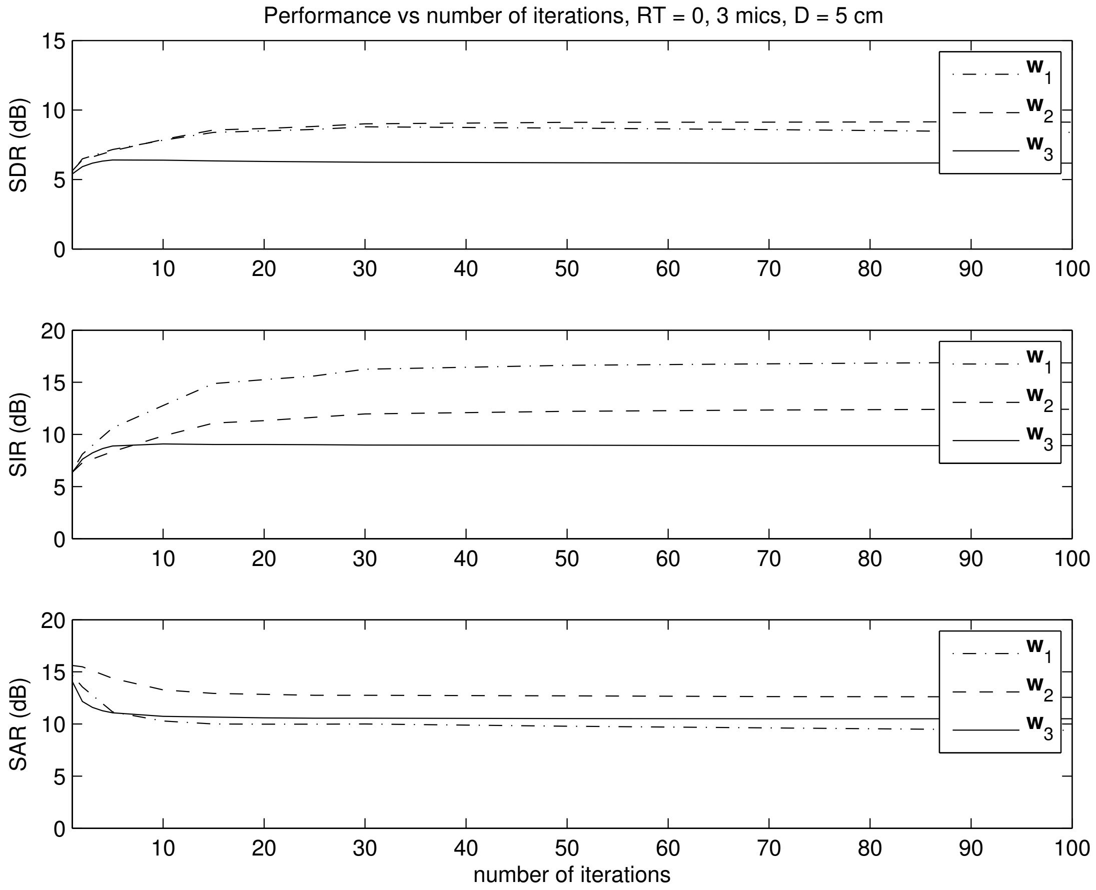

Figure 3.8: *Separation using three microphones: average performance of* $\mathbf{w}_1$*,* $\mathbf{w}_2$*, and* $\mathbf{w}_3$ *in the anechoic case as a function of the number of EM iterations. Used parameters:* $k_s = 2$ *and* $k_v = 5$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$*,* $k_x = 5$ *in* $\mathbf{w}_3$*.*

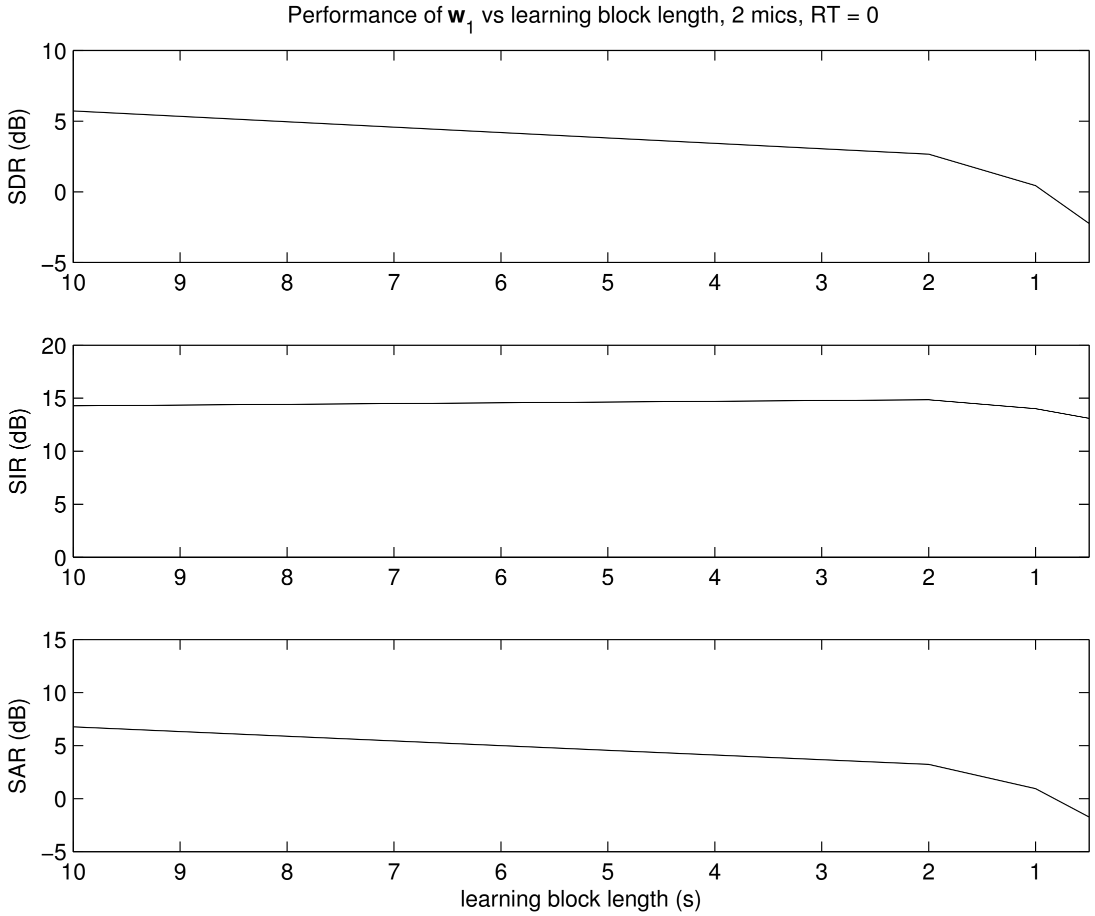

Figure 3.9: *Separation using two microphones: Average performance of* $\mathbf{w}_1$ *in the anechoic case vs learning block length in seconds. Used parameters:* $k_s = 2$ *and* $k_v = 15$.

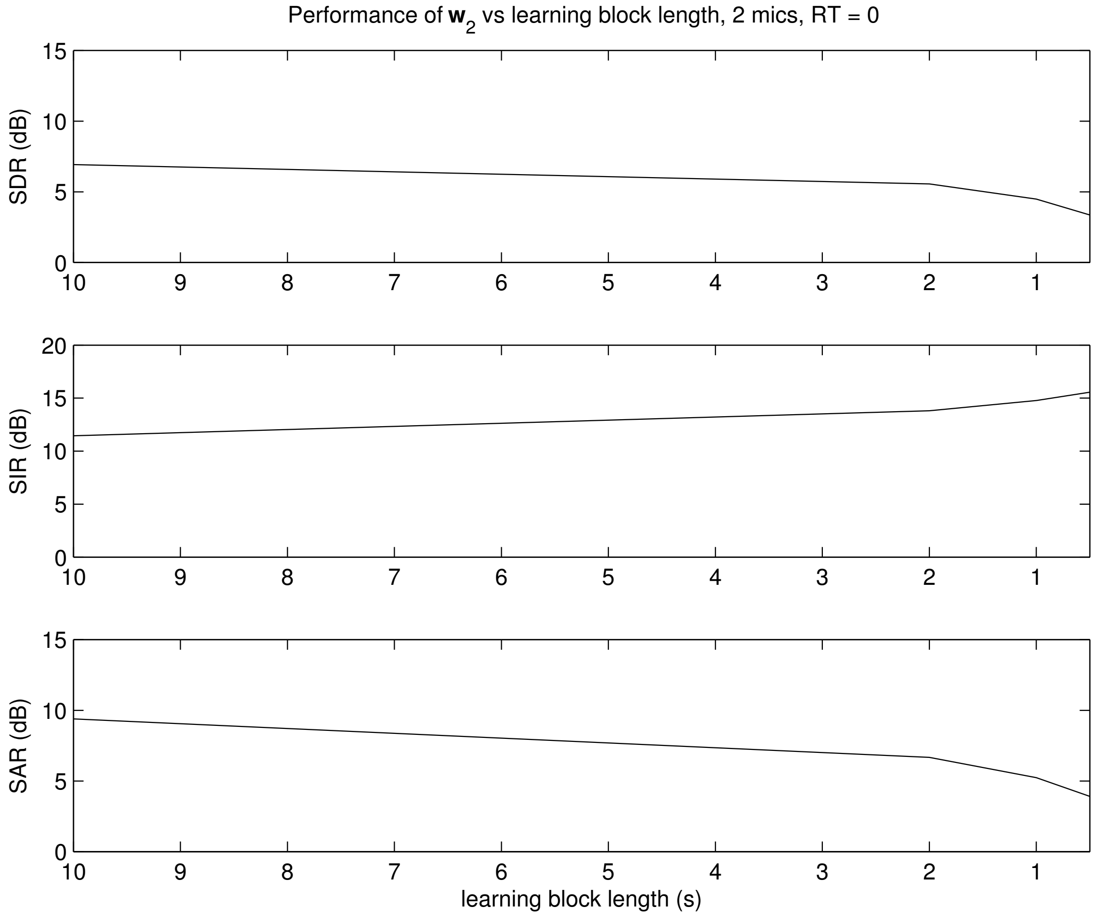

Figure 3.10: *Separation using two microphones: Average performance of* $\mathbf{w}_2$ *in the anechoic case vs learning block length in seconds. Used parameters:* $k_s = 2$ *and* $k_v = 15$.

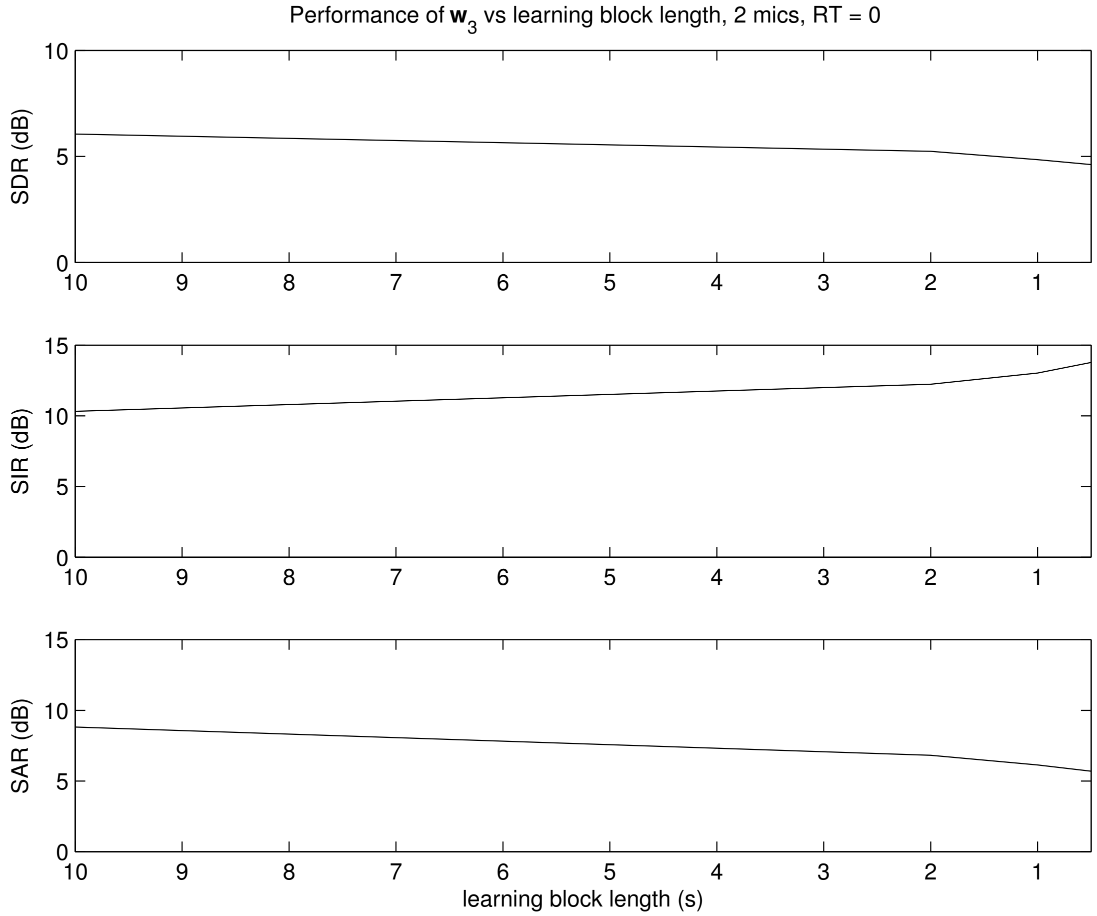

Figure 3.11: *Separation using two microphones: Average performance of* $\mathbf{w}_3$ *in the anechoic case vs learning block length in seconds. Used parameter:* $k_x = 15$.

3.5.4.3 Effect of the Learning Block Size

The EM algorithm used in our experiments is a batch learning algorithm. We studied the effect of varying the size of learning data on the performance of the non-linear beamformers. Figures 3.9, 3.10, and 3.11 show the average performance at the output of the non-linear beamformers in the anechoic case as a function of the EM learning block length. For the mixtures of MVDR and MVDR beamformers ($\mathbf{w}_2$ and $\mathbf{w}_3$), the performance is fairly consistent even when using short learning blocks of 0.5 seconds. However, the performance of $\mathbf{w}_1$ deteriorated when using learning blocks of 1 seconds or less. Note that the FMV algorithm can be considered as a special case of the non-linear beamformer $\mathbf{w}_3$, with $k_x = 1$ and very short learning blocks ($\approx$ 100 ms). The batch mode with short blocks of data can be used in applications where short delays are permissible, such as in human-computer interaction or surveillance. However, it is not appropriate for real-time applications. In these applications, online model learning is essential [114]. The online model learning should have a forgetting factor, and a mechanism for adding, deleting, and reassigning Gaussians to handle changes in the environment [115].

3.5.5 Directivity Patterns

In this section, we plot the directivity patterns for the non-linear beamformer $\mathbf{w}_3$ defined in (3.34) in the anechoic case. The microphone array used has two microphones with a 5 cm microphone spacing. The directivity patterns are defined as the magnitude of the response of the beamformer at frequency f for a far-field signal coming from direction Φ:

$$\mathrm{D}(f, \Phi) = \left| \sum_{j=1}^{N} w_j(f).e^{\iota 2\pi f (j-1) d c^{-1} \sin\Phi} \right| \tag{3.43}$$

Figure 3.12 shows four examples of directivity patterns for the non-linear beamformer $\mathbf{w}_3$. In this experiment, the desired source was at an angle of $10°$, and the interfering sources at $\{-45,\ -15,\ 50\}°$. The four examples are at four different time frames at the frequency of 453 Hz. In the first example (first row), the desired source and the interferer at angle $-45°$ were active. In the second example (second row), the interferer at angle $-15°$ was active. In the third example (third row), the desired source was active, and in the fourth example (fourth row), the interferer at angle $50°$ was active. The non-linear beamformer effectively reduces the contribution of the active interferer while having a distortionless response in the direction of

	time (s)	parameters
$\mathbf{w}_1$ (2 mics)	498	10 s block, $k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_2$ (2 mics)	498	10 s block, $k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_3$ (2 mics)	87	10 s block, $k_x = 15$, 50 iterations
$\mathbf{w}_1$ (3 mics)	193	10 s block, $k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_2$ (3 mics)	193	10 s block, $k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_3$ (3 mics)	14	10 s block, $k_x = 5$, 20 iterations
$\mathbf{w}_1$ (2 mics)	67	0.5 s block, $k_s = 2$, $k_v = 15$, 20 iterations
$\mathbf{w}_2$ (2 mics)	67	0.5 s block, $k_s = 2$, $k_v = 15$, 20 iterations
$\mathbf{w}_3$ (2 mics)	3	0.5 s block, $k_x = 15$, 3 iterations
$\mathbf{w}_1$ (3 mics)	17	0.5 s block, $k_s = 2$, $k_v = 5$, 20 iterations
$\mathbf{w}_2$ (3 mics)	17	0.5 s block, $k_s = 2$, $k_v = 5$, 20 iterations
$\mathbf{w}_3$ (3 mics)	1	0.5 s block, $k_x = 5$, 3 iterations

Table 3.2: *Computational time for* $\mathbf{w}_1$*,* $\mathbf{w}_2$*, and* $\mathbf{w}_3$.

the desired source.

Figure 3.13 shows examples of directivity patterns for the non-linear beamformer $\mathbf{w}_3$ when the active source is source 1. We show the directivity patterns which were designed to extract the four sources. We can see that when the active source is not the desired source, the non-linear beamformer attenuates the active source. Figures 3.14, 3.15, and 3.16 show more examples of the directivity patterns when the active source is source 2, source 3, and source 4 respectively . We can see that the non-linear beamformer attenuates the active source if it was not the desired source while always having a unity gain in the direction of the desired source.

Figure 3.17 shows directivity patterns when sources 2 and 3 are active at the same time-frequency point. We can see that when the desired source is either source 1 or source 4, the non-linear beamformer attempts to attenuate the two active sources using the single degree of freedom. when the desired source is source 2, the non-linear beamformer attempts to attenuate source 3, and similarly, when the desired source is source 3, the non-linear beamformer attempts to attenuate source 2.

3.5.6 Computational Time

In this section, we report the time it took for our MATLAB implementation of the non-linear beamformers to run on 2.5 GHz CPU. We report the time our implementation took for the extraction of one 10 s speech source, and one 0.5 s speech source. Table 3.2 shows the results.

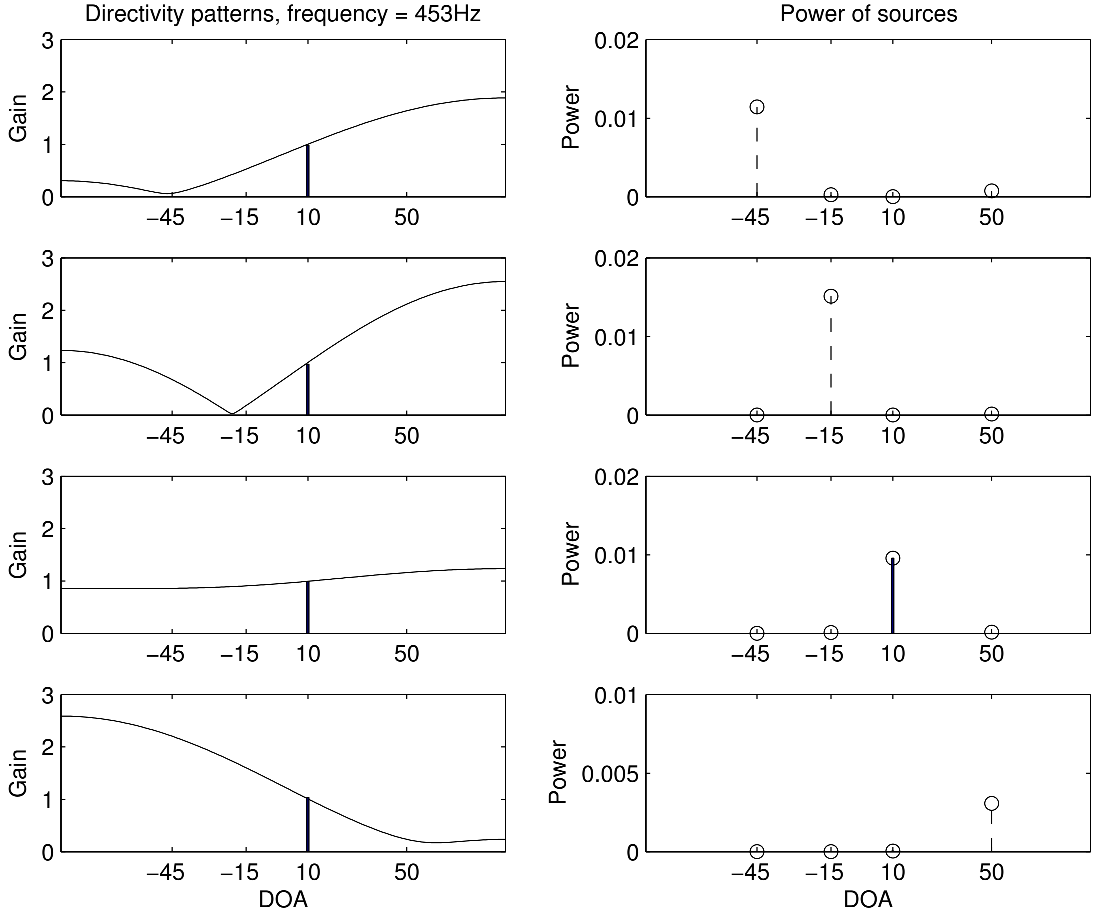

Figure 3.12: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$*. The desired source is at angle* 10°*, and the interfering sources are at* $\{-45,\ -15,\ 50\}^\circ$*. Left column: directivity patterns. Right column: power of sources.*

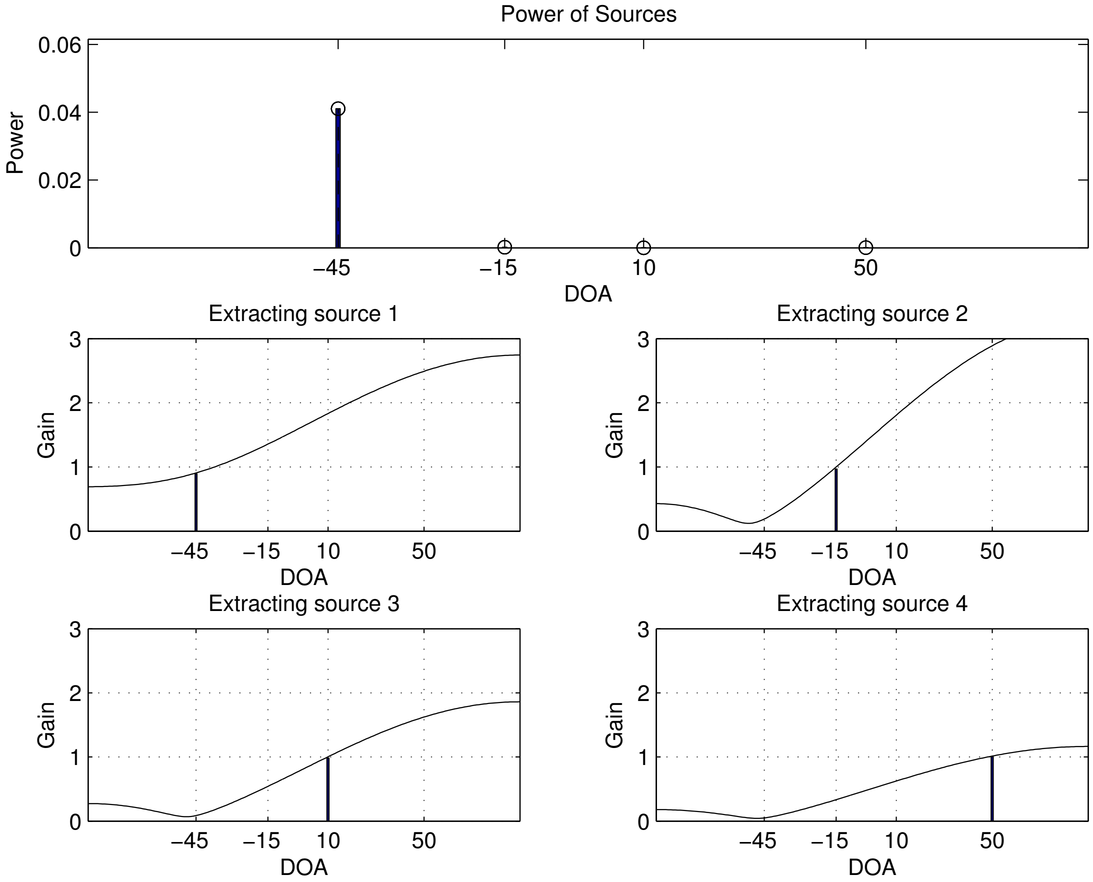

Figure 3.13: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$*. The active source is at angle* $-45°$.

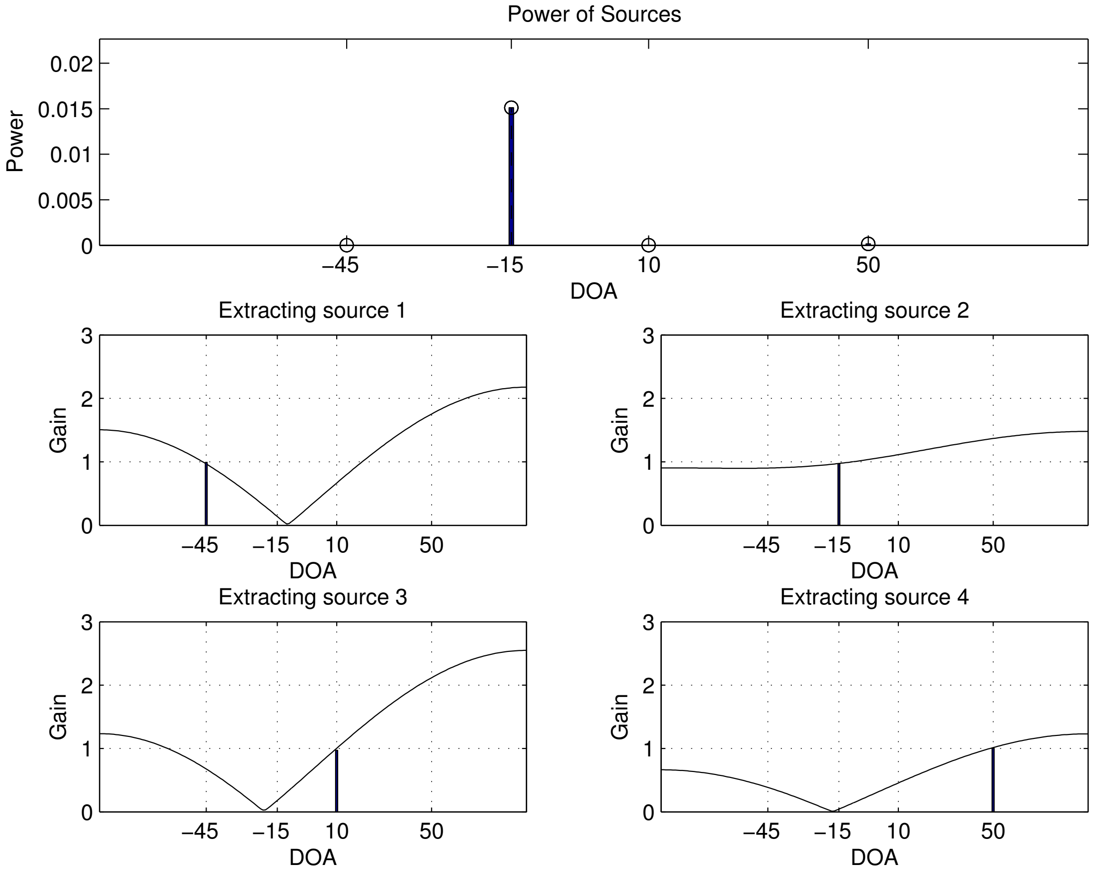

Figure 3.14: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$*. The active source is at angle* $-15°$.

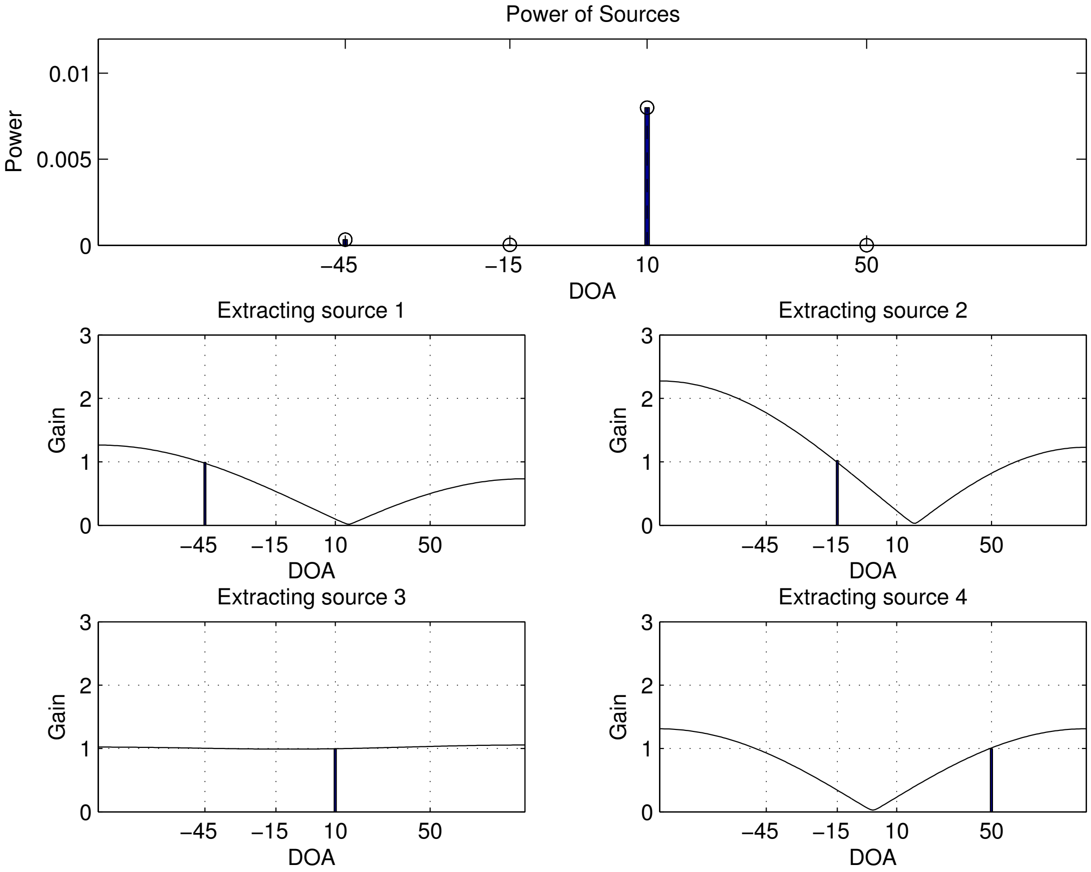

Figure 3.15: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$*. The active source is at angle* $10°$.

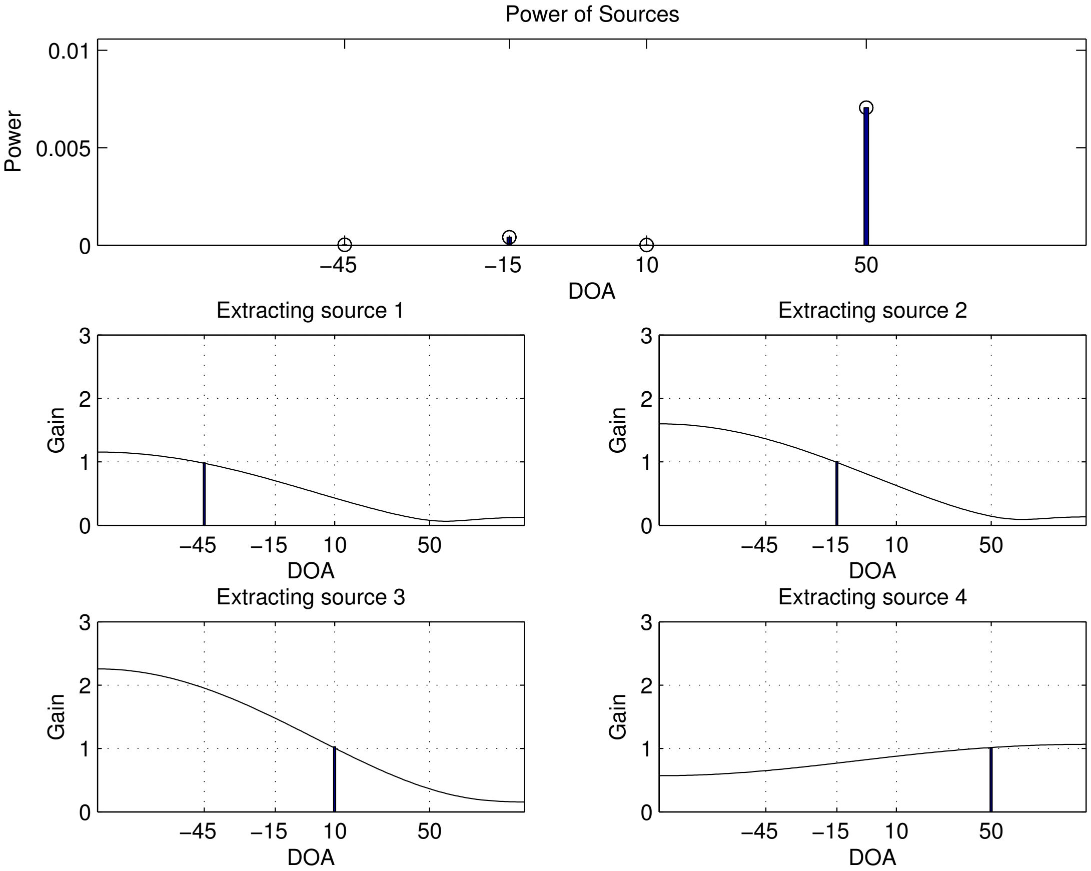

Figure 3.16: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$*. The active source is at angle* $50°$.

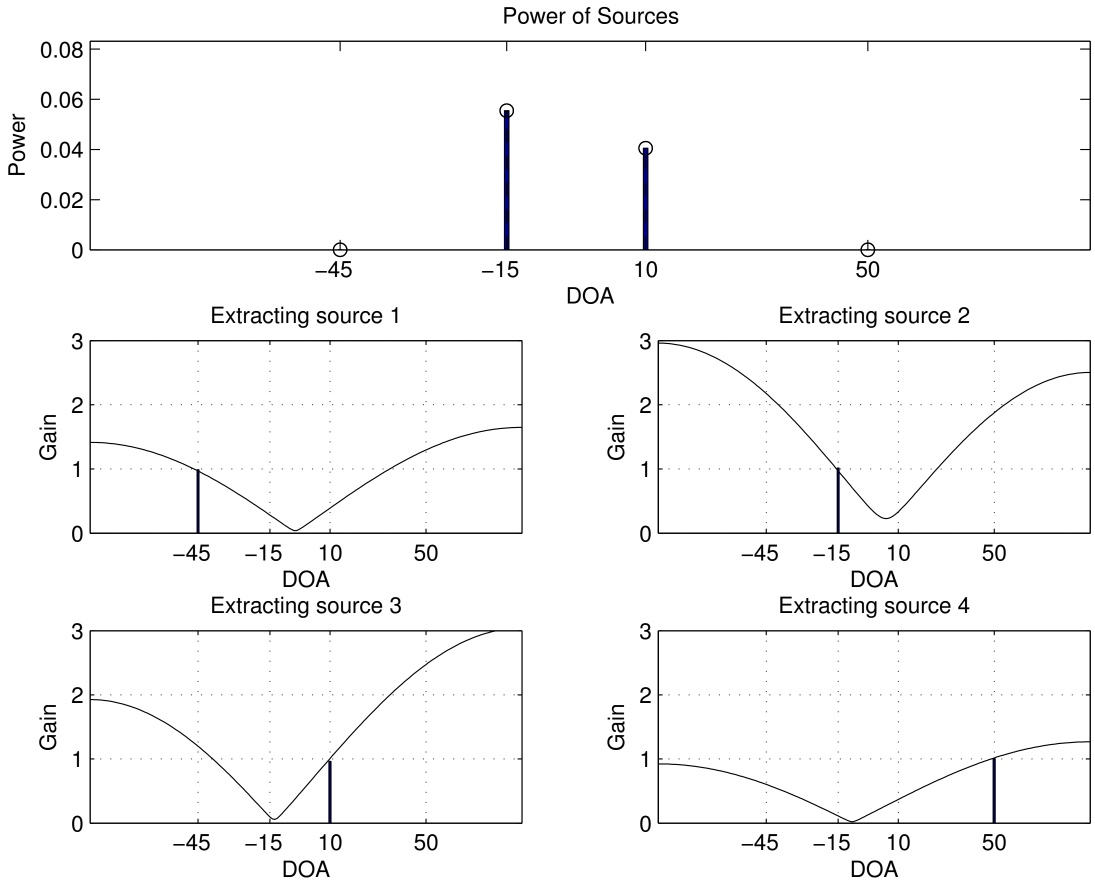

Figure 3.17: *Separation using two microphones: Examples of directivity patterns of* $\mathbf{w}_3$ *in the anechoic case at 453 Hz. Used parameter:* $k_x = 15$. *The sources at angles* $-15°$ *and* $10°$*are active.*

We can see that the beamformer $\mathbf{w}_3$ took less time than the other two beamformers. Beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ have similar computational time because they use the same learning algorithm (Section 3.4.2). The computational time for the three microphones case is lower than that of the two microphones case. This is because the performance of the beamformers using the three microphones array peak using a smaller number of Gaussian components than the two microphone case, and require less number of iterations.

3.6 Chapter Summary

Frequency-domain non-linear mixture of beamformers were introduced and applied to the extraction of a desired speech source from a known direction in underdetermined speech mixtures. The system model assumes an anechoic desired source signal, but no assumptions are made about the interferers, which can be of any nature such as point sources, spatial extended sources, diffuse sources, or a combination of them. The beamformers are derived assuming non-Gaussian interference signals modeled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers and dynamically finds suitable directivity patterns in order to reduce active interfering signals.

The non-linear beamformers require the location of the target speech source to be known or estimated in advance, but they have the following advantages:

- No need to know - or estimate - the number of interfering sources.
- Can be applied to underdetermined speech mixtures.
- The number of components in the GMM model controls the flexibility of the model, and can be used to trade-off complexity with performance, which can be good for hardware implementations with fixed computational constraints. When using a larger number of microphones, the performance peaks with a small number of GMM components.
- Can be applied to microphone arrays with two or more microphones.

The non-linear beamformers have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_1$ defined in (3.15) gives better interference rejection at the expense of higher artifacts. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ defined in (3.34) and (3.16) are distortionless beamformers (constant gain in the look-direction), and have significantly lower artifacts.

In terms of computational complexity, non-linear beamformer $\mathbf{w}_3$ employs the simplest learning algorithm and requires fewer iterations than non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$. Furthermore, the model learning for non-linear beamformer $\mathbf{w}_3$ is independent of the DOA of the desired source, which makes this non-linear beamformer suitable in applications where scanning for the source direction is needed.

Appendices

3.A Model Learning for $\mathbf{w}_1$ and $\mathbf{w}_2$

In this section, the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$ of the interference $\mathbf{v}$ and desired source s are estimated using the EM algorithm [30]. These parameters are required for the non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ defined in (3.15) and (3.16). Let us define a complete data set $D_c = \{\mathbf{x}, s, q_s, q_v\}$ composed of both the observed data $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$ and the latent data. If we were to actually have such a complete data set, we could define its log likelihood as:

$$l_c(\theta|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), s, q_s, q_v|\theta) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), s, q_s, q_v|\theta) \tag{3.44}$$

Given an initial value θ^0, the EM algorithm performs the following steps at each iteration l:

E-step: In the E-step, we compute the expectation of the complete data log likelihood:

$$Q(\theta, \theta^{l-1}) = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln p(\mathbf{x}, s, q_s, q_v|\theta) \tag{3.45}$$

The expectation is taken for each observed $\mathbf{x}$ with respect to the conditional probability of the hidden data, given the observed data and the parameter values obtained in the previous iteration. The result should then be averaged over all observed $\mathbf{x}$.

M-step: In the M-step, we maximise the expected complete log likelihood with respect to the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$.

Below we provide the derivation of the EM learning rules:

The Expectation Step:

To simplify $Q(\theta, \theta^{l-1})$, we first expand the probability of the joint event $p(\mathbf{x}, s, q_s, q_v)$. Applying the chain rule on $p(\mathbf{x}, s, \mathbf{v}, q_s, q_v)$, and marginalising $\mathbf{v}$, we can expand $p(\mathbf{x}, s, q_s, q_v)$ as follows:

$$\begin{aligned} p(\mathbf{x}, s, q_s, q_v) &= \int p(\mathbf{x}, s, \mathbf{v}, q_s, q_v) d\mathbf{v} \\ &= \int p(\mathbf{x}|s, \mathbf{v}).p_s(s|q_s).p_v(\mathbf{v}|q_v).p(q_s).p(q_v) d\mathbf{v} \qquad (3.46) \end{aligned}$$

Given both s and $\mathbf{v}$, $\mathbf{x}$ is deterministic ($\mathbf{a}$ is known). Therefore, $p(\mathbf{x}|s, \mathbf{v}) = \delta(\mathbf{x} - (\mathbf{a}s + \mathbf{v}))$, where δ is the Dirac's delta function. So using the identity: $\int f(\tau)\delta(t-T-\tau)d\tau = f(t-T))$, we get:

$$\begin{aligned} p(\mathbf{x}, s, q_s, q_v) &= \int \delta(\mathbf{x} - (\mathbf{a}s + \mathbf{v})).p_s(s|q_s).p_v(\mathbf{v}|q_v).p(q_s).p(q_v) d\mathbf{v} \\ &= p_s(s|q_s).p_v(\mathbf{x}|s, q_v).p(q_s).p(q_v) \qquad (3.47) \end{aligned}$$

where $p_v(\mathbf{x}|s, q_v)$=$G\left(\mathbf{a}s, \mathbf{R}_{v,q_v}\right)$. Substituting (3.47) in (3.45), we get:

$$\begin{aligned} Q(\theta, \theta^{l-1}) &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln (\\ & \quad p_v(\mathbf{x}|s, q_v, \mathbf{R}_v).p_s(s|q_s, \mathbf{R}_s^2).p(q_v).p(q_s)) \\ &= Q_A\left(\mathbf{R}_v|\theta^{l-1}\right) + Q_B\left(\sigma_s^2|\theta^{l-1}\right) + Q_C\left(\mathbf{c}_v|\theta^{l-1}\right) + Q_D\left(\mathbf{c}_s|\theta^{l-1}\right) \end{aligned} \qquad (3.48)$$

where $\sigma_s^2 = \left\{\sigma_{s,1}^2, \ldots, \sigma_{s,k_s}^2\right\}, \mathbf{R}_v = \{\mathbf{R}_{v,1}, \ldots, \mathbf{R}_{v,k_v}\}, \mathbf{c}_v = \{c_{v,1}, \ldots, c_{v,k_v}\}, \mathbf{c}_s = \{c_{s,1},$ $\ldots, c_{s,k_s}\}$ and:

$$Q_A\left(\mathbf{R}_v|\theta^{l-1}\right) = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln p_v(\mathbf{x}|s, q_v) \qquad (3.49)$$

$$Q_B\left(\sigma_s^2|\theta^{l-1}\right) = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln p_s(s|q_s) \qquad (3.50)$$

$$Q_C\left(\mathbf{c}_v|\theta^{l-1}\right) = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln p(q_v) \qquad (3.51)$$

$$Q_D\left(\mathbf{c}_s|\theta^{l-1}\right) = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) \ln p(q_s) \qquad (3.52)$$

Each of these terms involves only one of the parameters θ and may be maximised independently of each other. Next, we simplify the terms Q_A, Q_B, Q_C, and Q_D. Note that the terms Q_A, Q_B, Q_C, and Q_D are evaluated for each observed $\mathbf{x}$. The results should then be averaged over all observed $\mathbf{x}$

Evaluation of Q_A

$$\begin{aligned}
Q_A &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v}\int ds\, p\left(s, q_s, q_v|\mathbf{x}\right)\ln p_v(\mathbf{x}|s, q_v) \\
&= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v}\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right) p\left(q_s, q_v|\mathbf{x}\right)\ln p_v(\mathbf{x}|s, q_v) \\
&= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right)\ln p_v(\mathbf{x}|s, q_v) \qquad (3.53)
\end{aligned}$$

Substitute $p_v(\mathbf{x}|s, q_v) = G\left(\mathbf{a}s, \mathbf{R}_{v,q_v}\right) = \frac{1}{\pi^N \det \mathbf{R}_{v,q_v}}\exp\left(-\left(\mathbf{x}-\mathbf{a}s\right)^H \mathbf{R}_{v,q_v}^{-1}\left(\mathbf{x}-\mathbf{a}s\right)\right)$ in Q_A to get:

$$\begin{aligned}
Q_A = {} & \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right)\left(-\ln\det\mathbf{R}_{v,q_v} - \ln\pi^N - \mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x}\right. \\
& \left. + s^*\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x} + s\mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a} - s^*s\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\right) \qquad (3.54)
\end{aligned}$$

Define

$$\begin{aligned}
\langle s|\mathbf{x}, q_s, q_v\rangle &= \int ds\, p\left(s|q_s, q_v, \mathbf{x}\right) s \qquad (3.55) \\
\langle ss^*|\mathbf{x}, q_s, q_v\rangle &= \int ds\, p\left(s|q_s, q_v, \mathbf{x}\right) ss^* \qquad (3.56)
\end{aligned}$$

Substituting (3.55) and (3.56) in (3.54), we get:

$$\begin{aligned} Q_A &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\big(-\ln\det\mathbf{R}_{v,q_v} - \ln\pi^N - \mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x} \\ &\quad + \langle s^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x} + \langle s|\mathbf{x}, q_s, q_v\rangle\,\mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a} \\ &\quad - \langle ss^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\big) \end{aligned} \tag{3.57}$$

Using the trace property [116]:

$$\text{Tr[scalar]=scalar} \tag{3.58}$$

and the cyclic property of the trace [116]:

$$\text{Tr[ABC]=Tr[CAB]=Tr[BCA]} \tag{3.59}$$

we can write Q_A as follows:

$$\begin{aligned} Q_A &= -\sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\big(\ln\det\mathbf{R}_{v,q_v} + \ln\pi^N + \text{Tr}\left(\mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x}\right) \\ &\quad -\text{Tr}\left(\langle s^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{x}\right) - \text{Tr}\left(\langle s|\mathbf{x}, q_s, q_v\rangle\,\mathbf{x}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\right) \\ &\quad + \text{Tr}\left(\langle ss^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\right)\big) \\ &= -\sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\big(\ln\det\mathbf{R}_{v,q_v} + \ln\pi^N + \text{Tr}\left(\mathbf{R}_{v,q_v}^{-1}\mathbf{x}\mathbf{x}^H\right) \\ &\quad -\text{Tr}\left(\mathbf{R}_{v,q_v}^{-1}\mathbf{x}\,\langle s^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\right) \text{-Tr}\left(\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\,\langle s|\mathbf{x}, q_s, q_v\rangle\,\mathbf{x}^H\right) \\ &\quad + \text{Tr}\left(\mathbf{R}_{v,q_v}^{-1}\mathbf{a}\,\langle ss^*|\mathbf{x}, q_s, q_v\rangle\,\mathbf{a}^H\right)\big) \end{aligned} \tag{3.60}$$

The posterior probability of the GMM states $p\left(q_s, q_v|\mathbf{x}\right)$ can be evaluated as follows for each data vector $\mathbf{x}(n)$ given the current set of parameters $\theta^{(l-1)}$:

$$\begin{aligned}
\tau^{(l)}_{q_s,q_v}(n) &= p\left(q_s, q_v|\mathbf{x}(n), \theta^{(l-1)}\right) \\
&= \frac{p\left(q_s, q_v, \mathbf{x}(n)|\theta^{(l-1)}\right)}{p\left(\mathbf{x}(n)|\theta^{(l-1)}\right)} \\
&= \frac{p\left(q_s, q_v|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s, q_v, \theta^{(l-1)}\right)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p\left(q_s', q_v'|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s', q_v', \theta^{(l-1)}\right)} \qquad (3.61)
\end{aligned}$$

In order to evaluate (3.61), we need to evaluate $p(\mathbf{x}|q_s, q_v)$. This can be evaluated as follows:

$$\begin{aligned}
p(\mathbf{x}|q_s, q_v) &= \int p(\mathbf{x}, s|q_s, q_v) ds \\
&= \int p(\mathbf{x}|s, q_s, q_v).p(s|q_s, q_v) ds \\
&= \int p(\mathbf{x}|s, q_v).p(s|q_s) ds \\
&= \int G\left(\mathbf{a}s, \mathbf{R}_{v,q_v}\right).G\left(0, \sigma^2_{s,q_s}\right).ds \\
&= \int \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v}\right)}.\frac{1}{\pi\sigma^2_{s,q_s}}.\exp\left(-\left(\mathbf{x}-\mathbf{a}s\right)^H \mathbf{R}^{-1}_{v,q_v}\left(\mathbf{x}-\mathbf{a}s\right) - \frac{s.s^*}{\sigma^2_{s,q_s}}\right) ds \\
&= \int \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v}\right)}.\frac{1}{\pi\sigma^2_{s,q_s}}.\exp\Bigg(-\mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} + \mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}s + s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \\
&\qquad - s^*s\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} - \frac{ss^*}{\sigma^2_{s,q_s}}\Bigg) ds \\
&= \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v}\right)}.\frac{1}{\pi\sigma^2_{s,q_s}}.\exp\left(-\mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right).\int \exp\Bigg(2\Re\left(\mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}s\right) \\
&\qquad - s^*s\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right) ds\Bigg) \qquad (3.62)
\end{aligned}$$

To evaluate the integral, we use this identity:

$$\int \exp\left(-y^*yb + 2\Re\left(c^*y\right)\right) dy = b^{-1}\pi \exp\left(c^*\left(b^{-1}\right)^* c\right) \qquad (3.63)$$

So we get:

$$\begin{aligned} p(\mathbf{x}|q_s,q_v) &= \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v}\right)} . \frac{1}{\pi\sigma^2_{s,q_s}} . \frac{\pi}{\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right)} \exp\left(-\mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right. \\ &\quad \left. + \mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right) \\ &= \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v}\right)\left(\sigma^2_{s,q_s}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}+1\right)} \exp\left(-\mathbf{x}^H\left(\mathbf{R}^{-1}_{v,q_v}\right.\right. \\ &\quad \left.\left. -\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\right)\mathbf{x}\right) \end{aligned} \tag{3.64}$$

Using $\det(X)\left(1 + b^H X^{-1} a\right) = \det\left(X + ab^H\right)$ [117], we can simplify the term before the exponential and get:

$$\begin{aligned} p(\mathbf{x}|q_s,q_v) &= \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v} + \mathbf{a}\sigma^2_{s,q_s}\mathbf{a}^H\right)} \exp\left(\right. \\ &\quad \left. -\mathbf{x}^H\left(\mathbf{R}^{-1}_{v,q_v} - \mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\right)\mathbf{x}\right) \end{aligned} \tag{3.65}$$

Using the matrix inversion lemma [118],

$$B - BC\left(C^H BC + D\right)^{-1} C^H B = \left(B^{-1} + CD^{-1}C^H\right)^{-1}, \tag{3.66}$$

we can simplify

$$\mathbf{R}^{-1}_{v,q_v} - \mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \frac{1}{\sigma^2_{s,q_s}}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v} = \left(\mathbf{R}_{v,q_v} + \mathbf{a}\sigma^2_{s,q_s}\mathbf{a}^H\right)^{-1} \tag{3.67}$$

and get

$$\begin{aligned} p(\mathbf{x}|q_s,q_v) &= \frac{1}{\pi^N \det\left(\mathbf{R}_{v,q_v} + \mathbf{a}\sigma^2_{s,q_s}\mathbf{a}^H\right)} \exp\left(-\mathbf{x}^H\left(\mathbf{R}_{v,q_v} + \mathbf{a}\sigma^2_{s,q_s}\mathbf{a}^H\right)^{-1}\mathbf{x}\right) \\ &= G\left(0\,,\ \mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right) \end{aligned} \tag{3.68}$$

To complete the evaluation of Q_A we must express the conditional averages in (3.55) and (3.56) in terms of the parameters $\theta^{(l-1)}$. We first evaluate $p(s|\mathbf{x}, q_s, q_v)$:

$$\begin{aligned}
p(s|\mathbf{x}, q_s, q_v) &= \frac{p(\mathbf{x}, s, q_s, q_v)}{p(\mathbf{x}, q_s, q_v)} \\
&= \frac{p_s(s|q_s).p_v(\mathbf{x}|s, q_v).p(q_s).p(q_v)}{p(\mathbf{x}|q_s, q_v).p(q_s).p(q_v)} \\
&= \frac{G(0, \sigma^2_{s,q_s}).G(\mathbf{a}s, \mathbf{R}_{v,q_v})}{G\left(0, \mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)} \\
&= \frac{\det\left(\mathbf{R}_{v,q_v} + \mathbf{a}\sigma^2_{s,q_s}\mathbf{a}^H\right)}{\pi\sigma^2_{s,q_s}\det\mathbf{R}_{v,q_v}} \exp\Big(-s^* s\sigma^{-2}_{s,q_s} - \mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \\
&\quad + 2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right) - s^* s\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \mathbf{x}^H\left(\mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)^{-1}\mathbf{x}\Big)
\end{aligned} \tag{3.69}$$

Using $\det(X)\left(1 + b^H X^{-1} a\right) = \det\left(X + ab^H\right)$ [117], we can simplify the term before the exponential and get:

$$\begin{aligned}
p(s|\mathbf{x}, q_s, q_v) &= \frac{\det\left(\mathbf{R}_{v,q_v}\right)\left(1 + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\sigma^2_{s,q_s}\right)}{\pi\sigma^2_{s,q_s}\det\mathbf{R}_{v,q_v}} \exp\Big(-s^* s\sigma^{-2}_{s,q_s} - \mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \\
&\quad + 2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right) - s^* s\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \mathbf{x}^H\left(\mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)^{-1}\mathbf{x}\Big) \\
&= \frac{1}{\pi\left(\sigma^{-2}_{s,q_s} + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}} \exp\Big(-s^* s\sigma^{-2}_{s,q_s} - \mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \\
&\quad + 2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right) - s^* s\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a} + \mathbf{x}^H\left(\mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)^{-1}\mathbf{x}\Big)
\end{aligned} \tag{3.70}$$

Rearranging the terms inside the exponential, we get:

$$\begin{aligned}
p(s|\mathbf{x}, q_s, q_v) &= \frac{1}{\pi\left(\sigma^{-2}_{s,q_s} + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}} \exp\Big(-s^* s\left(\sigma^{-2}_{s,q_s} + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right) \\
&\quad + 2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right) - \mathbf{x}^H\Big(\mathbf{R}^{-1}_{v,q_v} - \left(\mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)^{-1}\Big)\mathbf{x}\Big)
\end{aligned} \tag{3.71}$$

If we expand $\left(\mathbf{R}_{v,q_v}+\sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)^{-1}$using the matrix inversion lemma, we get:

$$\begin{aligned}
p(s|\mathbf{x},q_s,q_v) &= \frac{1}{\pi\left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}}\exp\Big(-s^*s\left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)\\
&\quad +2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right)\\
&\quad -\mathbf{x}^H\Big(\mathbf{R}^{-1}_{v,q_v}-\mathbf{R}^{-1}_{v,q_v}+\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}+\sigma^{-2}_{s,q_s}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\Big)\mathbf{x}\Big)\\
&= \frac{1}{\pi\left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}}\exp\Big(-s^*s\left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)\\
&\quad +2\Re\left(s^*\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\right)-\mathbf{x}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\left(\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}+\sigma^{-2}_{s,q_s}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\Big)
\end{aligned} \tag{3.72}$$

We can see that $p(s|\mathbf{x},q_s,q_v)$ is Gaussian with the following parameters:

$$\text{variance} = \left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1} \tag{3.73}$$

$$\text{mean} = \left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \tag{3.74}$$

so

$$p(s|\mathbf{x},q_s,q_v) = G\left(\left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x}\,,\ \left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}\right) \tag{3.75}$$

We can now evaluate the conditional moments $\langle s|\mathbf{x}(n),q_s,q_v\rangle$ and $\langle ss^*|\mathbf{x}(n),q_s,q_v\rangle$:

$$\langle s|\mathbf{x},q_s,q_v\rangle = \left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \tag{3.76}$$

$$\langle ss^*|\mathbf{x},q_s,q_v\rangle = \left(\sigma^{-2}_{s,q_s}+\mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1}+\langle s|\mathbf{x},q_s,q_v\rangle\langle s^*|\mathbf{x},q_s,q_v\rangle \tag{3.77}$$

Evaluation of Q_B

$$\begin{aligned} Q_B &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v}\int ds\, p\left(s, q_s, q_v|\mathbf{x}\right)\ln p_s(s|q_s) \\ &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v}\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right) p\left(q_s, q_v|\mathbf{x}\right)\ln p_s(s|q_s) \\ &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right)\ln p_s(s|q_s) \end{aligned} \tag{3.78}$$

Substituting $p_s(s|q_s) = G\left(0, \sigma^2_{s,q_s}\right) = \frac{1}{\pi\sigma^2_{s,q_s}}\text{exp}\left(\frac{-|s|^2}{\sigma^2_{s,q_s}}\right)$ in Q_B , we get:

$$\begin{aligned} Q_B &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\int ds\, p\left(s|q_s, q_v, \mathbf{x}\right)\left(-\ln\pi - \ln\sigma^2_{s,q_s} - \frac{s^*s}{\sigma^2_{s,q_s}}\right) \\ &= -\sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\left(+\ln\pi + \ln\sigma^2_{s,q_s} + \frac{\langle ss^*|\mathbf{x}, q_s, q_v\rangle}{\sigma^2_{s,q_s}}\right) \end{aligned} \tag{3.79}$$

Evaluation of Q_C

$$\begin{aligned} Q_C &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v}\int ds\, p\left(s, q_s, q_v|\mathbf{x}\right)\ln p(q_v) \\ &= \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\ln p(q_v) \end{aligned} \tag{3.80}$$

Substituting $p_s(q_v) = c_{v,q_v}$ in Q_C, we get:

$$Q_C = \sum_{q_s=1}^{k_s}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\ln c_{v,q_v} \tag{3.81}$$

Evaluation of Q_D

If we follow the same steps as in the evaluation of Q_C, we get:

$$Q_D = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) \ln c_{s,q_s} \tag{3.82}$$

Summary of the Expectation Step Learning Rules:

We can see that to evaluate $Q(\theta, \theta^{l-1})$, we have to compute the posterior probability of the GMM states, and the first and second moments of the desired source given both the observed mixture and the GMM states. The posterior probability of the GMM states can be evaluated as follows:

$$\begin{aligned} \tau_{q_s,q_v}^{(l)}(n) &= p\left(q_s, q_v|\mathbf{x}(n), \theta^{(l-1)}\right) \\ &= \frac{p\left(q_s, q_v|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s, q_v, \theta^{(l-1)}\right)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p\left(q_s', q_v'|\theta^{(l-1)}\right) p\left(\mathbf{x}(n)|q_s', q_v', \theta^{(l-1)}\right)} \end{aligned} \tag{3.83}$$

where $p(\mathbf{x}|q_s, q_v)$ is equal to:

$$p(\mathbf{x}|q_s, q_v) = G\left(0\,,\ \mathbf{R}_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{a}\mathbf{a}^H\right) \tag{3.84}$$

The first and second moments of the desired source given both the observed mixture and the GMM states, which are denoted by $\langle s|\mathbf{x}(n), q_s, q_v\rangle$ and $\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle$ respectively, can be evaluated as follows:

$$\langle s|\mathbf{x}, q_s, q_v\rangle = \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}\right)^{-1} \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{x} \tag{3.85}$$

$$\langle ss^*|\mathbf{x}, q_s, q_v\rangle = \left(\sigma_{s,q_s}^{-2} + \mathbf{a}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{a}\right)^{-1} + \langle s|\mathbf{x}, q_s, q_v\rangle \langle s^*|\mathbf{x}, q_s, q_v\rangle \tag{3.86}$$

The Maximisation Step:

In the M-step, we maximise the expected complete log likelihood with respect to the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma_{s,q_s}^2, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. This can be done by taking derivatives with respect to θ and setting them to be equal to zero (under the constraints

$\sum_{q_s=1}^{k_s} c_{s,q_s} = 1$ and $\sum_{q_v=1}^{k_v} c_{v,q_v} = 1$).

Maximisation of Q_A

The term Q_A is a function of the interference model covariances $\mathbf{R}_{v,q_v}$. Using the following two identities [117, 119]:

$$\frac{\partial \ln \det (X)}{\partial X} = \left(X^{-1}\right)^H = \left(X^{-H}\right)^{-1} \tag{3.87}$$

$$\frac{\partial \text{Tr}\left(AX^{-1}B\right)}{\partial X} = -\left(X^{-1}BAX^{-1}\right)^H = -X^{-H}A^H B^H X^{-H} \tag{3.88}$$

Taking derivatives with respect to $\mathbf{R}_{v,q_v}$, we get:

$$\begin{aligned}\frac{\partial Q_A}{\partial \mathbf{R}_{v,q_v}} = & -\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\left(\mathbf{R}_{v,q_v}^{-1} - \mathbf{R}_{v,q_v}^{-1}\left(\mathbf{x}\mathbf{x}^H - \mathbf{a}\left\langle s|\mathbf{x}, q_s, q_v\right\rangle \mathbf{x}^H\right.\right. \\ & \left.\left. -\mathbf{x}\left\langle s^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H + \mathbf{a}\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H\right)\mathbf{R}_{v,q_v}^{-1}\right)\end{aligned} \tag{3.89}$$

Equating the derivative to zero, we get:

$$\begin{aligned}\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\mathbf{R}_{v,q_v}^{-1} = & \sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\mathbf{R}_{v,q_v}^{-1}\left(\mathbf{x}\mathbf{x}^H - \mathbf{a}\left\langle s|\mathbf{x}, q_s, q_v\right\rangle \mathbf{x}^H\right. \\ & \left. -\mathbf{x}\left\langle s^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H + \mathbf{a}\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H\right)\mathbf{R}_{v,q_v}^{-1}\end{aligned} \tag{3.90}$$

$$\begin{aligned}\mathbf{R}_{v,q_v}^{-1}\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right) = & \mathbf{R}_{v,q_v}^{-1}\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\left(\mathbf{x}\mathbf{x}^H - \mathbf{a}\left\langle s|\mathbf{x}, q_s, q_v\right\rangle \mathbf{x}^H\right. \\ & \left. -\mathbf{x}\left\langle s^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H + \mathbf{a}\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle \mathbf{a}^H\right)\mathbf{R}_{v,q_v}^{-1}\end{aligned} \tag{3.91}$$

$$\mathbf{R}_{v,q_v} = \frac{\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\left(\Lambda_{q_s,q_v}(n)\right)}{\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)} \tag{3.92}$$

where

$$\begin{aligned}\Lambda_{q_s,q_v}(n) \quad &= \quad \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n)\left\langle s^*|\mathbf{x}(n), q_s, q_v\right\rangle \mathbf{a}^H \\ & \quad\quad -\mathbf{a}\left\langle s|\mathbf{x}(n), q_s, q_v\right\rangle \mathbf{x}(n)^H \\ & \quad\quad +\mathbf{a}\left\langle ss^*|\mathbf{x}(n), q_s, q_v\right\rangle \mathbf{a}^H \end{aligned} \tag{3.93}$$

Averaging over the observed signal $\mathbf{x}$:

$$\mathbf{R}_{v,q_v} = \frac{\frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)\left(\Lambda_{q_s,q_v}(n)\right)}{\frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right)} \tag{3.94}$$

Maximisation of Q_B

The term Q_B is a function of the source model variances σ^2_{s,q_s}. Taking derivatives with respect to σ^2_{s,q_s}, we get:

$$\frac{\partial Q_B}{\partial \sigma^2_{s,q_s}} = -\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\left(\frac{1}{\sigma^2_{s,q_s}} - \frac{\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle}{\sigma^4_{s,q_s}}\right) \tag{3.95}$$

equating the derivative to zero, we get:

$$\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) = \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\frac{\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle}{\sigma^2_{s,q_s}} \tag{3.96}$$

$$\sigma^2_{s,q_s} = \frac{\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle}{\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)} \tag{3.97}$$

Averaging over the observed signal $\mathbf{x}$:

$$\sigma^2_{s,q_s} = \frac{\frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)\left\langle ss^*|\mathbf{x}, q_s, q_v\right\rangle}{\frac{1}{\eta}\sum_{n=1}^{\eta}\sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right)} \tag{3.98}$$

Maximisation of Q_C

The term Q_C is a function of the interference model mixing probabilities c_{v,q_v}. We should ensure that the probabilities c_{v,q_v} satisfy the non-negativity $c_{v,q_v} \geq 0$ and normalisation

$\sum_{q_v=1}^{k_v} c_{v,q_v} = 1$ constraints. Both constraints can be enforced automatically by working with new parameters $\overline{c_{v,q_v}}$, related to the mixing proportions through [88]:

$$c_{v,q_v} = \frac{\exp\left(\overline{c_{v,q_v}}\right)}{\sum_{q_v^{'}=1}^{k_v} \exp\left(\overline{c_{v,q_v}}\right)} \tag{3.99}$$

The gradient is then taken with respect to the new parameters:

$$\begin{aligned}
\frac{\partial Q_C}{\partial \overline{c_{v,q_v}}} &= \frac{\partial}{\partial \overline{c_{v,q_v}}} \left(\sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) \ln \exp\left(\overline{c_{v,q_v}}\right) \right. \\
&\qquad \left. - \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) \ln \sum_{q_v^{'}=1}^{k_v} \exp\left(\overline{c_{v,q_v}}\right) \right) \\
&= \frac{\partial}{\partial \overline{c_{v,q_v}}} \left(\sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) \left(\overline{c_{v,q_v}}\right) \right) \\
&\qquad - \frac{\partial}{\partial \overline{c_{v,q_v}}} \left(\ln \sum_{q_v^{'}=1}^{k_v} \exp\left(\overline{c_{v,q_v}}\right) \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p\left(q_s, q_v|\mathbf{x}\right) \right) \\
&= \sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right) - \frac{\partial}{\partial \overline{c_{v,q_v}}} \left(\ln \sum_{q_v^{'}=1}^{k_v} \exp\left(\overline{c_{v,q_v}}\right) \right)
\end{aligned} \tag{3.100}$$

Using the identity [117]:

$$\frac{\partial \ln f\left(X\right)}{\partial X} = \frac{\frac{\partial f(X)}{\partial X}}{f\left(X\right)} \tag{3.101}$$

we get:

$$\begin{aligned}
\frac{\partial Q_C}{\partial \overline{c_{v,q_v}}} &= \sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right) - \frac{\exp\left(\overline{c_{v,q_v}}\right)}{\sum_{q_v^{'}=1}^{k_v} \exp\left(\overline{c_{v,q_v}}\right)} \\
&= \sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right) - c_{v,q_v}
\end{aligned} \tag{3.102}$$

Equate to zero:

$$c_{v,q_v} = \sum_{q_s=1}^{k_s} p\left(q_s, q_v|\mathbf{x}\right) \tag{3.103}$$

Averaging over the observed signal $\mathbf{x}$:

$$c_{v,q_v} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p\left(q_s, q_v | \mathbf{x}\right) \tag{3.104}$$

Maximisation of Q_D

The term Q_D is a function of the source model mixing probabilities c_{s,q_s} if we follow the same steps as in the maximisation of Q_C, we get:

$$c_{s,q_s} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p\left(q_s, q_v | \mathbf{x}\right) \tag{3.105}$$

Summary of the Maximisation Step Learning Rules:

The maximisation of

$$Q(\theta, \theta^{l-1}) = Q_A\left(\mathbf{R}_v | \theta^{l-1}\right) + Q_B\left(\sigma_s^2 | \theta^{l-1}\right) + Q_C\left(\mathbf{c}_v | \theta^{l-1}\right) + Q_D\left(\mathbf{c}_s | \theta^{l-1}\right) \tag{3.106}$$

results in the following update rules:

$$c_{v,q_v}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n) \tag{3.107}$$

$$c_{s,q_s}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n) \tag{3.108}$$

$$\sigma_{s,q_s}^{2(l)} = \frac{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n)\ \langle ss^* | \mathbf{x}(n), q_s, q_v \rangle}{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n)} \tag{3.109}$$

$$\mathbf{R}_{v,q_v}^{(l)} = \frac{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n)\, \Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n)} \tag{3.110}$$

where

$$\begin{aligned}\Lambda_{q_s,q_v}(n) &= \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n)\left\langle s^*|\mathbf{x}(n),q_s,q_v\right\rangle \mathbf{a}^H \\ &\quad -\mathbf{a}\left\langle s|\mathbf{x}(n),q_s,q_v\right\rangle \mathbf{x}(n)^H \\ &\quad +\mathbf{a}\left\langle ss^*|\mathbf{x}(n),q_s,q_v\right\rangle \mathbf{a}^H \end{aligned} \tag{3.111}$$

3.B Model Learning for $\mathbf{w}_3$

In this section, the parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$ of the observed mixture $\mathbf{x}$ are estimated using the EM algorithm. These parameters are required for the non-linear beamformer $\mathbf{w}_3$ defined in (3.34). Let us define a complete data set $D_c = \{\mathbf{x}, q_x\}$ composed of both the observed data $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$ and the latent data. If we were to actually have such a complete data set, we define its log likelihood as:

$$l_c(\theta_x|D_c) = \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), q_x|\theta_x) = \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), q_x|\theta_x) \tag{3.112}$$

The EM algorithm may be executed as follows:

E-step: In the E-step, we compute the expectation of the complete data log likelihood, where, for each observed data vector, the expectation is taken over the latent data using the posterior $p\left(q_x|\mathbf{x}(n), \theta_x^{(l-1)}\right)$; $\theta_x^{(l-1)}$ are the parameters obtained in the previous iteration.

$$\begin{aligned} Q(\theta_x, \theta_x^{(l-1)}) &= \mathrm{E}\left[\ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), q_x|\theta_x)|\mathbf{x}(n)\right] \\ &= \mathrm{E}\left[\sum_{n=1}^{\eta} \ln\, p(\mathbf{x}(n), q_x|\theta_x)|\mathbf{x}(n)\right] \\ &= \sum_{n=1}^{\eta} \mathrm{E}\left[\ln\, p(\mathbf{x}(n), q_x|\theta_x)|\mathbf{x}(n)\right] \\ &= \sum_{n=1}^{\eta}\sum_{q_x=1}^{k_x} p\left(q_x|\mathbf{x}(n), \theta_x^{(l-1)}\right) \ln\, p\left(\mathbf{x}(n), q_x|\theta_x\right) \end{aligned} \tag{3.113}$$

This reduces to calculating $p\left(q_x|\mathbf{x}(n),\theta_x^{(l-1)}\right)$, the posterior probability of the latent variables given the observed data and the current estimates of the parameters.

$$\begin{aligned}
\tau_{q_x}^{(l)}(n) &= p\left(q_x|\mathbf{x}(n),\theta_x^{(l-1)}\right) \\
&= \frac{p\left(q_x,\mathbf{x}(n)|\theta_x^{(l-1)}\right)}{p\left(\mathbf{x}(n)|\theta_x^{(l-1)}\right)} \\
&= \frac{p\left(q_x|\theta_x^{(l-1)}\right)\,p\left(\mathbf{x}(n)|q_x,\theta_x^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} p\left(q_x'|\theta_x^{(l-1)}\right)\,p\left(\mathbf{x}(n)|q_x',\theta_x^{(l-1)}\right)} \\
&= \frac{c_{x,q_x}^{(l-1)}\,G\left(0,\mathbf{R}_{x,q_x}^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} c_{x,q_x'}^{(l-1)}\,G\left(0,\mathbf{R}_{x,q_x'}^{(l-1)}\right)} \qquad (3.114)
\end{aligned}$$

M-step: In the M-step, we maximise the expected complete log likelihood with respect to the parameters $\theta_x = \{c_{x,q_x},\mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$. This can be done by taking derivatives with respect to θ_x and setting them to be equal to zero, while also including a Lagrangian term to account for the constraint that $\sum_{q_x=1}^{k_x} c_{x,q_x} = 1$. First, we expand the probability of the joint event in $Q(\theta_x,\theta_x^{l-1})$:

$$\begin{aligned}
Q(\theta_x,\theta_x^{(l-1)}) &= \sum_{n=1}^{\eta}\sum_{q_x=1}^{k_x} p\left(q_x|\mathbf{x}(n),\theta_x^{(l-1)}\right)\,\ln p\left(\mathbf{x}(n),q_x|\theta_x\right) \\
&= \sum_{n=1}^{\eta}\sum_{q_x=1}^{k_x} p\left(q_x|\mathbf{x}(n),\theta_x^{(l-1)}\right)\,\ln\left(p\left(q_x|\theta_x\right)p\left(\mathbf{x}(n)|q_x,\theta_x\right)\right) \\
&= \sum_{n=1}^{\eta}\sum_{q_x=1}^{k_x} \tau_{q_x}^{(l)}(n)\,\ln\left(c_{x,q_x}p\left(\mathbf{x}(n)|q_x,\theta_x\right)\right) \qquad (3.115)
\end{aligned}$$

If we add a Lagrange multiplier, and expand the density, we get:

$$\begin{aligned}
\mathcal{L}(\theta_x) &= \sum_{n=1}^{\eta}\sum_{q_x=1}^{k_x}\tau_{q_x}^{(l)}(n)\Bigg(\ln c_{x,q_x} - N\ln\pi - \ln\det\mathbf{R}_{x,q_x} - \mathbf{x}(n)^H\mathbf{R}_{x,q_x}^{-1}\mathbf{x}(n)\Bigg) \\
&\quad -\lambda\Bigg(\sum_{q_x=1}^{k_x} c_{x,q_x} - 1\Bigg) \qquad (3.116)
\end{aligned}$$

We find the new estimate $\theta_x^{(l)} = \{c_{x,q_x}^{(l)}, \mathbf{R}_{x,q_x}^{(l)} : 1 \leq q_x \leq k_x)$ at a maximum where $\frac{\delta \mathcal{L}(\theta_x)}{\delta \theta_x} = 0$.

A new estimate of the covariance matrix $\mathbf{R}_{x,q_x}$:

$$\begin{aligned}
\frac{\delta \mathcal{L}(\theta_x)}{\delta \mathbf{R}_{x,q_x}} &= \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \left(-\frac{\delta}{\delta \mathbf{R}_{x,q_x}} \ln \det \mathbf{R}_{x,q_x} - \frac{\delta}{\delta \mathbf{R}_{x,q_x}} \mathbf{x}(n)^H \mathbf{R}_{x,q_x}^{-1} \mathbf{x}(n) \right) \\
&= \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \left(-\mathbf{R}_{x,q_x}^{-H} + \mathbf{R}_{x,q_x}^{-H} \mathbf{x}(n) \mathbf{x}(n)^H \mathbf{R}_{x,q_x}^{-H} \right) \\
&= 0
\end{aligned} \tag{3.117}$$

So, we have:

$$\begin{aligned}
\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \mathbf{R}_{x,q_x}^{-1} &= \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \mathbf{R}_{x,q_x}^{-1} \mathbf{x}(n) \mathbf{x}(n)^H \mathbf{R}_{x,q_x}^{-1} \\
\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) &= \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \mathbf{R}_{x,q_x}^{-1} \mathbf{x}(n) \mathbf{x}(n)^H
\end{aligned}$$

$$\mathbf{R}_{x,q_x}^{(l)} = \frac{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)\, \mathbf{x}(n)\, \mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)} \tag{3.118}$$

A new estimate of the mixture components weights $c_{x,q_x} = p\left(q_x | \theta_x\right)$:

$$\begin{aligned}
\frac{\delta \mathcal{L}(\theta_x)}{\delta c_{x,q_x}} &= \left(\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \frac{\delta \ln c_{x,q_x}}{\delta c_{x,q_x}} \right) - \lambda \left(\frac{\delta c_{x,q_x}}{\delta c_{x,q_x}} \right) \\
&= \left(\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \frac{1}{c_{x,q_x}} \right) - \lambda
\end{aligned} \tag{3.119}$$

So, we have:

$$c_{x,q_x} = \frac{1}{\lambda} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \tag{3.120}$$

Substituting this into the constraint:

$$\sum_{q_x=1}^{k_x} c_{x,q_x} = \sum_{q_x=1}^{k_x} \frac{1}{\lambda} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) = 1 \tag{3.121}$$

$$\lambda = \sum_{q_x=1}^{k_x} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \tag{3.122}$$

Inserting λ into our estimate:

$$\begin{aligned} c_{x,q_x}^{(l)} &= \frac{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)}{\sum_{q_x=1}^{k_x} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)} \\ &= \frac{1}{\eta} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \end{aligned} \tag{3.123}$$

Chapter 4
Combining Time-Frequency Masking and Mixtures of Beamformers

4.1 Introduction

In this chapter, we present a modification to the mixture of MPDR beamformers presented in the previous chapter. In the previous chapter, we used the EM algorithm to learn a GMM in each frequency band. The EM algorithm is computationally expensive, and the model learning depends on the selection of appropriate initial values for the covariance matrices (local maxima problems). In this chapter, we investigate the use of other clustering algorithms used in time-frequency masking instead of the EM algorithm.

The algorithm presented in this chapter combines time-frequency masking techniques and mixture of beamformers. The proposed algorithm has two main stages. In the first stage, the mixture time-frequency points are partitioned into a sufficient number of clusters using one of the time-frequency masking techniques described in Chapter 2. In the second stage, we will use the clusters obtained in the first stage to calculate covariance matrices, one for each cluster in each frequency bin. These covariance matrices and the time-frequency masks are then used in the mixture of MPDR beamformers. The resulting non-linear beamformer has low computational complexity and removes the musical noise found in time-frequency masked outputs at the expense of lower interference attenuation. The mixture of MPDR beamformers stage can be regarded as a post-processing step for sources separated by time-frequency masking. Two variants of the proposed method will be described and compared. The first one uses binary time-frequency masks, and the second one uses soft time-frequency masks.

Most time-frequency masking approaches use all the time-frequency bins at the same time to perform clustering [19, 98]. This is done to avoid the permutation ambiguity. If clustering was performed in each frequency bin (or group of frequencies) independently, the permutation problem has to be solved after clustering. However, our approach avoids the permutation ambiguity problem. This is because our approach assumes the knowledge of the desired source location and extracts the signal arriving from that direction. This avoids the permutation ambiguity problem, and the clustering stage can be performed in each frequency bin independently, in a group of frequencies, or using all the frequencies.

Clustering can be performed by using, for example, weighted k-means on the smoothed weighted histogram of the IPD/ILD as in DUET [19, 20], probabilistic modelling of the IPD/ILD as in [21, 94], or by clustering the normalised observation vector using a k-means algorithm as in MENUET [98, 99]. These methods are discussed in more detail in Chapter 2 and in the references therein. In this chapter, we will use the clustering method used in the MENUET speech separation algorithm to perform the time-frequency masking stage. The clustering method used in the MENUET algorithm has the advantage that it can utilise the information provided by more than two microphones.

The remainder of this chapter is structured as follows. Section 4.2 presents the proposed method. We present some simulation results that illustrate the performance of the proposed methods in Section 4.3. In Section 4.4, we describe an existing method that resembles the methods we proposed in this chapter, and we then compare it with our proposed mixture of beamformers approach. More experiments and comparisons of the proposed methods with other source separation algorithms can also be found in Chapter 5. In Section 4.5, we give the conclusions and the chapter summary.

4.2 Algorithm Steps

In this section, we outline the steps of the proposed method. The STFT, normalisation, clustering, and time-frequency mask design steps are common with the MENUET [98, 99] time-frequency masking algorithm. In the final two steps we design the mixture of beamformers based on the obtained time-frequency masks.

STFT

The time domain signals $\mathbf{x}(t)$ are converted into the time-frequency domain with a STFT to give $\mathbf{x}(n, f)$.

Normalisation

The normalisation is performed by selecting a reference microphone I and evaluating for each channel:

$$\bar{x_i}(n,f) = |x_i(n,f)| \exp\left[j\frac{\angle\left(x_i(n,f)/x_I(n,f)\right)}{4fc^{-1}d_{\text{max}}}\right] \tag{4.1}$$

where c is the speed of sound, and d_{max} is the maximum separation between the reference microphone I and any microphone i. This normalisation avoids the frequency dependence in the phase difference and holds the time delay information in the exponent term [99]. Then a unit-norm normalisation is applied to prevent outliers in the level ratio affecting the clustering performance [99]:

$$\bar{\mathbf{x}}(n,f) \leftarrow \frac{\bar{\mathbf{x}}(n,f)}{||\bar{\mathbf{x}}(n,f)||} \tag{4.2}$$

By this normalisation, $\bar{\mathbf{x}}(n,f)$ is dependent only on the source geometry [98, 99]. In $\bar{\mathbf{x}}(n,f)$ the phase difference information is held in the argument term, and the level ratio is normalised by the vector norm normalisation. $\bar{\mathbf{x}}(n,f)$ are N-dimensional complex vectors, and therefore clustering will be performed in an N-dimensional complex space.

Clustering

In this step, we partition the normalised observations $\bar{\mathbf{x}}(n,f)$ into k_x clusters $C_1, ..., C_{k_x}$. We use a clustering algorithm such as k-means to minimise the total sum of the squared distances between cluster members $\bar{\mathbf{x}} \in C_{q_x}$ and their centroid $\mathbf{c}_{q_x}$:

$$\mathcal{J} = \sum_{q_x=1}^{k_x} \sum_{\bar{\mathbf{x}} \in C_{q_x}} ||\bar{\mathbf{x}} - \mathbf{c}_{q_x}||^2 \tag{4.3}$$

The unit-norm normalisation makes the distance calculation in the clustering easier, because it projects the vectors on a unit hyper sphere [99]. If $\bar{\mathbf{x}}$ and the cluster centroids are projected on the unit hyper sphere, the square distance $||\bar{\mathbf{x}} - \mathbf{c}_{q_x}||^2 = 2(1 - \Re(\mathbf{c}_{q_x}^H \bar{\mathbf{x}}))$. In this case, the minimisation of the distance is equivalent to the maximisation of the real part of $\mathbf{c}_{q_x}^H \bar{\mathbf{x}}$, whose calculation is less demanding in terms of computational complexity [99]. The centroid $\mathbf{c}_{q_x}$ of a cluster C_{q_x} is calculated by:

$$\mathbf{c}_{q_x} = \sum_{\bar{\mathbf{x}} \in C_{q_x}} \frac{\bar{\mathbf{x}}}{|C_{q_x}|},\ \mathbf{c}_{q_x} = \frac{\mathbf{c}_{q_x}}{||\mathbf{c}_{q_x}||} \tag{4.4}$$

and where $|C_{q_x}|$ is the number of members in C_{q_x}. As in the GMM case (Chapter 3), the number of clusters does not need to be equal to the number of sources (which we assume unknown). We investigate the effect of the number of clusters in the experimental evaluation section later in this Chapter.

Most existing time-frequency masking methods use all the time-frequency bins at the same time to perform clustering [19, 98]. This is done to avoid the permutation ambiguity. In our method, the clustering stage can be performed in each frequency bin independently, in a group of frequencies, or using all the frequencies. This can be done because our method, as will be described in the next steps, extracts a desired source from a known location and therefore avoids the permutation ambiguity problem.

Designing Time-frequency masks

In this step, we consider two types of time-frequency masks. We either design time-frequency binary masks that selects the time-frequency points in one of the clusters:

$$\mathbf{M}_{q_x}(n,f) = \begin{cases} 1 & \bar{\mathbf{x}} \in C_{q_x} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

or soft masks that contain values between 0 and 1 that are computed based on the distance to cluster centroids. There are many possible distance functions that can be used to compute the soft masks. We use this heuristic function:

$$\rho_{q_x}(n,f) = \frac{1}{||\bar{\mathbf{x}}(n,f) - \mathbf{c}_{q_x}|| + \epsilon} \tag{4.6}$$

where ϵ is a very small number that prevents division by zero. Soft masks are then computed by normalising ρ as follows:

$$\mathbf{M}_{q_x}(n,f) = \frac{\rho_{q_x}(n,f)}{\sum_{q_x'=1}^{k_x} \rho_{q_x'}(n,f)} \tag{4.7}$$

We then compute the vector signals representing each cluster:

$$\mathbf{y}_{q_x}(n,f) = \mathbf{M}_{q_x}(n,f)\mathbf{x}(n,f) \tag{4.8}$$

Correlation matrix $\mathbf{R}_{q_x}(f)$ estimation

In this step, we calculate for each cluster the spatial covariance matrix in each frequency:

$$\mathbf{R}_{q_x}(f) = \sum_{n} \mathbf{y}_{q_x}(n, f)\mathbf{y}_{q_x}(n, f)^H \tag{4.9}$$

there is no need to scale the matrix $\mathbf{R}_{q_x}(f)$ with the number of samples as scaling has no effect on the MPDR beamformer weights. In order to avoid ill-conditioned or singular matrices, we regularise the correlation matrix by multiplying each diagonal element with $1 + \beta$, where β is very small number (we use $\beta = 1e - 3$). The effect of various values of β and other regularisation methods is studied in Section 5.2.6.

Mixture of beamformers

For each time-frequency point $(n,\ f)$, we calculate a weighted mixture of MPDR beamformers in the direction of the desired source. The estimate of the desired source is:

$$\hat{s}(n, f) = \sum_{q_x=1}^{k_x} \mathbf{M}_{q_x}(n, f)\, \mathbf{w}_{q_x}(f)\, \mathbf{x}(n, f) \tag{4.10}$$

where $\mathbf{M}_{q_x}(n, f)$ can be binary or soft masks and:

$$\mathbf{w}_{q_x}(f) = \frac{\mathbf{a}(f)^H\, \mathbf{R}_{q_x}^{-1}(f)}{\mathbf{a}(f)^H\, \mathbf{R}_{q_x}^{-1}(f)\, \mathbf{a}(f)} \tag{4.11}$$

In the following, we shall use the notation $\mathbf{w}_4$ to represent the mixture of beamformers which use binary masks, and $\mathbf{w}_5$ to represent the mixture of beamformers which use soft masks.

4.3 Experimental Evaluation

In this section we study the performance of the hard and soft versions of the non-linear beamformer proposed in this chapter. We use the setup, speech data, and evaluation measures described in Sections 3.5.1 and 3.5.2.

	$\mathbf{w}_4$	$\mathbf{w}_5$
STFT frame	1024	1024
STFT step	256	256
Clusters (2 mics)	$k_x = 15$	$k_x = 15$
(3 mics)	$k_x = 5$	$k_x = 5$
Frequency block size	512	512
Learning block duration	10 s	10 s

Table 4.1: *Algorithm parameters for* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

4.3.1 Algorithm Parameters

Unless mentioned otherwise, we use the values listed in Table 4.1 for the STFT frame size, STFT step size, number of clusters, frequency block size, and learning block duration. It will become clear later in this chapter why these values where chosen. In particular, increasing the number of clusters, or reducing the STFT step size or the frequency block size beyond the values listed in the table requires more computational time with an insignificant performance improvement.

4.3.2 Effect of Design Parameters

We investigate the effect of various parameters on the performance of the non-linear beamformers. We study the effect of the number of clusters, the size of groups of frequencies we perform the clustering at, and the duration of blocks we work on. In this experiment, we study the performance of the non-linear beamformer when four sources are operating in an anechoic environment, and the microphone array used has two or three microphones with a total array size $D = 5$ cm. We assume that the location of a desired source is known.

4.3.2.1 Effect of the Number of Clusters

Figure 4.1 shows the average performance at the output of the mixture of beamformers $\mathbf{w}_4$ as a function of the number of clusters k_x. The SIR increases with k_x, but the improvement is insignificant at $k_x > 10$.

Figure 4.2 shows the average performance at the output of the mixture of MVDR beamformers $\mathbf{w}_5$ as a function of the number of clusters k_x. We can see that using the soft masks instead of binary masks causes an increase in the SAR at the expense of a decrease of the SIR.

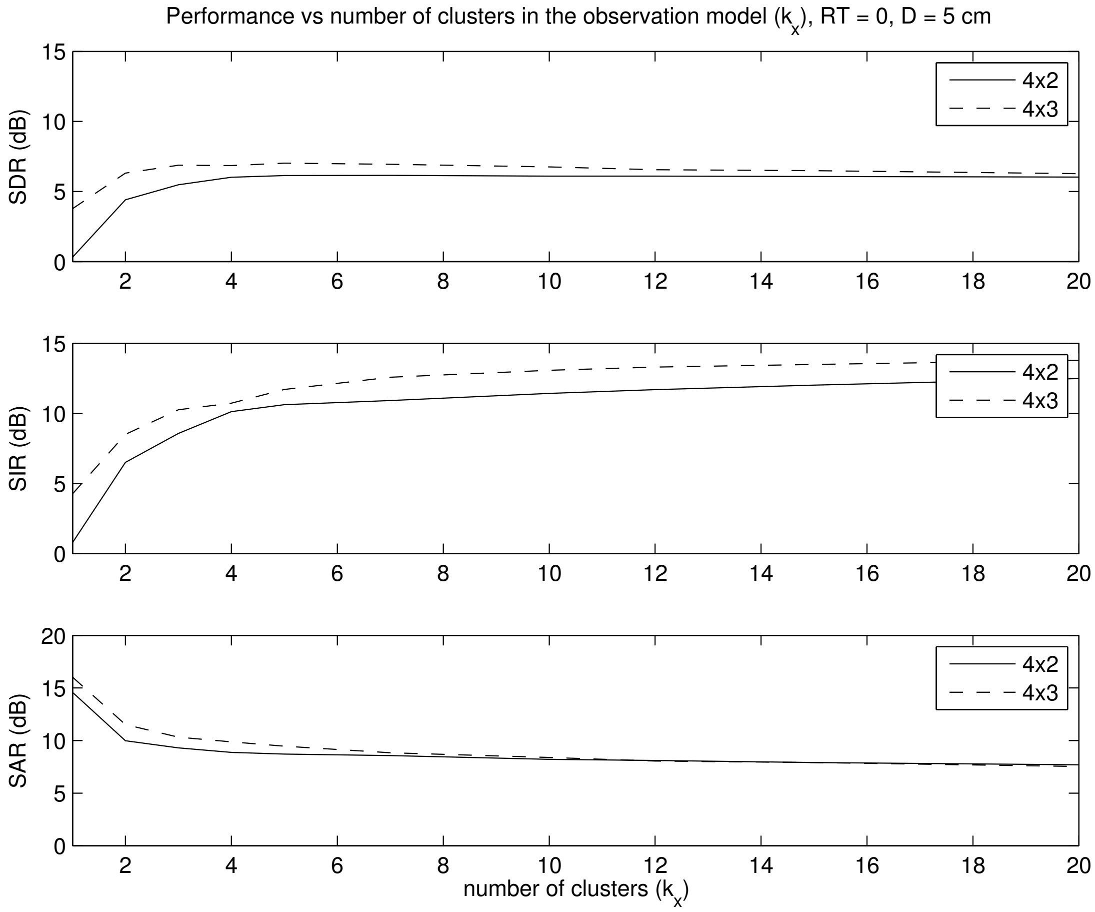

Figure 4.1: *Average performance of* $\mathbf{w}_4$ *in the anechoic case as a function of the number of clusters* k_x*.*

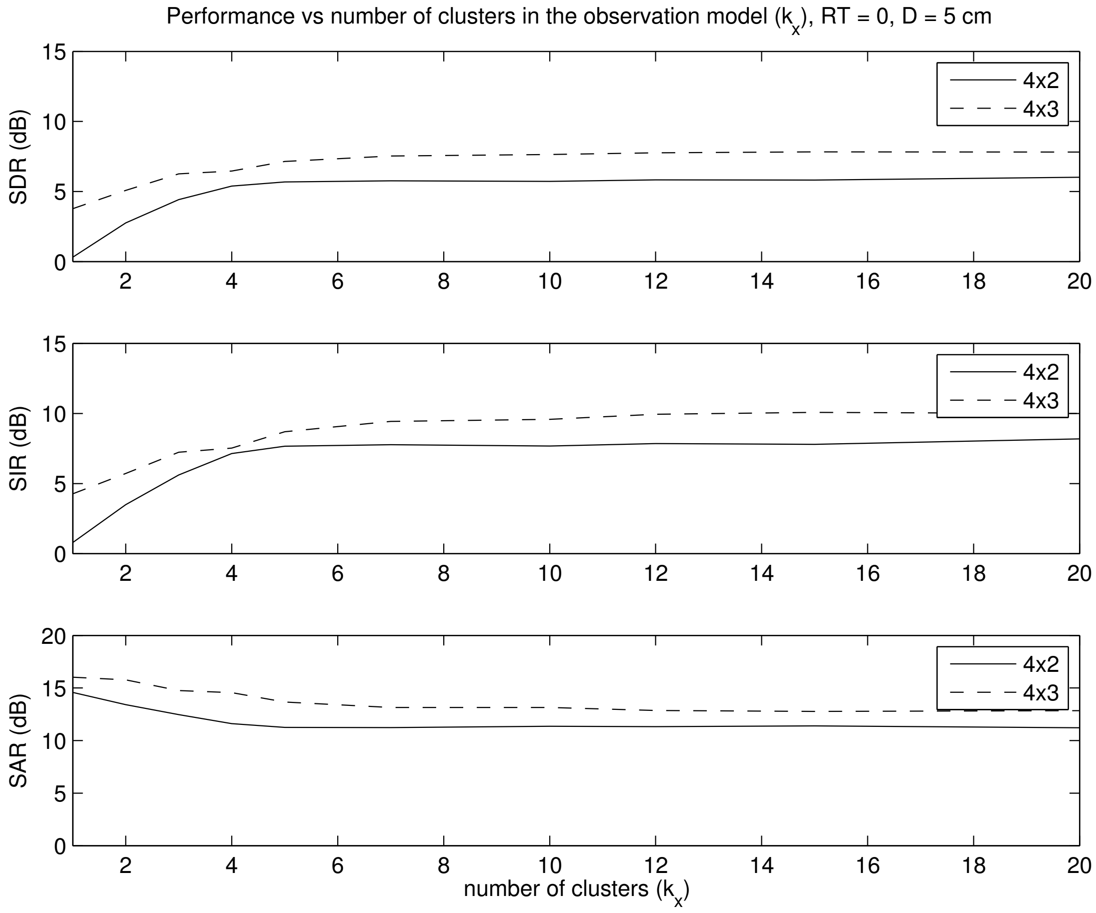

Figure 4.2: *Average performance of* $\mathbf{w}_5$ *in the anechoic case as a function of the number of clusters* k_x.

	time (s)	parameters
$\mathbf{w}_4$ (2 mics)	9.7	10 s block, $k_x = 15$
$\mathbf{w}_5$ (2 mics)	10.8	10 s block, $k_x = 15$
$\mathbf{w}_4$ (3 mics)	5.2	10 s block, $k_x = 5$
$\mathbf{w}_5$ (3 mics)	5.6	10 s block, $k_x = 5$
$\mathbf{w}_4$ (2 mics)	1.3	0.5 s block, $k_x = 15$
$\mathbf{w}_5$ (2 mics)	1.9	0.5 s block, $k_x = 15$
$\mathbf{w}_4$ (3 mics)	0.6	0.5 s block, $k_x = 5$
$\mathbf{w}_5$ (3 mics)	0.83	0.5 s block, $k_x = 5$

Table 4.2: *Computational time for* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

4.3.2.2 Effect of the Frequency Block Size

We study the effect of varying the number of frequency bins used to calculate the masks on the performance of the non-linear beamformers. Figures 4.3 and 4.4 show the average performance at the output of the non-linear beamformers in the anechoic case as a function of the number of frequency bins used to calculate the masks. The performance is fairly constant.

4.3.2.3 Effect of the Learning Block Duration

In this section, we study the effect of varying the duration of learning data on the performance of the non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$. Figures 4.5 and 4.6 show the average performance at the output of the non-linear beamformers in the anechoic case as a function of the learning block length. We can see that when using short learning blocks, the SIR increases, the SAR decreases (this had no audible artifacts), and the SDR remains fairly constant. The batch mode with short blocks of data can be used in applications where short delays are permissible, such as in human-computer interaction or surveillance.

4.3.3 Computational Time

In this section, we report the time it took for our MATLAB implementation of the non-linear beamformers to run on a 2.5 GHz CPU. The time reported is for the extraction of one 10 s speech source. We report the time our implementation took for the extraction of one 10 s speech source, and one 0.5 s speech source. Table 4.2 shows the results.

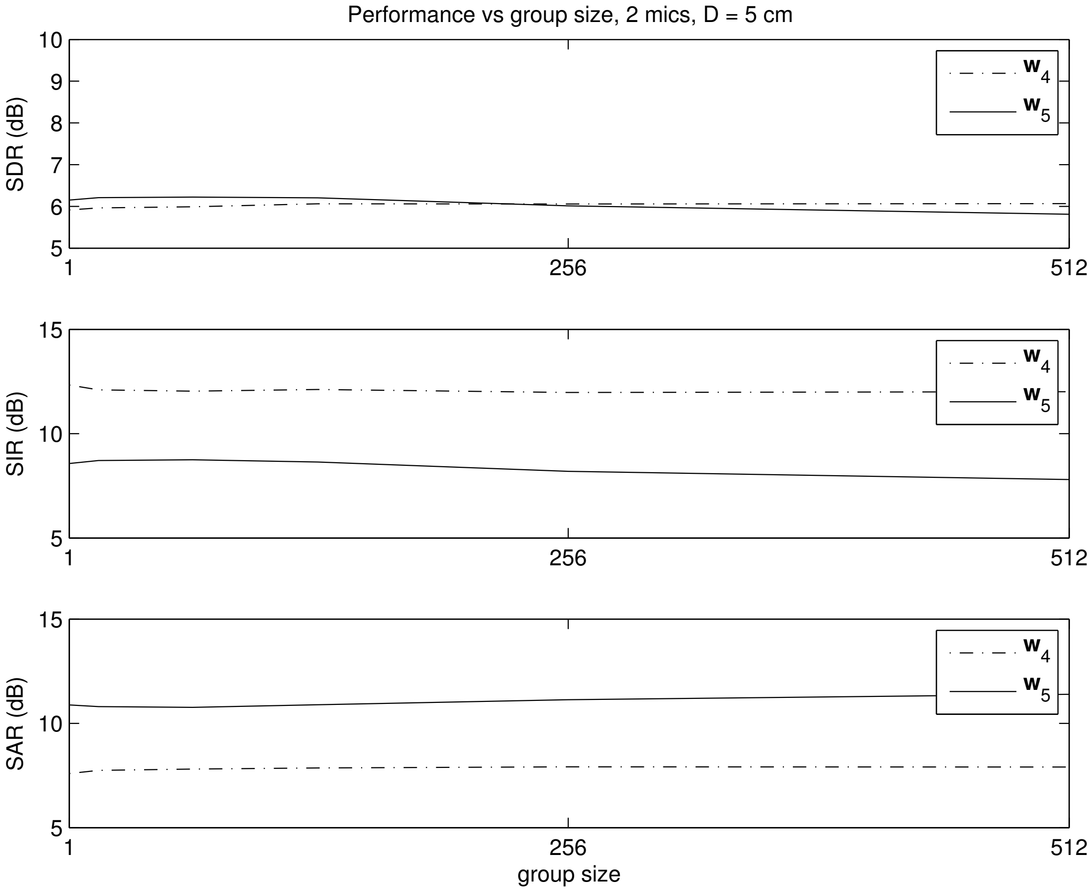

Figure 4.3: *Separation using two microphones: Average performance of* $\mathbf{w}_4$ *and* $\mathbf{w}_5$ *in the anechoic case vs number of frequency bins used to calculate the masks. Used parameter:* $k_x = 15$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

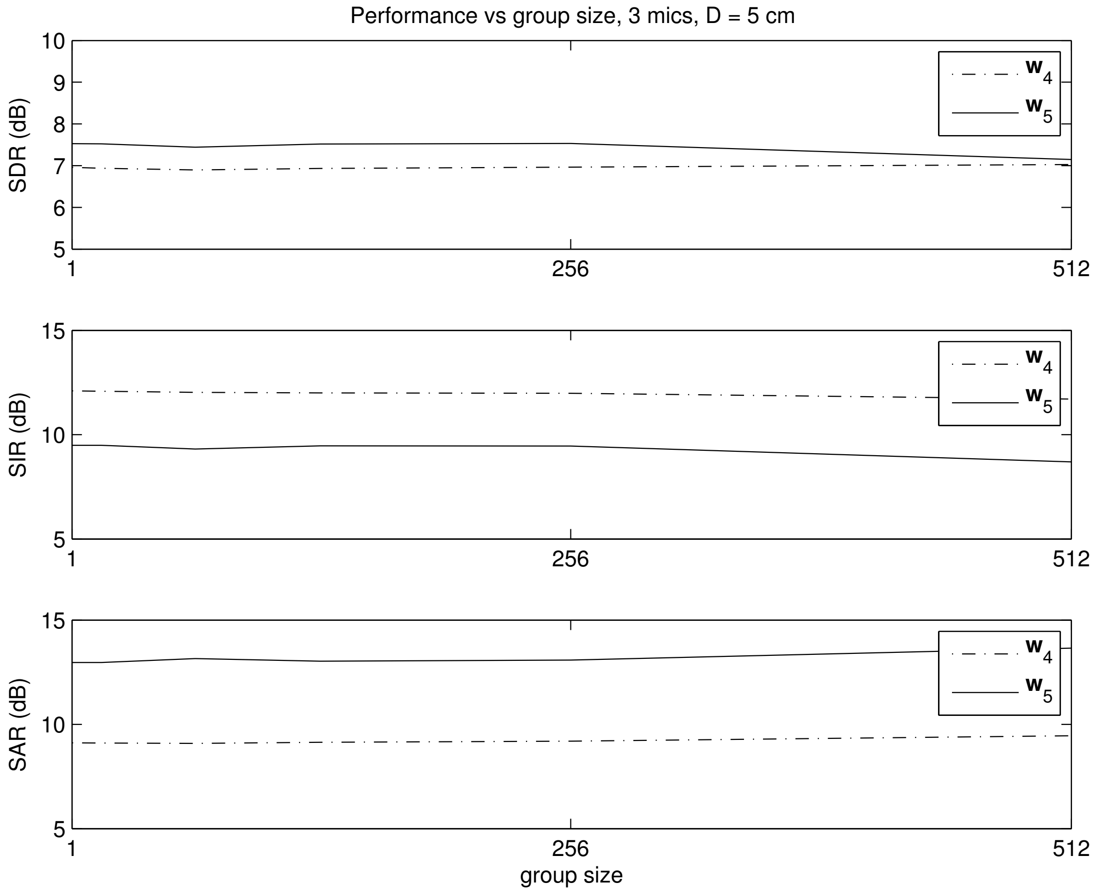

Figure 4.4: *Separation using three microphones: Average performance of* $\mathbf{w}_4$ *and* $\mathbf{w}_5$ *in the anechoic case vs number of frequency bins used to calculate the masks. Used parameter:* $k_x = 5$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

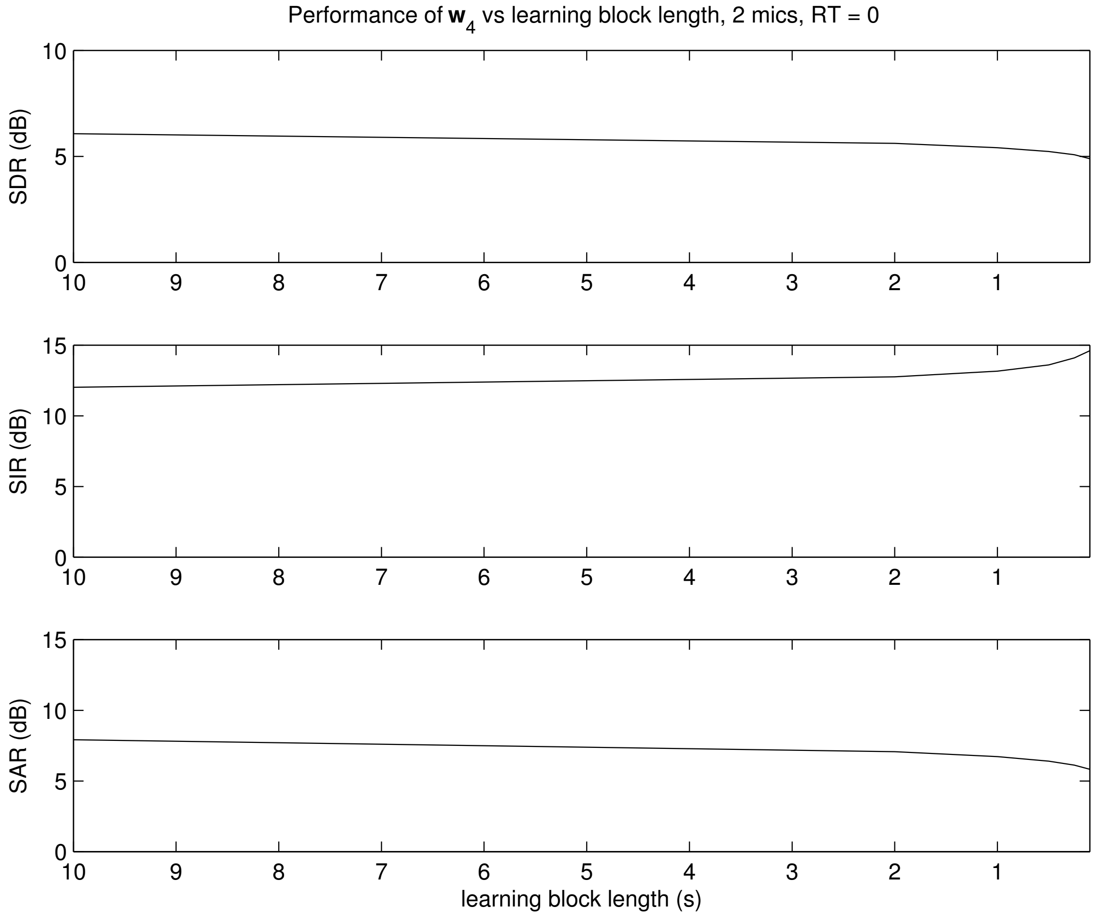

Figure 4.5: *Separation using two microphones: Average performance of* w_4 *in the anechoic case vs learning block length in seconds. Used parameter:* $k_x = 15$.

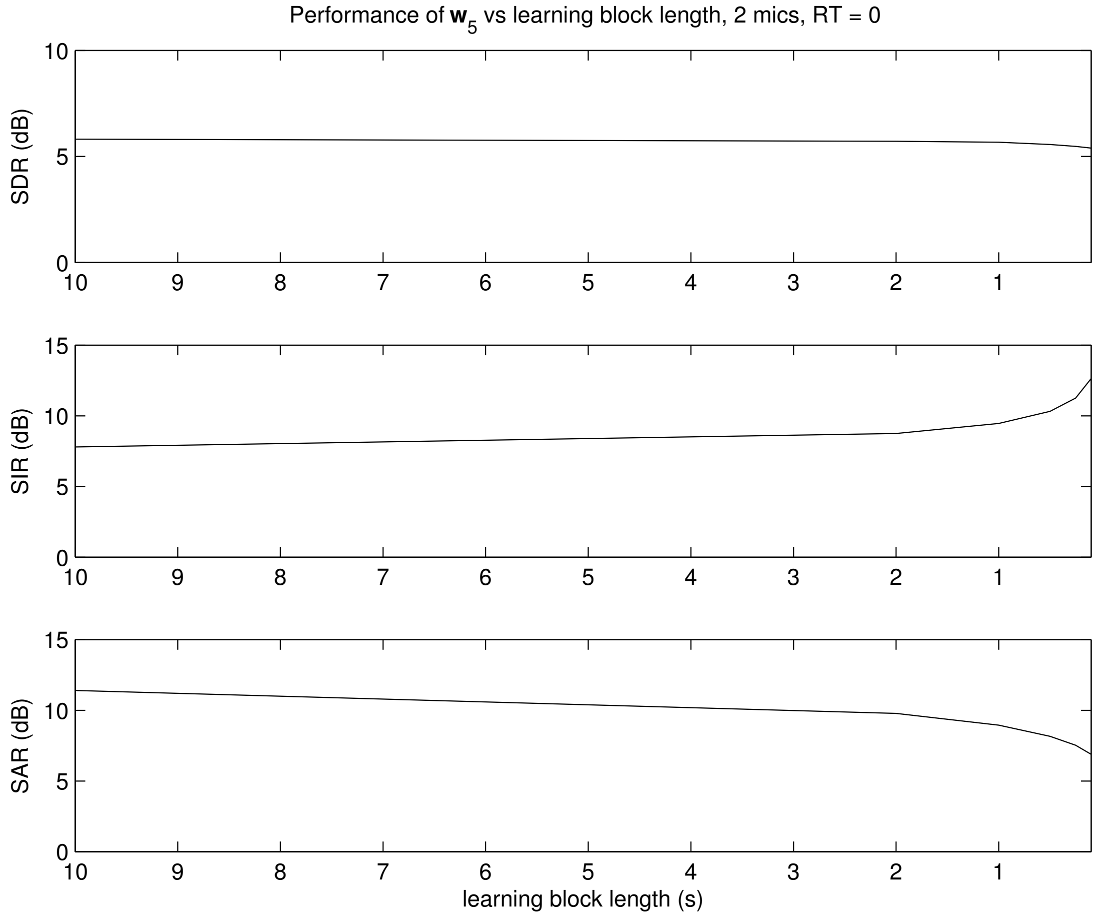

Figure 4.6: *Separation using two microphones: Average performance of* $\mathbf{w}_5$ *in the anechoic case vs learning block length in seconds. Used parameter:* $k_x = 15$.

4.4 Comparison with an Existing Method

During this research, we noticed that the method proposed in this chapter resembles an existing method proposed by Cermak et al. in [120]. The method described in [120] combines time-frequency binary masking and a set of beamformers, termed a beamformer array by the authors. In this section, we describe this existing method, and then compare it with our proposed mixture of beamformers approach.

4.4.1 Time-Frequency Masking and Beamformer Array

We start our comparison by highlighting the assumptions used in our method and the method used in [120]. Our method assumes that the direction of the desired source is known or estimated a priori, and there is no need to know or estimate the number of interferers. The interferers can be point sources, extended sources, or any form of noise. The goal of our proposed method is to extract a desired point source from a known location (source extraction). The method proposed in [120] assumes that the number of sources is known, but does not assume that their location is known. The goal in [120] is to estimate all the sources (source separation). To achieve source separation in [120], time-frequency masking is first employed, followed by the estimation of the mixing vector (steering vector) for each source. A beamformer array is then designed using the estimated mixing vector and the time-frequency masks. The method proposed in [120] has similar steps to the method we proposed in Section 4.2, with the difference being in the design of the beamformers. In [120], the authors estimate binary time-frequency masks $\mathbf{M}_j,\ j = 1, ..., M$ (M is the known number of sources), and then estimate the mixing vector $\hat{\mathbf{a}}_j(f)$ for source j using a least square estimate:

$$\hat{\mathbf{a}}_j(f) = \frac{\mathrm{E}\left[\mathbf{x}(n,f)\mathbf{M}_j(n,f)x_I^*(n,f)\right]}{\mathrm{E}\left[|\mathbf{M}_j(n,f)x_I(n,f)|^2\right]} \tag{4.12}$$

where x_I is the I^{th} channel of the observed mixture, I is a reference channel, and $(.)^*$ denotes the complex conjugate. A beamformer array is then designed using the estimated mixing vector and the time-frequency masks.

Let us now focus on separating one source (with index d). The main idea behind the beamformer array is to compose B different mixtures (B is defined shortly) from the pre-separated signals:

$$\mathbf{y}_{q_x}(n,f) = \mathbf{M}_{q_x}(n,f)\mathbf{x}(n,f) : 1 \leq q_x \leq M \tag{4.13}$$

provided by time-frequency masking, which are then filtered by B MVDR beamformers. The input to each one of B beamformers is a mixture which includes the pre-separated target signal ($\mathbf{y}_d$) and different pre-separated interferers (a selection of $\mathbf{y}_{q_x}, q_x \neq d$). All interferes are used at least once in these mixtures. Each one of the B beamformers point to the direction of the target signal, and attenuate the selected pre-separated interferers. Finally, we add together all the outputs of the beamformers.

The number of beamformers B depends on $Z \in \{1, ..., N-1\}$, the number of interferers selected in each mixture (N is the number of channels). For obvious reasons, the value of Z should not be higher than the number of degrees of freedom in MVDR beamformers ($N-1$). Once a value for Z is selected, B can be estimated by:

$$B = \frac{(M-1)!}{Z!(M-1-Z)!} \tag{4.14}$$

We compose B mixtures. Each mixture $\mathbf{u}_{d,b}$ includes the pre-separated target signal and a selection of the pre-separated interferers:

$$\mathbf{u}_{d,b}(n,f) = \mathbf{y}_d + \sum_{g \in z_b} \mathbf{y}_g \tag{4.15}$$

where z_b represents the Z interferers selected for mixture $\mathbf{u}_{d,b}$. We design B beamformers, each given by:

$$\mathbf{w}_{\hat{d},b}(f) = \frac{\hat{\mathbf{a}}_d(f)^H \, \mathbf{R}_b^{-1}(f)}{\hat{\mathbf{a}}_d(f)^H \, \mathbf{R}_b^{-1}(f) \, \hat{\mathbf{a}}_d(f)} \tag{4.16}$$

where the beamformer $\mathbf{w}_{\hat{d},b}$ [1] is designed to filter $\mathbf{u}_{d,b}$, and the correlation matrix $\mathbf{R}_b(f)$ is given by:

$$\mathbf{R}_b(f) = \sum_n \left(\sum_{g \in z_b} \mathbf{y}_g(n,f) \right) \left(\sum_{g \in z_b} \mathbf{y}_g(n,f) \right)^H \tag{4.17}$$

Finally, we add together all the outputs of the beamformers.

[1] We use the notation $\hat{d}$ in $\mathbf{w}_{\hat{d},b}$ to remind the reader that an estimate of the mixing vector $\hat{\mathbf{a}}_d$ is used.

4.4.2 Comparison between the Mixture of Beamformers and the Beamformer Array

The comparison between the two approaches can be easily understood by an example. Let us take an example where the number of sources $M = 4$ and the number of microphones $N = 2$. In this section, we compare the equations involved in the extraction of one source with index d using the beamformer array method, and our proposed mixture of beamformers method. To facilitate the comparison, we assume that the mixture of beamformers will cluster $k_x = 4$ components. Therefore, for both methods, the time-frequency masking stage will produce four signals corresponding to the original sources (up to an arbitrary ordering). Binary time-frequency masks will be used in this comparison. We choose the desired source to be source number 4 (This particular index $d = 4$ reduces the confusion between a source index and a beamformer index b).

4.4.2.1 Beamformer Array Approach

In the two microphone case, there is only one permissible value for Z, and we have $Z = 1$, $B = 3$. We first estimate the mixing vector for the desired source at each frequency:

$$\hat{\mathbf{a}}_4(f) = \frac{\mathrm{E}\left[\mathbf{x}(n,f)\mathbf{M}_4(n,f)x_1^*(n,f)\right]}{\left[|\mathbf{M}_4(n,f)x_1(n,f)|^2\right]} \tag{4.18}$$

where we used the first channel as the reference channel I. The input signals of the beamformers are:

$$\begin{aligned} \mathbf{u}_{4,1}(n,f) &= \mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_1(n,f)\mathbf{x}(n,f) && (4.19)\\ \mathbf{u}_{4,2}(n,f) &= \mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_2(n,f)\mathbf{x}(n,f) && (4.20)\\ \mathbf{u}_{4,3}(n,f) &= \mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_3(n,f)\mathbf{x}(n,f) && (4.21) \end{aligned}$$

The beamformer array consists of the following beamformers:

$$\mathbf{w}_{\hat{4},1}(f) \quad = \quad \frac{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_1^{-1}(f)}{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_1^{-1}(f) \, \hat{\mathbf{a}}_4(f)} \tag{4.22}$$

$$\mathbf{w}_{\hat{4},2}(f) \quad = \quad \frac{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_2^{-1}(f)}{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_2^{-1}(f) \, \hat{\mathbf{a}}_4(f)} \tag{4.23}$$

$$\mathbf{w}_{\hat{4},3}(f) \quad = \quad \frac{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_3^{-1}(f)}{\hat{\mathbf{a}}_4(f)^H \, \mathbf{R}_3^{-1}(f) \, \hat{\mathbf{a}}_4(f)} \tag{4.24}$$

where

$$\mathbf{R}_1(f) \quad = \quad \sum_n \left(\mathbf{M}_1(n,f)\mathbf{x}(n,f)\right) \left(\mathbf{M}_1(n,f)\mathbf{x}(n,f)\right)^H \tag{4.25}$$

$$\mathbf{R}_2(f) \quad = \quad \sum_n \left(\mathbf{M}_2(n,f)\mathbf{x}(n,f)\right) \left(\mathbf{M}_2(n,f)\mathbf{x}(n,f)\right)^H \tag{4.26}$$

$$\mathbf{R}_3(f) \quad = \quad \sum_n \left(\mathbf{M}_3(n,f)\mathbf{x}(n,f)\right) \left(\mathbf{M}_3(n,f)\mathbf{x}(n,f)\right)^H \tag{4.27}$$

The estimate of source four is:

$$\hat{s}_4^{\text{BA}}(n,f) = \mathbf{w}_{\hat{4},1}(f)\mathbf{u}_{4,1}(n,f) + \mathbf{w}_{\hat{4},2}(f)\mathbf{u}_{4,2}(n,f) + \mathbf{w}_{\hat{4},3}(f)\mathbf{u}_{4,3}(n,f) \tag{4.28}$$

4.4.2.2 Mixture of Beamformers Approach

Our proposed method uses a mixture of beamformers in the (known) direction of the desired source $\mathbf{a}_4(f)$:

$$\hat{s}_4^{\text{MOB}}(n,f) = \sum_{q_x=1}^{4} \mathbf{M}_{q_x}(n,f) \, \mathbf{w}_{4,q_x}(f) \, \mathbf{x}(n,f) \tag{4.29}$$

where:

$$\mathbf{w}_{4,q_x}(f) = \frac{\mathbf{a}_4(f)^H \, \mathbf{R}_{q_x}^{-1}(f)}{\mathbf{a}_4(f)^H \, \mathbf{R}_{q_x}^{-1}(f) \, \mathbf{a}_4(f)} \tag{4.30}$$

and where:

$$\begin{aligned}
\mathbf{R}_1(f) &= \sum_n \left(\mathbf{M}_1(n,f)\mathbf{x}(n,f)\right)\left(\mathbf{M}_1(n,f)\mathbf{x}(n,f)\right)^H && (4.31)\\
\mathbf{R}_2(f) &= \sum_n \left(\mathbf{M}_2(n,f)\mathbf{x}(n,f)\right)\left(\mathbf{M}_2(n,f)\mathbf{x}(n,f)\right)^H && (4.32)\\
\mathbf{R}_3(f) &= \sum_n \left(\mathbf{M}_3(n,f)\mathbf{x}(n,f)\right)\left(\mathbf{M}_3(n,f)\mathbf{x}(n,f)\right)^H && (4.33)\\
\mathbf{R}_4(f) &= \sum_n \left(\mathbf{M}_4(n,f)\mathbf{x}(n,f)\right)\left(\mathbf{M}_4(n,f)\mathbf{x}(n,f)\right)^H && (4.34)
\end{aligned}$$

Note that $\mathbf{R}_1(f)$, $\mathbf{R}_2(f)$, $\mathbf{R}_3(f)$ are similar to what we calculate in the beamformer array approach (Equations (4.25)-(4.27)).

4.4.2.3 Comparison

Let us compare our proposed mixture of beamformers approach with the beamformer array approach of [120]. We can write (4.28) as follows:

$$\begin{aligned}
\hat{s}_4^{\mathrm{BA}}(n,f) = {} & \mathbf{w}_{\hat{4},1}(f)\left(\mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_1(n,f)\mathbf{x}(n,f)\right) + \\
& \mathbf{w}_{\hat{4},2}(f)\left(\mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_2(n,f)\mathbf{x}(n,f)\right) + \\
& \mathbf{w}_{\hat{4},3}(f)\left(\mathbf{M}_4(n,f)\mathbf{x}(n,f) + \mathbf{M}_3(n,f)\mathbf{x}(n,f)\right) \qquad (4.35)
\end{aligned}$$

which can be further simplified to:

$$\begin{aligned}
\hat{s}_4^{\mathrm{BA}}(n,f) = {} & \mathbf{M}_1(n,f)\mathbf{w}_{\hat{4},1}(f)\mathbf{x}(n,f) + \\
& \mathbf{M}_2(n,f)\mathbf{w}_{\hat{4},2}(f)\mathbf{x}(n,f) + \\
& \mathbf{M}_3(n,f)\mathbf{w}_{\hat{4},3}(f)\mathbf{x}(n,f) + \\
& \mathbf{M}_4(n,f)\left(\mathbf{w}_{\hat{4},1}(f) + \mathbf{w}_{\hat{4},2}(f) + \mathbf{w}_{\hat{4},3}(f)\right)\mathbf{x}(n,f) \qquad (4.36)
\end{aligned}$$

We remind the reader that one and only one of the binary masks $\mathbf{M}_1(n,f)$, $\mathbf{M}_2(n,f)$, $\mathbf{M}_3(n,f)$, and $\mathbf{M}_4(n,f)$ can have a non-zero value at any time-frequency point.

The mixture of beamformers estimate can be written as:

$$\begin{aligned}\hat{s}_4^{\text{MOB}}(n,f) &= \mathbf{M}_1(n,f)\mathbf{w}_{4,1}(f)\mathbf{x}(n,f) + \\ &\quad \mathbf{M}_2(n,f)\mathbf{w}_{4,2}(f)\mathbf{x}(n,f) + \\ &\quad \mathbf{M}_3(n,f)\mathbf{w}_{4,3}(f)\mathbf{x}(n,f) + \\ &\quad \mathbf{M}_4(n,f)\mathbf{w}_{4,4}(f)\mathbf{x}(n,f) \end{aligned} \tag{4.37}$$

We can see that main two differences between (4.36) and (4.37) are the use of the known or estimated mixing vectors, and the beamformer used when the target component mask ($\mathbf{M}_4(n,f)$ in our case) is equal to one. The mixture of beamformers approach use $\mathbf{w}_{4,4}(f)$, while the beamformer array approach use $\left(\mathbf{w}_{\hat{4},1}(f) + \mathbf{w}_{\hat{4},2}(f) + \mathbf{w}_{\hat{4},3}(f)\right)$. Contrary to what was claimed in [120], we can see that the beamformer array is not distortionless. When the target component mask $\mathbf{M}_4(n,f)$ is equal to one, the beamformer output is equal to $(\mathbf{w}_{\hat{4},1}(f)+ \mathbf{w}_{\hat{4},2}(f) + \mathbf{w}_{\hat{4},3}(f))\mathbf{x}(n,f)$, which has a gain of three for a signal arriving from the target estimated direction. This scaling can be seen as a form of post-processing the output to increase the interference rejection while introducing artifacts.

We now compare the performance of the beamformer array approach with the mixture of beamformers approach ($\mathbf{w}_4$). We use similar setup and performance measures to the experiments described in Section 4.3. We use two microphones with a spacing $d = 5$ cm, a STFT frame of 1024 samples, a STFT step size of 256 samples, a learning block duration of 10 s, and a frequency block size of 512 bins (all frequency bins). We report the performance of $\mathbf{w}_4$ when using $k_x = 4$, and $k_x = 15$ (more results can be found in Section 4.3). We also compare the performance of the two approaches before and after a post-processing step as proposed in [108] (method described in Section 2.8.7). In the post-processing stage, we use an ENSIR threshold of 3 dB (i.e. we null a time-frequency point if the ENSIR is less than 3 dB).

The performance results are summarised in Table 4.3. The beamformer array approach can achieve a SDR of 7.3 dB, a SIR of 13.5 dB, and a SAR of 9.0 dB. In comparison, the mixture of beamformers $\mathbf{w}_4$ with $k_x = 4$ can achieve a SDR of 6.0 dB, a SIR of 10.1 dB, and a SAR of 8.9 dB. The mixture of beamformers $\mathbf{w}_4$ with $k_x = 15$ can achieve a SDR of 6.1 dB, a SIR of 12.0 dB, and a SAR of 7.9 dB. When an estimated mixing vector is used in the mixture of beamformers ($k_x = 4$), we get a SDR of 5.7 dB, a SIR of 9.6 dB, and a SAR of 8.7 dB. We can see that the beamformer array approach outperforms the mixture of beamformers approach, and

	SDR	SIR	SAR
beamformer array, $k_x = 4$	7.3	13.5	9.0
mixture of beamformers $\mathbf{w}_4$, $k_x = 4$	6.0	10.1	8.9
mixture of beamformers $\mathbf{w}_4$, $k_x = 4$, estimated $\mathbf{a}$	5.7	9.6	8.7
mixture of beamformers $\mathbf{w}_4$, $k_x = 15$	6.1	12.0	7.9
beamformer array, $k_x = 4$, with post-processing	6.5	17.3	7.2
mixture of beamformers $\mathbf{w}_4$, $k_x = 4$, with post-processing	6.5	16.9	7.3
mixture of beamformers $\mathbf{w}_4$, $k_x = 15$, with post-processing	6.0	17.3	6.7

Table 4.3: *Performance comparison between the beamformer array and the mixture of beamformers approaches.*

that using an estimated mixing vector in the mixture of beamformers reduced its performance slightly. When we add a post-processing stage, the beamformer array approach can achieve a SDR of 6.5 dB, a SIR of 17.3 dB, and a SAR of 7.2 dB. In comparison, the mixture of beamformers $\mathbf{w}_4$ with $k_x = 4$ can achieve a SDR of 6.5 dB, a SIR of 16.9 dB, and a SAR of 7.3 dB. The mixture of beamformers $\mathbf{w}_4$ with $k_x = 15$ can achieve a SDR of 6.0 dB, a SIR of 17.3 dB, and a SAR of 6.7 dB. We can see that with a post-processing step the beamformer array approach and the mixture of beamformers approach have a comparable performance.

4.5 Chapter Summary

We presented a modification to the mixture of MPDR beamformers presented in the previous chapter. The presented methods combine clustering algorithms used in time-frequency masking with the mixture of beamformers. This results in a significant reduction in computational complexity.

The non-linear beamformers have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_4$ gives better interference rejection, while the non-linear beamformer $\mathbf{w}_5$ gives better SAR. In terms of computational complexity, both non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$ have similar computational complexity, and require significantly less time than non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$ (Section 3.5.6). We then compared the mixture of beamformers approach with an existing method proposed by Cermak et al. in [120]. The method proposed by Cermak et al. use a time-frequency binary masking stage followed by beamforming stage. We showed that the proposed method gives a comparable performance to the method in [120] when we use a post-processing step. More experiments and comparisons of the proposed methods with other source separation algorithms can be found in Chapter 5.

Chapter 5

Further Results and Comparisons

5.1 Introduction

In Chapters 3 and 4, we proposed different non-linear beamformers that can extract a desired source from a known location. In particular, we proposed three non-linear beamformers in Chapter 3 where the signal estimator is a mixture of MMSE beamformers (denoted as $\mathbf{w}_1$), a mixture of MVDR beamformers (denoted as $\mathbf{w}_2$), and a mixture of MPDR beamformers (denoted as $\mathbf{w}_3$). Then in Chapter 4, we presented a modification to the mixture of MPDR beamformers $\mathbf{w}_3$ where we employed clustering algorithms used in time-frequency masking instead of the EM algorithm. Two variants of the proposed method were described. The first one uses binary time-frequency masks (denoted as $\mathbf{w}_4$), and the second one uses soft time-frequency masks (denoted as $\mathbf{w}_5$). In Sections 3.5 and 4.3, we presented some experimental results that illustrate the performance of the proposed methods and we studied the effect of design parameters on the extraction performance.

In this chapter, we report more experimental results and compare the proposed non-linear beamformers with some other source separation methods. The chapter is structured as follows. In Section 5.2.1, we describe the rooms used in the experimental evaluation. The speech data used in the experiments is described in Section 5.2.2. In Section 5.2.4, we describe the evaluation metrics. In Section 5.2.5, the algorithm parameters we use unless mentioned otherwise are presented. In Section 5.2.6, we investigate the effect of DOA offset and regularisation. In Sections 5.2.7 and 5.2.8, we study the performance of the proposed non-linear beamformers in room reverberation and compare their performance with the MENUET and the FMV algorithms. Then, in Section 5.2.9, we report the performance of the non-linear beamformers in real life recordings and compare their performance with the MENUET and the FMV algorithms.

Samples of the speech files used in the experiments can be found in a CD attached to the thesis.

5.2 Experimental Evaluation

5.2.1 Setup

Our experiments were made in three rooms with different characteristics and arrangements of the sources and the microphones. In all rooms, the number of the sources was four. In the first two rooms, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [50]. The reverberation time was varied in order to study the effect of reverberation time on the performance. In the third room, multichannel recordings of several speech sources were recorded in a physical room with a reverberation time of 810 ms.

Figures 5.1 shows the first room. Two microphone arrays were used. The first has three microphones with a spacing $d = 2.5$ cm, and the second has two microphones with a spacing $d = 5$ cm. Both microphone arrays have a total length of $D = 5$ cm. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{-45,\, -15,\, 10,\, 50\}^\circ$. This room is used in all the experiments in this chapter except for the experiments in Sections 5.2.8 and 5.2.9.

In Section 5.2.8, simulated recordings in the room illustrated in Figure 5.2 were used. The room is similar to room 1, but the arrangements of the sources is different. The desired source was placed 10 cm away from the microphone array at angle $\phi = 10^\circ$, while the interferers were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{-45,\, -15,\, 50\}^\circ$. Although the desired source was in close proximity to the microphones, we scaled the power of the sources so that the received power of all the sources at the microphones are equal. In this setup, the desired source is closer to the microphones than the interfering sources, and therefore suffers from less reverberation. This arrangement of sources was considered because it represents many practical scenarios where the desired source is in close proximity to the microphones.

In Section 5.2.9, live recordings in the room illustrated in Figure 5.3 were used. In this room, the microphone array has two microphones with spacing $d = 7$ cm. The desired source was placed 30 cm away from the microphone array at $\phi = 0^\circ$, while the interferers were placed in a semi-circle of radius 1.5 m around the microphone arrays at angles $\phi = \{-60,\, -30,\, 50\}^\circ$. The power of the sources was scaled so that the received power of all the sources at the microphones are equal. In this setup, the desired source is closer to the microphones than the interfering

		$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
STFT frame		1024	1024	1024
STFT step		256	256	256
GMM components	(2 mics)	$k_s = 2, k_v = 15$	$k_s = 2, k_v = 15$	$k_x = 15$
	(3 mics)	$k_s = 2, k_v = 5$	$k_s = 2, k_v = 5$	$k_x = 5$
EM Iterations	(2 mics)	100	100	50
	(3 mics)	100	100	20
Learning block duration		10 s	10 s	10 s

Table 5.1: *Algorithm parameters for* $\mathbf{w}_1$, $\mathbf{w}_2$, *and* $\mathbf{w}_3$.

sources, and therefore suffers from less reverberation.

5.2.2 Speech Data

We used speech files taken from the TIMIT speech corpus [113] to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of the sources in each mixture was four.

5.2.3 Demonstration CD

Attached to this thesis is a CD with samples of the speech signals used in the experiments and the outputs of the non-linear beamformers ($\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$, and $\mathbf{w}_5$), the MENUET algorithm, and the FMV algorithm in the three rooms.

5.2.4 Evaluation Measures

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we used the SDR, SIR and the SAR calculated as defined in [53]. The computation of the evaluation measures is detailed in Section 2.7.

Unless mentioned otherwise, the SDR, SIR and SAR values in our results were averaged over all the sources and mixtures.

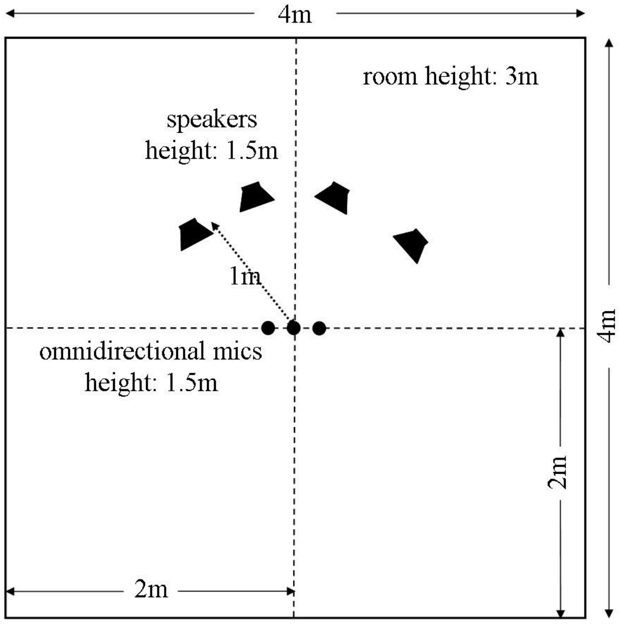

Figure 5.1: *Layout of room 1.*

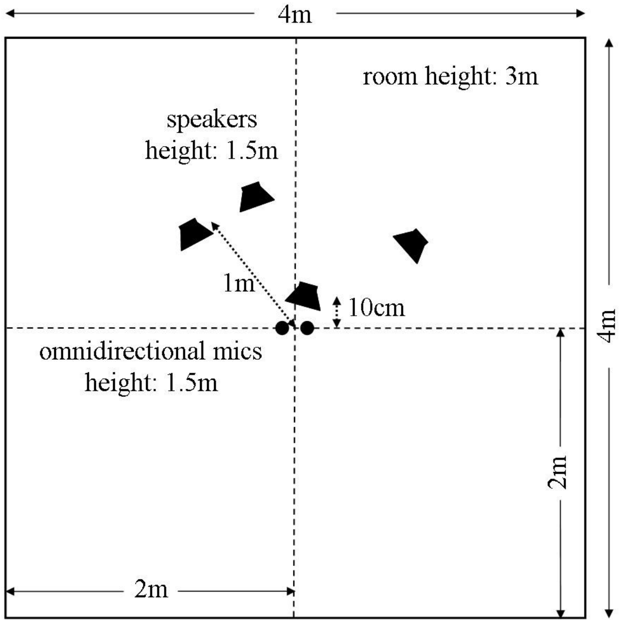

Figure 5.2: *Layout of room 2.*

		$\mathbf{w}_4$	$\mathbf{w}_5$
STFT frame		1024	1024
STFT step		256	256
Clusters	(2 mics)	$k_x = 15$	$k_x = 15$
	(3 mics)	$k_x = 5$	$k_x = 5$
Frequencies	(2 mics)	512	512
	(3 mics)	512	512
Learning block duration		10 s	10 s

Table 5.2: *Algorithm parameters for* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

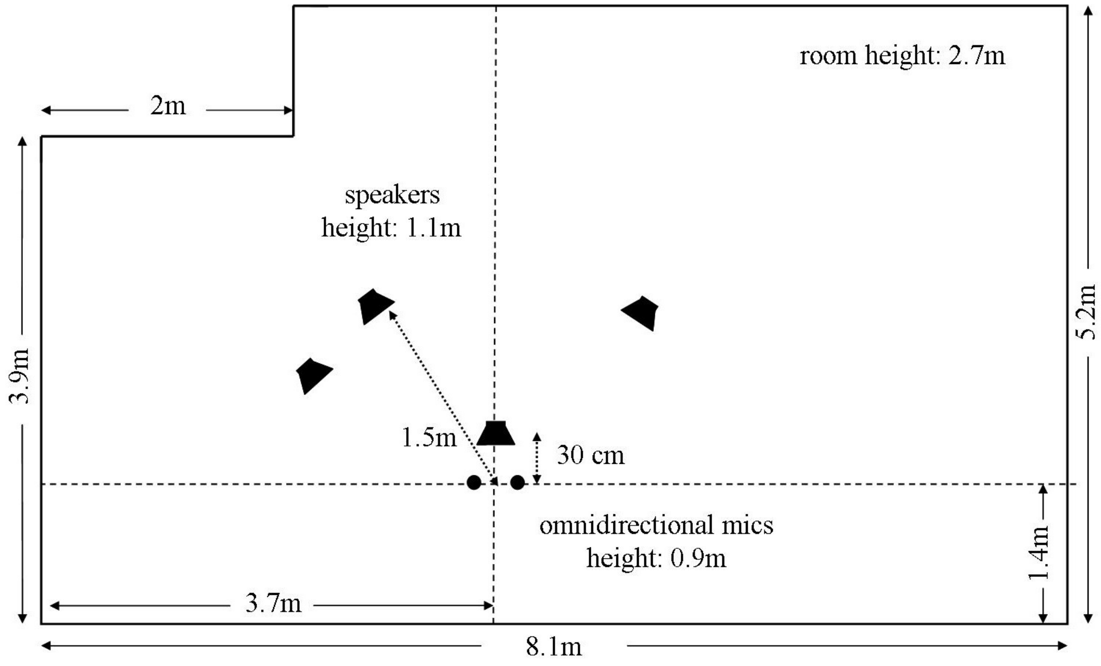

Figure 5.3: *Layout of room 3.*

5.2.5 Algorithm Parameters

Unless mentioned otherwise, we use the values listed in Table 5.1 for the STFT frame size, STFT step size, number of GMM components, and number of iterations for the non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$. Additionally, we use the values listed in Table 5.2 for the STFT frame size, STFT step size, number of clusters, and number of frequencies used in each learning block for the non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$. Increasing the number of GMM components/clusters or the number of EM iterations, or reducing the STFT step size or the frequency block size beyond the values listed in the table requires more computational time with an insignificant performance improvement.

5.2.6 Effect of DOA Offset and Regularisation

In some applications, the DOA of the desired source is scanned across a region of interest in space or estimated from the observed data. However, the desired signal can arrive from a different direction than that assumed. In this section, we test the effect of the mismatch between the assumed DOA of the desired source and the true one. There are many methods that have been developed to enhance the robustness of beamforming techniques against DOA mismatch, and other mismatches in the model, such as microphone gain and phase or location of microphones. Incorporating them in the mixture of beamformers framework should be straightforward. In all our results in this thesis we used multiplicative diagonal loading as suggested in [62]. In this section, we study and compare three regularisation methods: multiplicative diagonal loading [62], (additive) diagonal loading [52], and eigenvalue thresholding [110]. Another method for adding mismatch robustness to beamformers is incorporate additional linear constraints to the MVDR or the MPDR beamformers. The beamformers in which additional linear constraints are imposed are referred to as linear constrained minimum variance (LCMV) beamformers [52]. In LCMV beamformers, robustness against DOA offsets can be enhanced by forcing a flatter directivity patterns near the signal direction. This can be done using additional directional or derivative constraints [52]. Each constraint removes one of the remaining degrees of freedom available to attenuate interferers. We are not going to study LCMV methods in this thesis.

We first use multiplicative regularisation applied to the diagonal terms of the correlation matrix as suggested in [62]. In this regularisation method, we multiply each diagonal element in the correlation matrix with $1 + \beta$, where β is very small number. We first use $\beta = 1e - 3$. Figures 5.4, 5.5, 5.6, and 5.7 show the average performance at the output of the non-linear beamformers in the anechoic case as a function of the DOA offset. The non-linear beamformers appear to be robust against small DOA offsets.

We now study the effect of different regularisation methods and regularisation parameters on the performance of non-linear beamformers. We only show results for the non-linear beamformer $\mathbf{w}_3$, as results for other non-linear beamformers are similar.

Figure 5.8 show the effect additive diagonal loading on the average performance of the non-linear beamformer $\mathbf{w}_3$ in the anechoic case as a function of the DOA offset. In this regularisation method, we add a small value ϵ to each diagonal element in the correlation matrix [52]. This method penalises large values of beamformer weights and has the general effect of designing a beamformer for a higher white noise level than is actually present [52]. We can see that increasing the value of ϵ reduces the value of SIR significantly without any real benefits on the SDR or robustness against DOA offsets. We believe that the reason behind this is that there is wide range of values for the diagonal terms of the correlation matrices, and therefore it is impossible to pick a single value of ϵ for all Gaussian components or clusters at all frequencies. To avoid this problem, the value added to each diagonal term should be proportional to their values.

Figure 5.9 show the effect multiplicative diagonal loading on the average performance of the non-linear beamformer $\mathbf{w}_3$ in the anechoic case as a function of the DOA offset. In this regularisation method, we multiply each diagonal element in the correlation matrix with $1 + \beta$, where β is a small number [62]. This method is equivalent to additive diagonal loading , albeit the value added to each diagonal term is proportional to its value. We can see that increasing the value of β gives more robustness against DOA offsets, however very large values such as $\beta = 5e - 3$ can reduce the SIR significantly. In comparison with additive diagonal loading, this method gives better performance and robustness against DOA offsets.

Figure 5.10 show the effect eigenvalue thresholding on the average performance of the non-linear beamformer $\mathbf{w}_3$ in the anechoic case as a function of the DOA offset. In this regularisation method, we modify the matrix to ensure no eigenvalue is less than a factor γ times the largest, where $0 < \gamma < 1$ [110]. Specifically, let $\mathbf{QDQ}^{-1}$ denote the eigenvalue decomposition

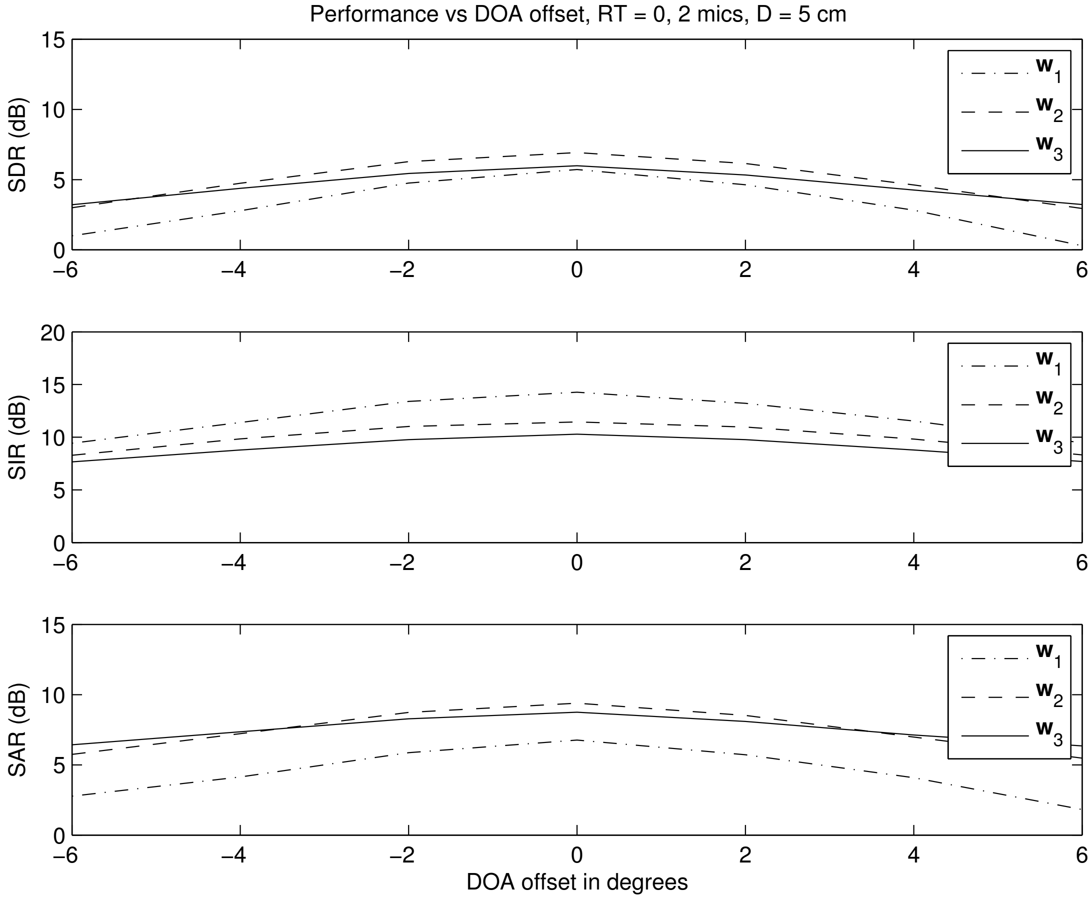

Figure 5.4: *Separation using two microphones: average performance of* $\mathbf{w}_1$*,* $\mathbf{w}_2$*, and* $\mathbf{w}_3$ *in the anechoic case as a function of DOA offset. Used parameters:* $k_s = 2$ *and* $k_v = 15$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$*,* $k_x = 15$ *in* $\mathbf{w}_3$*.*

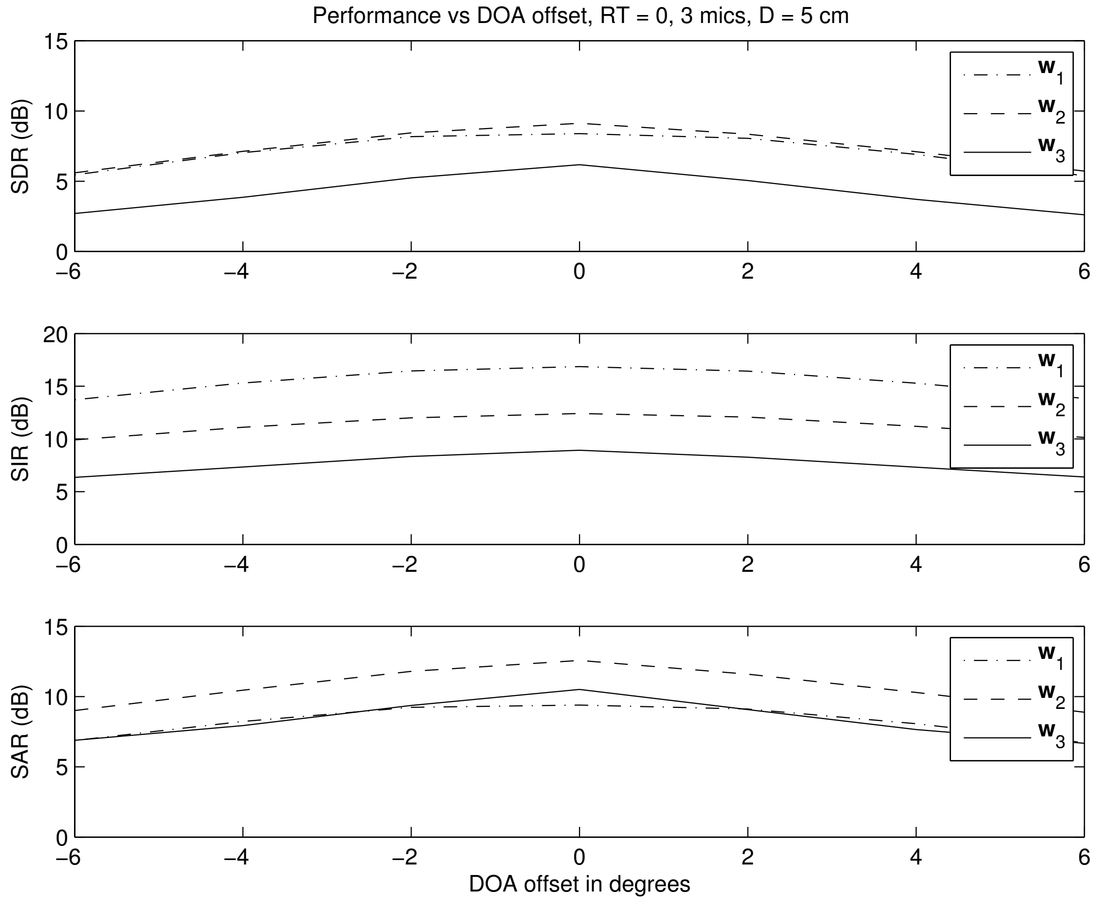

Figure 5.5: *Separation using three microphones: average performance of* $\mathbf{w}_1$, $\mathbf{w}_2$, *and* $\mathbf{w}_3$ *in the anechoic case as a function of DOA offset. Used parameters:* $k_s = 2$ *and* $k_v = 5$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$, $k_x = 5$ *in* $\mathbf{w}_3$.

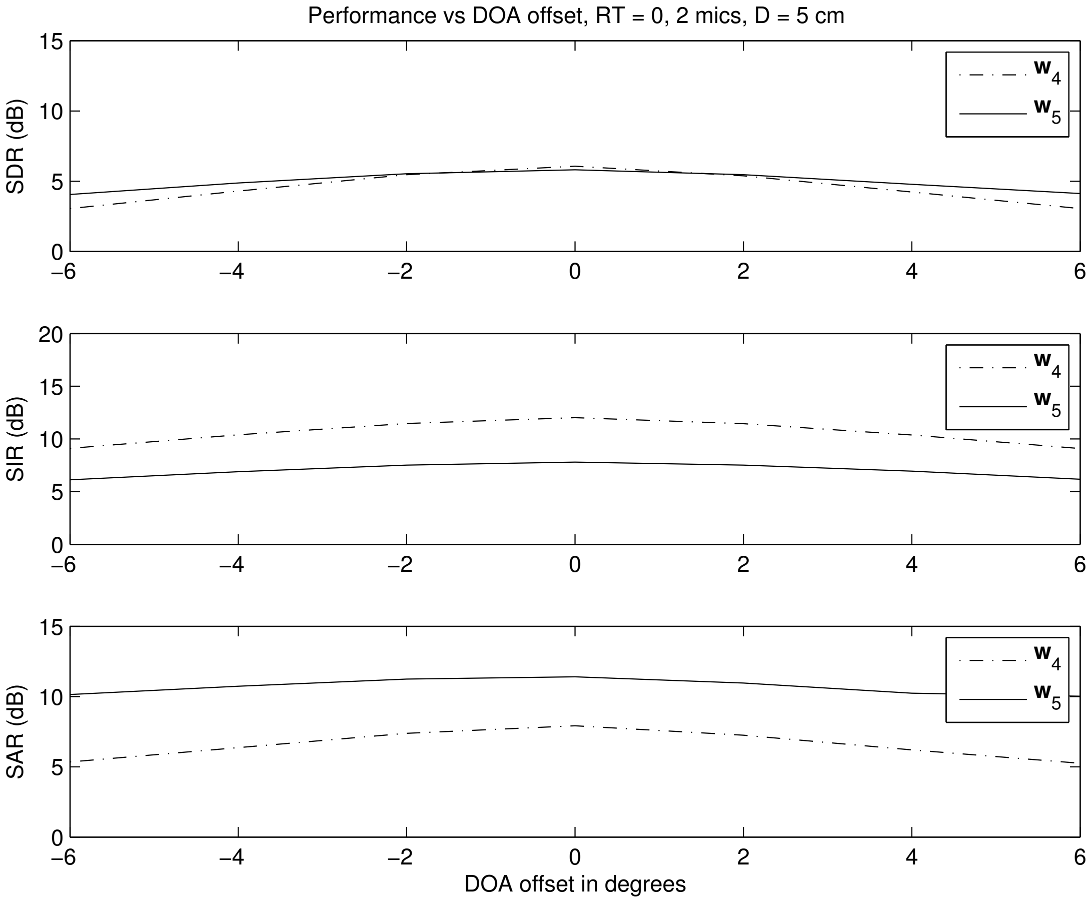

Figure 5.6: *Separation using two microphones: average performance of* $\mathbf{w}_4$ *and* $\mathbf{w}_5$ *in the anechoic case as a function of DOA offset. Used parameter:* $k_x = 15$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

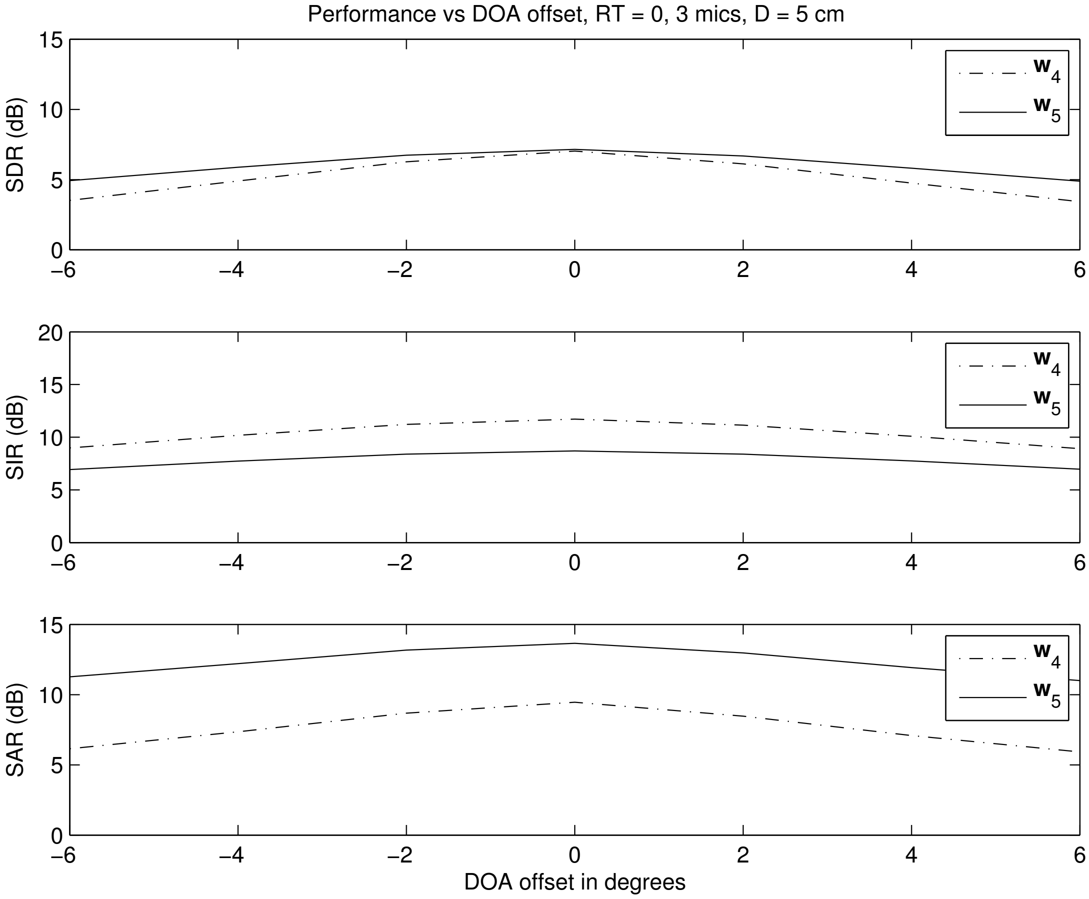

Figure 5.7: *Separation using three microphones: average performance of* $\mathbf{w}_4$ *and* $\mathbf{w}_5$ *in the anechoic case as a function of DOA offset. Used parameter:* $k_x = 5$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

of $\mathbf{R}$, where $\mathbf{D}$ is a diagonal matrix with the matrix eigenvalues on the diagonal:

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix} \tag{5.1}$$

We modify the matrix $\mathbf{D}$ to ensure no eigenvalue is less than $\gamma\lambda_{\text{max}}$, where λ_{max} is the largest eigenvalue in matrix $\mathbf{D}$. The modified matrix $\mathbf{D}$ is computed as follows:

$$\breve{\mathbf{D}} = \begin{bmatrix} \max\{\lambda_1, \gamma\lambda_{\text{max}}\} & & \\ & \ddots & \\ & & \max\{\lambda_N, \gamma\lambda_{\text{max}}\} \end{bmatrix} \tag{5.2}$$

The modified correlation matrix is computed according to:

$$\breve{\mathbf{R}} = \mathbf{Q}\breve{\mathbf{D}}\mathbf{Q}^{-1} \tag{5.3}$$

We can see from Figure 5.10 that increasing the value of γ gives more robustness against DOA offsets, however very large values such as $\gamma = 5e - 3$ can reduce the SIR significantly. In comparison with additive and multiplicative diagonal loading, this method gives better performance and robustness against DOA offsets.

In summary, multiplicative diagonal loading and eigenvalue thresholding methods gives robustness against DOA offsets. However, it is important not to pick too large values for the regularisation parameters in order to avoid a reduction in performance.

5.2.7 Effect of Reverberation

Figure 5.11 shows the average performance of the non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ as a function of the room reverberation time when four sources are operating, and the microphone array used has two microphones with a 5 cm microphone spacing. $k_x = 15$ was used in $\mathbf{w}_3$, and $k_s = 2, k_v = 15$ was used in the two other beamformers. We compared the performance of the three non-linear beamformers with the performance of the MENUET and FMV algorithms.

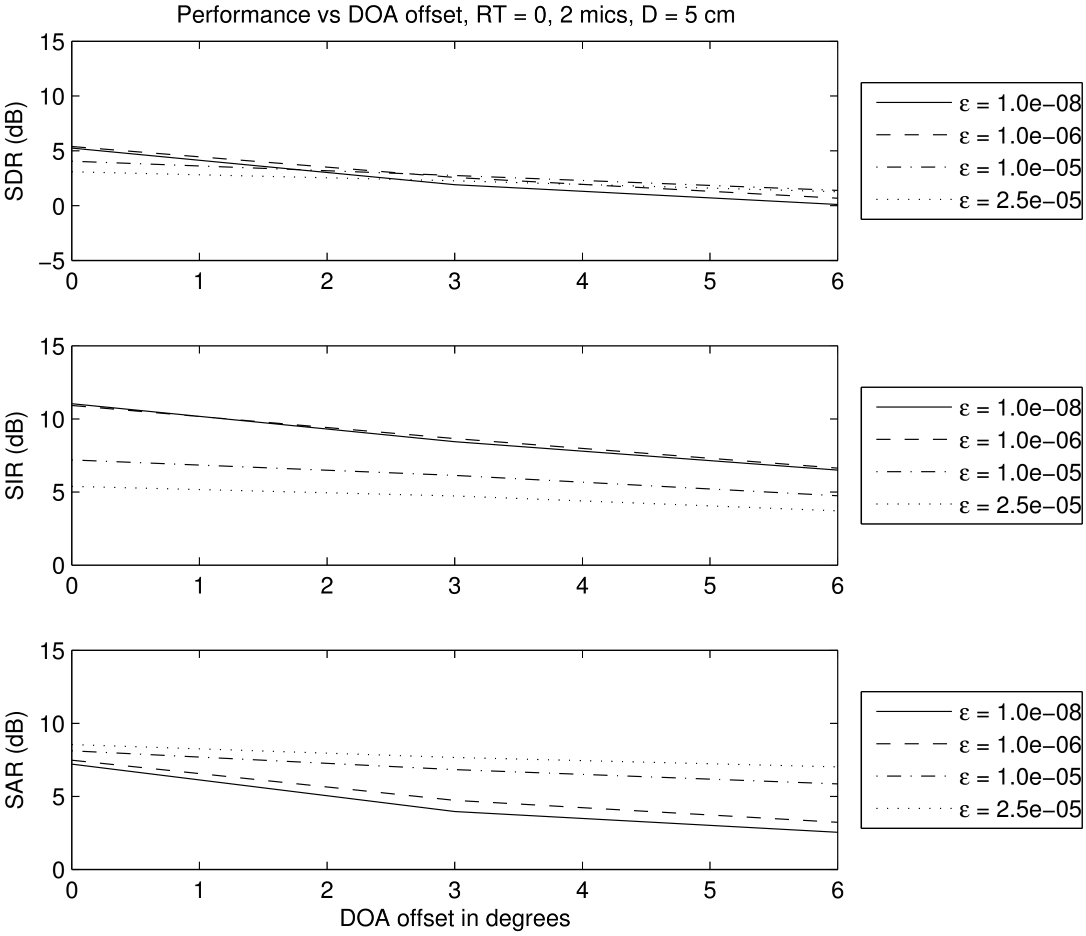

Figure 5.8: *Separation using two microphones: average performance of* $\mathbf{w}_3$ *in the anechoic case as a function of DOA offset when using additive diagonal loading. Used parameter:* $k_x = 15$.

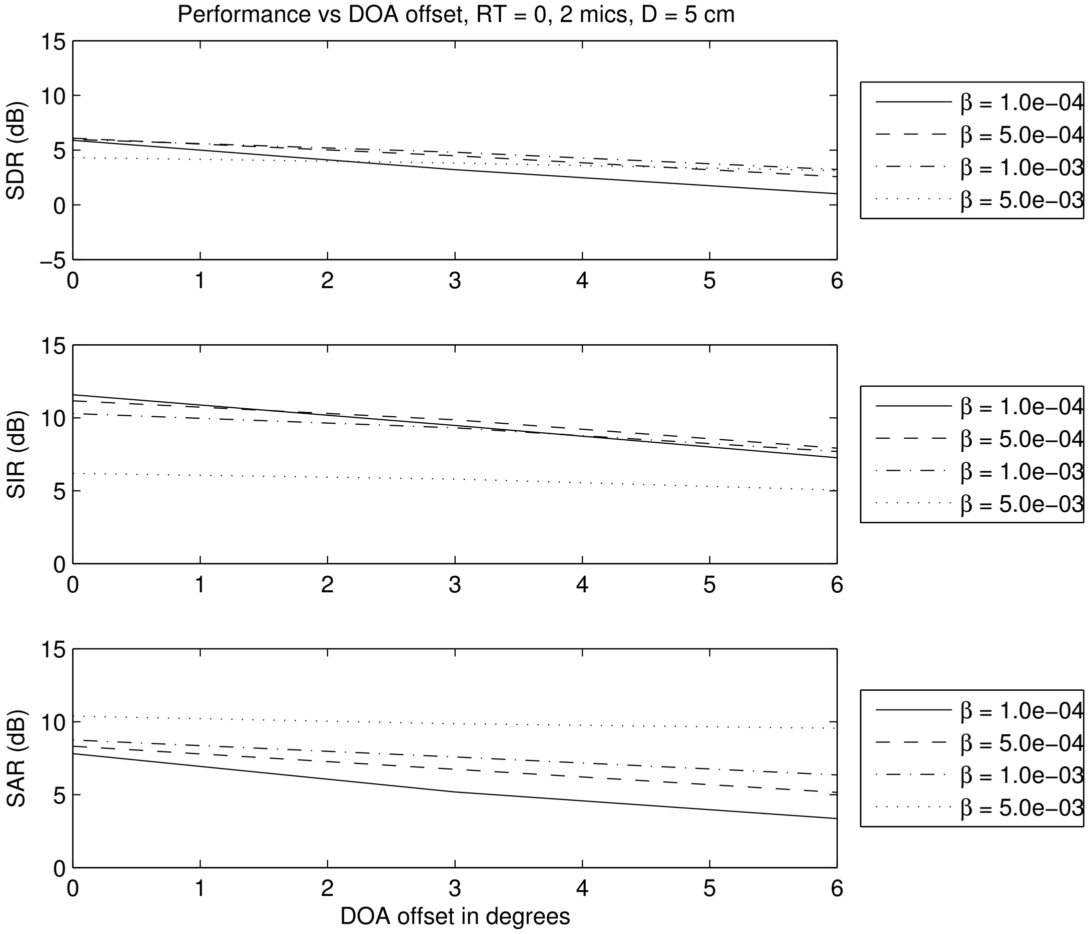

Figure 5.9: *Separation using two microphones: average performance of* $\mathbf{w}_3$ *in the anechoic case as a function of DOA offset when using multiplicative diagonal loading. Used parameter:* $k_x = 15$.

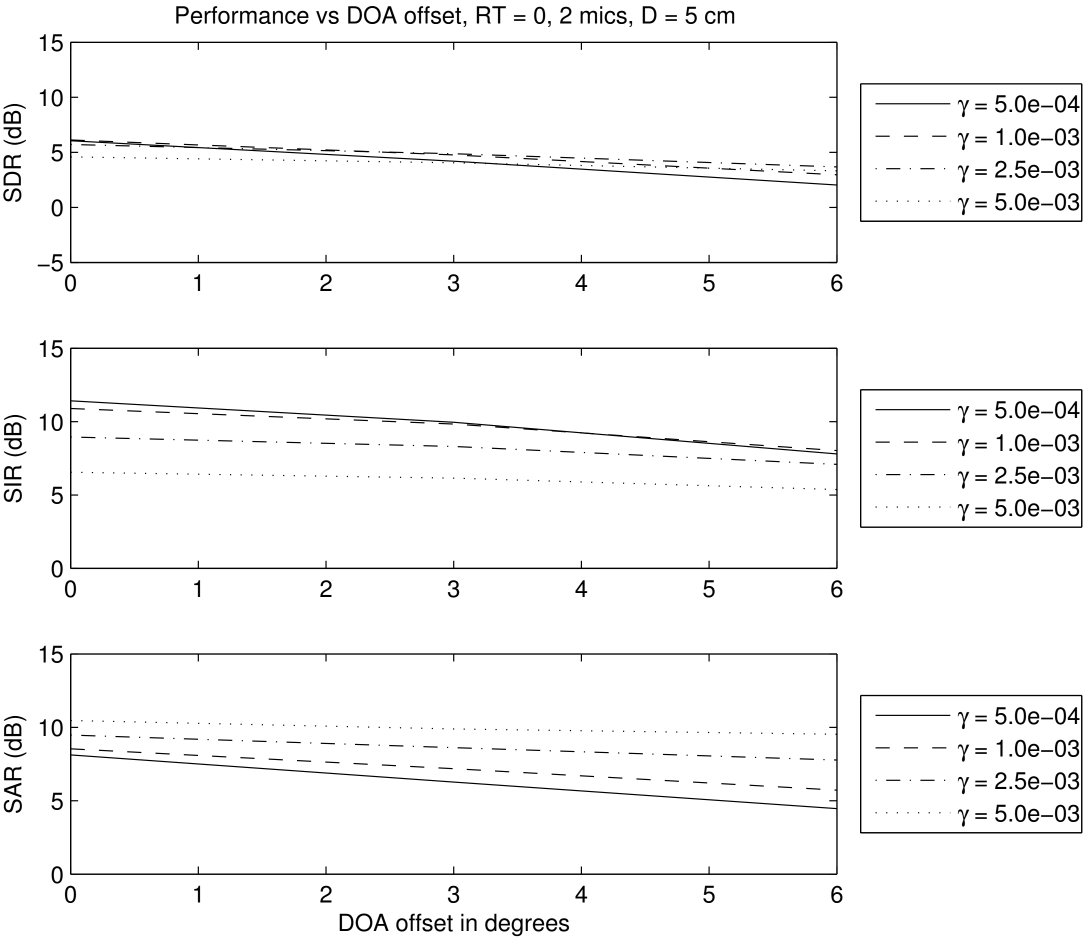

Figure 5.10: *Separation using two microphones: average performance of* $\mathbf{w}_3$ *in the anechoic case as a function of DOA offset when using eigenvalue thresholding. Used parameter:* $k_x = 15$.

In the FMV algorithm, a small step size of 16 samples is essential to obtain good separation performance, while a step size of 256 samples is sufficient in the non-linear beamformers. The MENUET algorithm and the mixture of MMSE beamformers ($\mathbf{w}_1$) gives a high SIR, but suffers from a very low SAR at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ defined in (3.16) and (3.34) respectively have significantly lower artifacts at higher reverberation times.

Figure 5.12 shows the average performance of the non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ as a function of the room reverberation time when four sources are operating, and the microphone array used has three microphones with a 2.5 cm microphone spacing. We compared the performance of the three non-linear beamformers with the performance of the MENUET and FMV algorithms. $k_x = 5$ was used in $\mathbf{w}_3$, and $k_s = 2, k_v = 5$ was used in the two other beamformers. Unsurprisingly, the performance of non-linear beamformers improved with the addition of the third microphone.

Figures 5.13 and 5.14 show the average performance of the non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$ as a function of the room reverberation time. The non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$ offer a better SAR compared to MENUET at the expense of a lower SIR.

5.2.8 Desired Source in Close Proximity to Array

In many applications, such as in human-computer interaction, the desired source is closer to the microphones than the interfering sources, and therefore suffers from less reverberation. In such a situation, the mixing model used in the development of the proposed non-linear beamformers (equation 3.1) is more accurate, as in this mixing model, the desired source signal is assumed to be anechoic, but no assumptions are made about the interferers, which can be reverberant.

In order to illustrate the performance of the non-linear beamformers in such situations, multichannel recordings of several speech sources were simulated in the room illustrated in Figure 5.2. The microphone array has two microphones with spacing $d = 5$ cm. We use the same speech files used in the simulations (five mixtures of male sources, and five mixtures of female sources). The number of the sources in each mixture was four. The desired source was placed 10 cm away from the microphone array at angle $\phi = 10°$, while the interferers were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{-45, -15, 50\}°$. Although the desired source was in close proximity to the microphones, we scaled the power

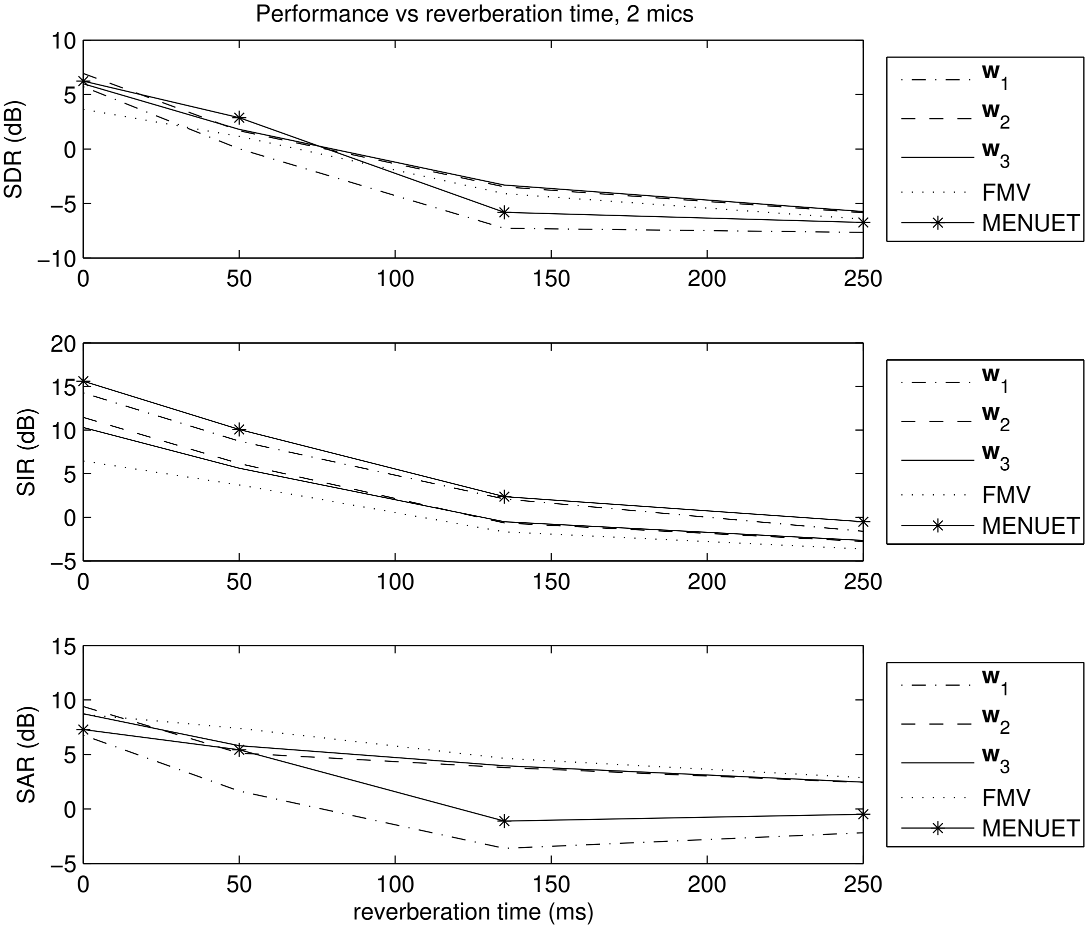

Figure 5.11: *Separation using two microphones: average performance of* $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, *FMV, and MENUET as a function of reverberation time. Used parameters:* $k_s = 2$ *and* $k_v = 15$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$, $k_x = 15$ *in* $\mathbf{w}_3$.

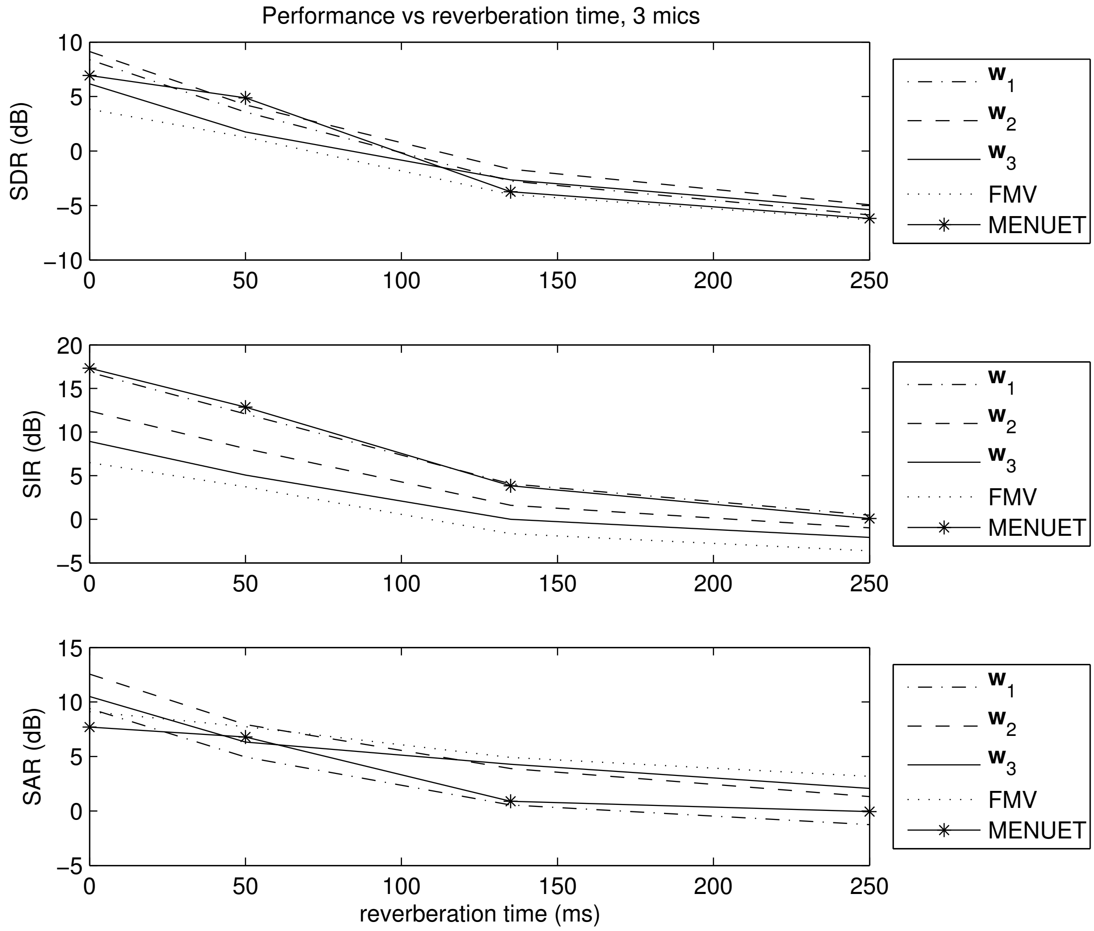

Figure 5.12: *Separation using three microphones: average performance of* $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, *FMV, and MENUET as a function of reverberation time. Used parameters:* $k_s = 2$ *and* $k_v = 5$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$, $k_x = 5$ *in* $\mathbf{w}_3$.

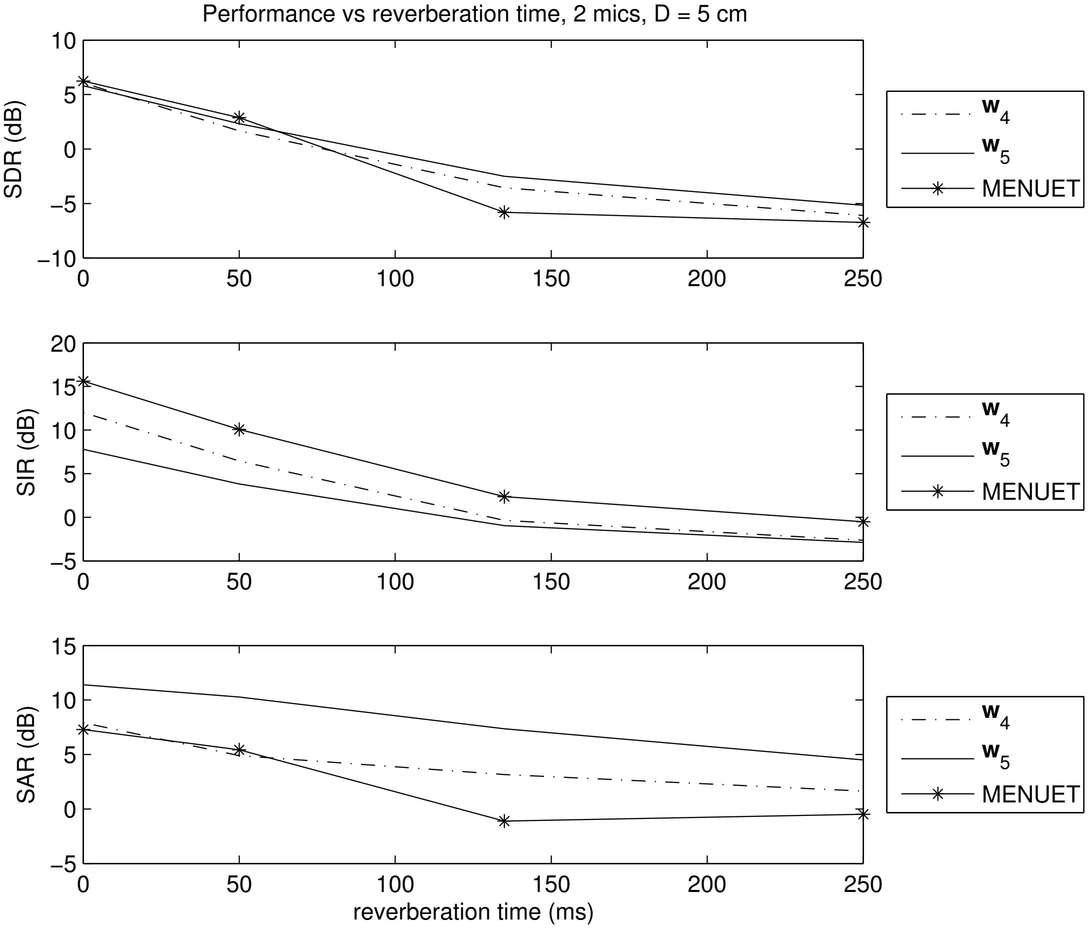

Figure 5.13: *Separation using two microphones: average performance of* $\mathbf{w}_4$*,* $\mathbf{w}_5$*, and MENUET as a function of reverberation time. Used parameter:* $k_x = 15$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$*.*

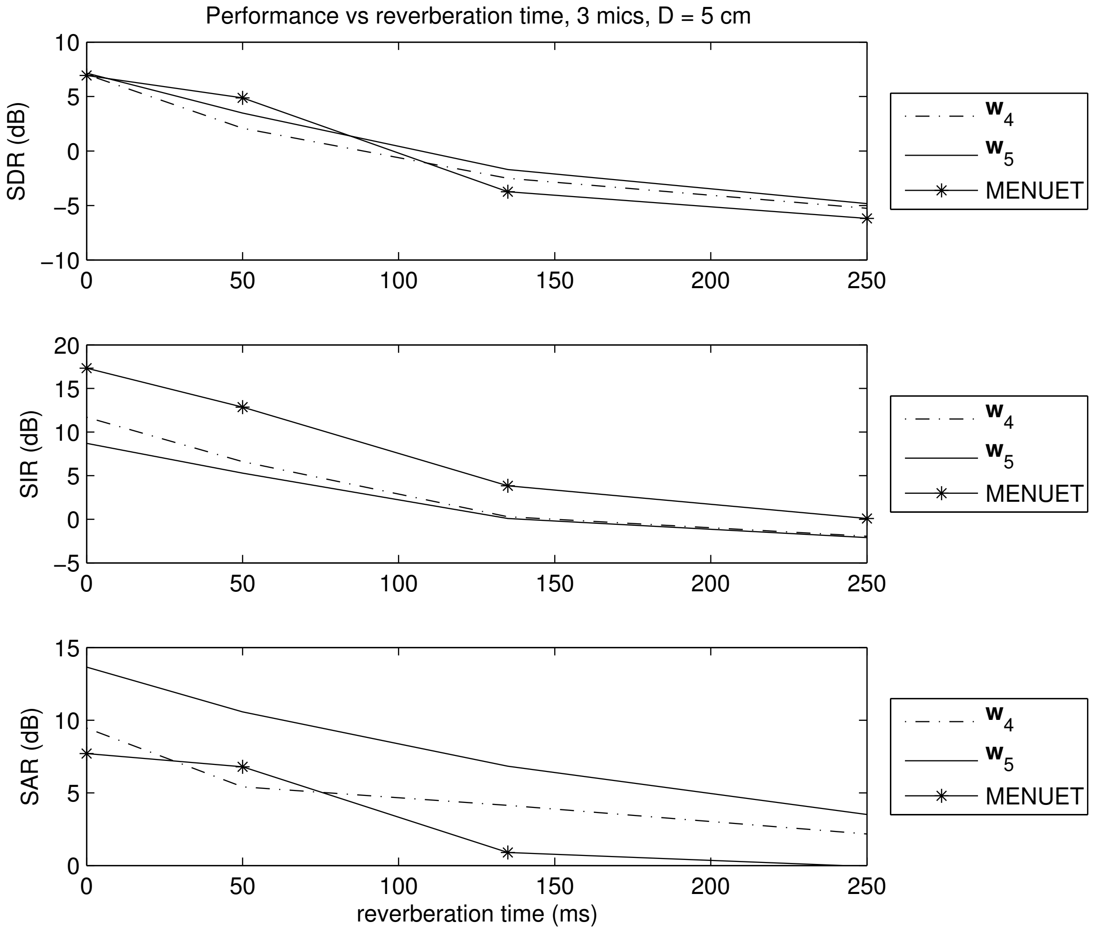

Figure 5.14: *Separation using three microphones: average performance of* $\mathbf{w}_4$*,* $\mathbf{w}_5$*, and MENUET as a function of reverberation time. Used parameter:* $k_x = 5$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$*.*

of the sources so that the received power of all the sources at the microphones are equal. The performance values were averaged over all the ten mixtures. Figures 5.15 and 5.16 show the results. As predicted, the performance of the non-linear beamformers in reverberation when the desired source is in close proximity to the array is better than when the desired source is further away (Figures 5.11 and 5.13). The mixture of MMSE beamformers ($\mathbf{w}_1$) gives the highest SIR and SDR, and it did not suffer from a low SAR at higher reverberation times.

5.2.9 Live Recordings

In order to illustrate the performance of the non-linear beamformers in real life recordings, multichannel recordings of several speech sources were recorded in a room with a reverberation time of 810 ms. The dimensions of the room and the positions of the microphones and the sources are illustrated in Figure 5.3. The microphone array has two microphones with spacing $d = 7$ cm. We use the speech files described in Section 5.2.2 (five mixtures of male sources, and five mixtures of female sources). The number of the sources in each mixture was four. The desired source was placed 30 cm away from the microphone array, while the interferers were placed in a semi-circle of radius 1.5 m around the microphone arrays at angles $\phi = \{-60, -30, 50\}^{\circ}$. We compared the three non-linear beamformers with the FMV and MENUET algorithms. The SDR, SIR and SAR values were averaged over all the mixtures. Table 5.3 shows the results. Due to the high reverberation times, all of the methods suffer from low SIR values, but they all afford SIR improvements over the input mixture (the mixture SIR is –4.7 dB). The non-linear beamformer $\mathbf{w}_1$ has the highest SIR and SDR performance, and also achieves better SAR than MENUET which had the second best SIR. However, when listening to the outputs, it is clear that both MENUET and $\mathbf{w}_1$ suffer from artifacts and musical noise. The distortionless response beamformers $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$, and $\mathbf{w}_5$ have no artifacts in the desired source, but the residual interference signal can be heard.

5.2.10 Time-Frequency Masks

To understand how the various beamformers are achieving their signal enhancement, we can look at equivalent time-frequency masks for each algorithm. Figure 5.17 compares the equivalent mask of the three non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$, and $\mathbf{w}_3$ with the time-frequency mask of MENUET on an example mixture. The equivalent mask was computed at each time-frequency point as the ratio of the energy of the desired signal estimate to the energy of the observed mix-

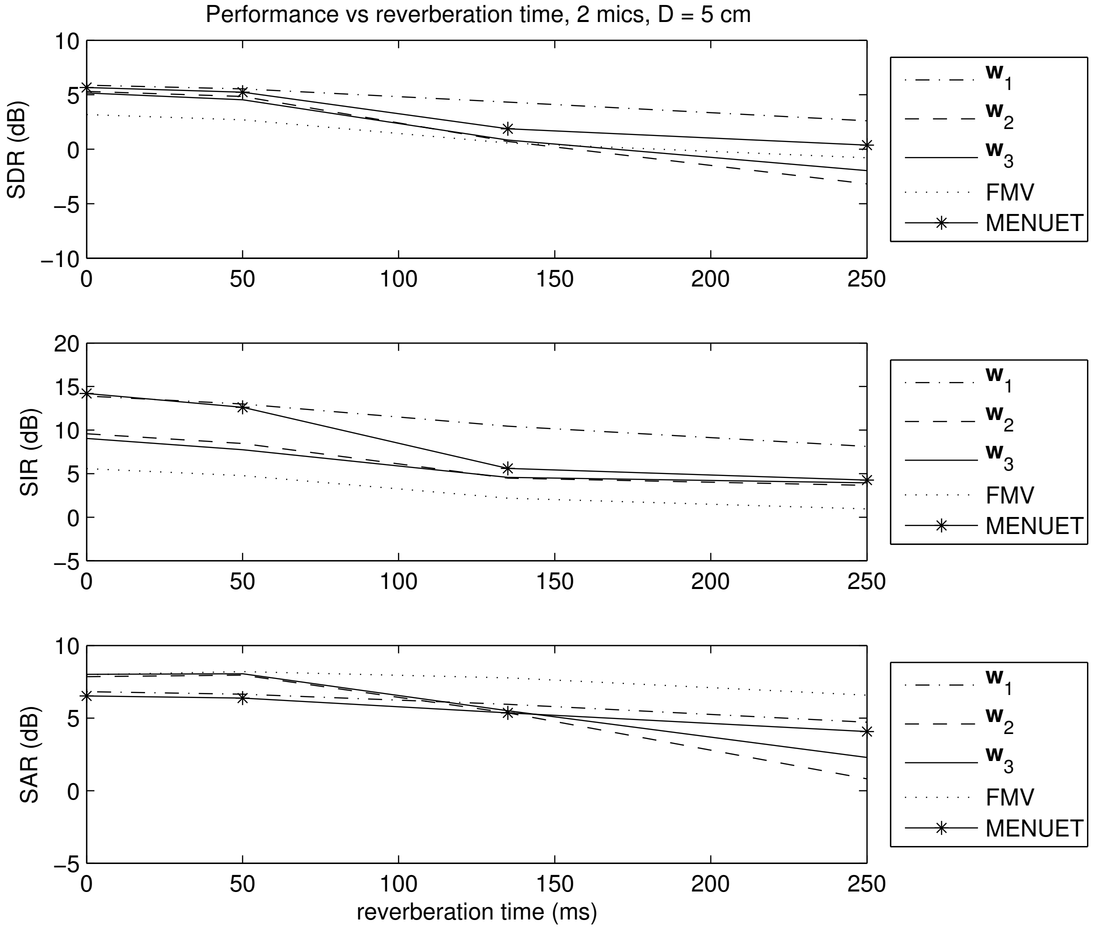

Figure 5.15: *Separation using two microphones: average performance of* $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, *FMV, and MENUET as a function of reverberation time when the desired source is in close proximity to the microphone array. Used parameters:* $k_s = 2$ *and* $k_v = 15$ *in* $\mathbf{w}_1$ *and* $\mathbf{w}_2$, $k_x = 15$ *in* $\mathbf{w}_3$.

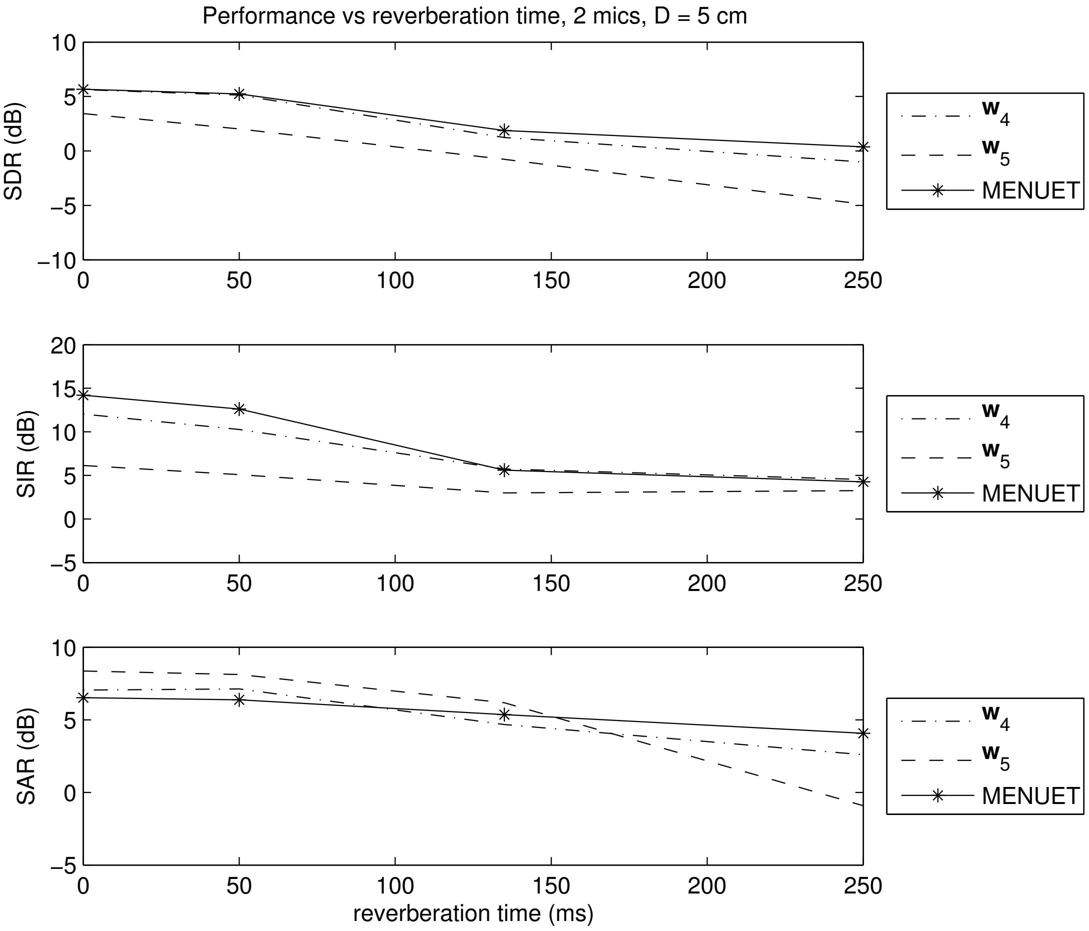

Figure 5.16: *Separation using two microphones: average performance of* $\mathbf{w}_4$*,* $\mathbf{w}_5$*, and MENUET as a function of reverberation time when the desired source is in close proximity to the microphone array. Used parameter:* $k_x = 15$ *in* $\mathbf{w}_4$ *and* $\mathbf{w}_5$.

	SDR	SIR	SAR	SIR gain
$\mathbf{w}_1$	**0.3**	**3.8**	4.5	**8.5**
$\mathbf{w}_2$	-1.7	-0.5	7.9	4.2
$\mathbf{w}_3$	-1.8	-0.3	7.0	4.4
$\mathbf{w}_4$	-0.8	1.6	5.3	6.3
$\mathbf{w}_5$	-1.9	-1.2	**9.8**	3.5
FMV	-2.1	-0.3	5.7	4.4
MENUET	-0.4	3.3	3.6	8

Table 5.3: *Average performance using real life recordings in a room with 810 ms reverberation time. Used parameters: $k_s = 2$ and $k_v = 15$ in $\mathbf{w}_1$ and $\mathbf{w}_2$, $k_x = 15$ in $\mathbf{w}_3$, $\mathbf{w}_4$, and $\mathbf{w}_5$. Best Performance values is in bold font.*

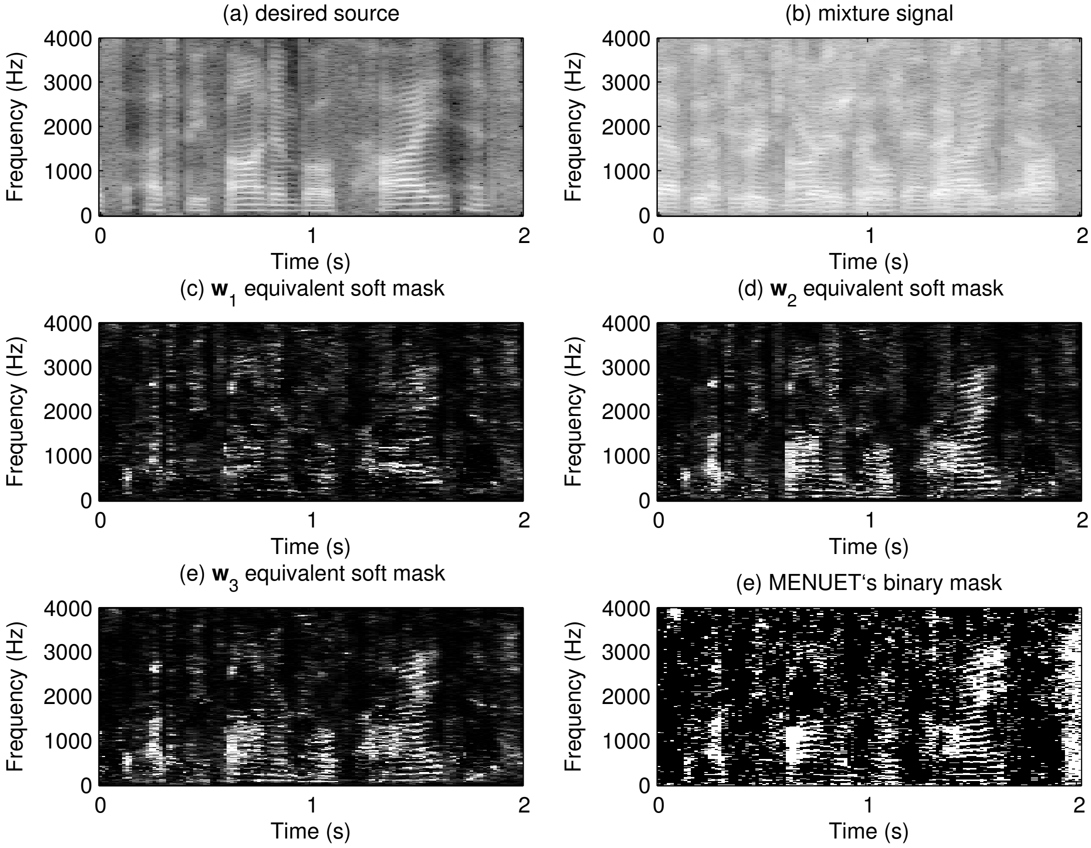

Figure 5.17: *(a) Spectrogram of a desired signal. (b) Spectrogram of a mixture of 4 sources. (c-f) t-f masks.*

ture. The non-linear beamformer's approach results in a soft decision mask for the observed signal.

5.3 Discussion

In this chapter, we first evaluated the performance of the non-linear beamformers when the direction of the arrival of the desired source is different from the presumed one. Regularisation methods used in beamforming techniques can be easily incorporated in the mixture of beamformers framework. It was shown that multiplicative diagonal loading and eigenvalue thresholding methods gives robustness against small DOA offsets. However, it is important not to pick large values for the regularisation parameters in order to avoid a reduction in performance.

In the second test, the non-linear beamformers were evaluated on underdetermined speech mixtures with room reverberation. It was shown that the mixture of MMSE beamformers $\mathbf{w}_1$ gives better interference rejection at the expense of higher artifacts. The non-linear beamformers $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$, and $\mathbf{w}_5$ are distortionless beamformers (constant gain in the look-direction), and have significantly lower artifacts at higher reverberation times. When the desired source is in close proximity the microphone array, the non-linear beamformer $\mathbf{w}_1$ did not suffer from higher artifacts at high reverberation time (as shown in Section 5.2.8). In real life recordings with a reverberation time of 810 ms and the desired source in close proximity to the array, the non-linear beamformer $\mathbf{w}_1$ has the highest SIR and SDR performance at the expense of audible artifacts. The distortionless response beamformers $\mathbf{w}_2$, $\mathbf{w}_3$, $\mathbf{w}_4$, and $\mathbf{w}_5$ have no artifacts in the desired source, but the residual interference signal can be heard.

Chapter 6
Conclusions and Future Work

6.1 Summary and Conclusions

The goal of this thesis was the investigation and development of new methods for the separation and extraction of audio signals, and more specifically of speech signals. Our proposed methods focused on the extraction of a desired source with known location from underdetermined mixtures, where the number of sources is larger than the number of microphones.

In this thesis, we first discussed the different possibilities of source separation and extraction environments and prior knowledge of the sources, mixing process, and the microphones. In general, source separation and extraction is more difficult when the room impulse response of the room is long, when the sources are moving, or when the number of sources is larger than the number of microphones. Some source separation and extraction methods requires some a priori information such as the microphone geometry, source geometry, number of sources, type of sources (speech, music...etc) and channel parameters. The availability of this a priori information depends on the application.

We then reviewed some basic properties of speech signals, and discussed their relevance to the speech separation and extraction problem. Regularities in the speech signals such as common onsets, common offsets, harmonicity, and temporal coherence are exploited by many methods, especially methods that emulate the way the human auditory system perform speech separation and extraction (CASA methods). Statistical properties of speech such as non-gaussianity, non-stationarity, non-whiteness, and sparseness are also exploited in many speech separation and extraction methods. Furthermore, the use of microphone arrays gives one the opportunity to exploit the fact that the sources originate at different points in space.

We then discussed performance measures that can be used to evaluate source separation and extraction methods. Many performance measures have been proposed. A very important factor to take into account when choosing a suitable performance evaluation measure is whether the output of the source separation/extraction method is a signal that is intended to be listened to or not. In many applications, the extracted sources are meant to be listened to straight after separation. In these applications, we require high quality source estimates. In other applications, the extracted sources are processed to obtain information, such as recognising the words pronounced by a speaker using a computer, or identifying the identity of a speaker or a musical instrument. In such applications, the effectiveness of source separation and extraction can be judged by the performance of the final application.

We then reviewed current solutions to the speech separation and extraction problem. Source separation methods initially only considered the case when the number of sources is equal to or less than the number of microphones (determined and overdetermined mixtures). ICA is one of the major methods used in this case. ICA is a blind method; it does not need additional a priori information, such as the array geometry or the positions of the desired and interfering sources. However, ICA methods require that the number of sources to be known a priori. In ICA, the mixing filters (or the separating filters) are determined so that the estimated source signals are as independent as possible. The separating filters act as null beamformers that attenuate the interfering signals. For this reason, standard ICA cannot deal with underdetermined mixtures.

Adaptive beamforming techniques have also been used successfully in source separation and extraction for determined and overdetermined mixtures. Unlike ICA, beamforming techniques require information about the microphone array configuration and the sources (for example, the direction of the desired source). However, beamforming techniques can attenuate spatially spread and reverberant interferers, and there is no need to determine their number.

In underdetermined mixtures, the assumption that the sources have a sparse representation under an appropriate transform is a very popular assumption. One popular approach to sparsity-based separation is time-frequency masking, where we process the mixture signal a time-frequency mask that attenuate interfering signals while preserving time-frequency points where the signal of interest is dominant. If more than one mixture signal is available, the spatial information at each time-frequency point can sometimes be used to determine which time-frequency points belong to each source. To estimate the masks, many methods assume that the speech source do not overlap in the time-frequency domain, and partition the mixture time-frequency coefficients based on the inter-channel level/phase difference. In practice, this disjointness assumption is not fully met and becomes less realistic in reverberation. Furthermore, time-frequency masking suffers from musical noise due to masking of time-frequency points where the sources overlap.

In order to combine the benefits of non-linear time-varying separation in time-frequency masking with the benefits of spatial filtering in the linear beamformers, we proposed to exploit the speech's sparsity in the time-frequency domain in order to extend the use of beamforming techniques to underdetermined speech mixtures. In Chapter 3, we used GMMs to model the speech non-Gaussianity and the spatial distribution of the sources. We presented three frequency domain non-linear beamformers that can extract a desired source from a known-direction. The first two non-linear beamformers are based on modeling the desired source signal and the interference separately. The desired source signal is modeled using a 1-dimensional GMM, and the observed interference is modeled using an N-dimensional GMM, where N is the number of microphones. The covariance matrices of each Gaussian state represent a spatial covariance matrix. The signal estimator in these beamformers comprises of a set of MMSE beamformers (termed $\mathbf{w}_1$) or MVDR beamformers (termed $\mathbf{w}_2$). The third non-linear beamformer is based on modeling the observed mixture signal (the desired source and interference together) using an N-dimensional GMM. In this case, the signal estimator comprises a set of MPDR beamformers (termed $\mathbf{w}_3$). The model learning is performed with an EM algorithm using the observed mixture signals only, and no prior training is required. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. These weights are specific to each time-frequency point. This results in the non-linear beamformer dynamically finding suitable directivity patterns in order to reduce active interfering signals. This allows the non-linear beamformer to deal with underdetermined mixtures.

An important feature of the proposed extraction system is that no assumptions on the number, location or size of the interferers were taken, and the interferers can be of any nature such as point sources, spatial extended sources, diffuse sources, or a combination of them. The proposed methods can be applied to microphone arrays with two or more microphones.

The non-linear beamformers have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_1$ gives better interference rejection at the expense of higher artifacts, especially at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ are distortionless beamformers (constant gain in the look-direction), and have significantly lower artifacts at higher reverberation times. When the desired source is in close proximity the microphone array, the non-linear beamformer $\mathbf{w}_1$ did not suffer from higher artifacts at high reverberation times.

In terms of computational complexity, non-linear beamformer $\mathbf{w}_3$ employs the simplest learning algorithm and requires fewer iterations than non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$. Furthermore, the model learning for non-linear beamformer $\mathbf{w}_3$ is independent of the DOA of the desired source, which makes this non-linear beamformer suitable in applications where scanning for the source direction is needed. However, it is clear that the three non-linear beamformers are computationally expensive and are not suitable for many real-time applications.

In Chapter 4, we presented a modification to the mixture of MPDR beamformers ($\mathbf{w}_3$). We suggested using simple clustering algorithms used in many popular time-frequency masking algorithms instead of the GMM model and the EM algorithm. The proposed algorithm has two main stages. In the first stage, the mixture time-frequency points are partitioned into a sufficient number of clusters using time-frequency masking techniques. In the second stage, we use the clusters obtained in the first stage to calculate covariance matrices, one for each cluster in each frequency bin. These covariance matrices and the time-frequency masks are then used in the mixture of MPDR beamformers. The resulting non-linear beamformer has low computational complexity and removes the musical noise found in time-frequency masked outputs at the expense of lower interference attenuation. The mixture of MPDR beamformers stage can be regarded as a post-processing step for sources separated by time-frequency masking. Two variants of the proposed method were described and compared. The first one uses binary time-frequency masks, and in this case the mixture of beamformers is termed $\mathbf{w}_4$. The second variant uses soft (real-valued) time-frequency masks, and in this case the mixture of beamformers is termed $\mathbf{w}_5$.

The non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$ have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_4$ gives better interference rejection, while the non-linear beamformer $\mathbf{w}_5$ gives better SAR. In terms of computational complexity, both non-linear beamformers $\mathbf{w}_4$ and $\mathbf{w}_5$ have similar computational complexity, and require significantly less computational time than non-linear beamformers $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$.

6.2 Suggestions for Future Work

There is still a large room for extensions to the algorithms presented. Research in these extensions might enhance the performance of the proposed methods, and increase their usefulness in practical applications.

Online Model Learning

In our current implementation, the EM algorithm used in $\mathbf{w}_1$, $\mathbf{w}_2$ and $\mathbf{w}_3$, and the k-means clustering algorithm used in $\mathbf{w}_4$ and $\mathbf{w}_5$ are in a batch learning mode. In sections 3.5.4.3 and 4.3.2.3, we studied the effect of using short blocks of data. The batch mode with short blocks of data can be used in applications where short delays are permissible, such as in human-computer interaction or surveillance. However, it is not appropriate for real-time applications. In these applications, online model learning is useful [114, 121]. The online model learning should have a forgetting factor, and a mechanism for adding, deleting, and reassigning clusters to handle changes in the environment, such as the number of active sources and their location [115].

Subband Filtering

In order to perform complete source separation in reverberant rooms, it is necessary to remove cross-frame interference, which arises from the use of time-frequency representations that use an analysis window length (frame length) shorter than the reverberation time. In order to remove the cross-frame interference, we can use unmixing filters in each subband instead of instantaneous separation (multipliers). The length of the subband filters should be chosen carefully; a short filter will fail to completely remove the cross-frame interference, while a filter that is too long degrades separation performance because the number of data samples per filter coefficient decreases, which results in inaccuracies in filter coefficient estimation. Furthermore, room reverberations typically last longer at low frequencies than they do at high frequencies. Therefore, it should be advantageous to use longer separating filters in the low frequency bands [122].

Auditory Filterbanks

Most speech source separation and extraction algorithms rely on the transformation of time-domain mixture signals received at the microphone array into the time-frequency domain to improve the disjointness of received signals. Most researchers transform the mixture signals using the STFT for computational efficiency. The STFT provides an equal frequency resolution for all frequencies. However, speech signals concentrate most of their energy at low frequencies and overlapping of source signals is more probable to be present in this frequency region. We can improve the frequency resolution by increasing the STFT window size, but this leads to a reduction in the temporal resolution and thus more overlap. Auditory filter banks provide a high frequency resolution for low frequencies. It was shown in [123] that the use of modulated Hann windows with centre frequencies spaced in the equivalent rectangular bandwidth (ERB) [124] and the Bark [125] auditory scales improve the disjointness of speech mixtures. Therefore, it is worth investigating the use of auditory representations in the mixture of beamformers framework instead of the STFT.

Additional Response Constraints

The MVDR and MPDR beamformers are designed by linearly constraining the beamformer weights to have a constant gain in the look-direction. It is possible to develop beamformers in which additional linear constraints are imposed. These beamformers are referred to as linear constrained minimum variance (LCMV) beamformers [52]. It would be worth investigating the use of multiple linear constraints in the mixture of beamformers framework for added robustness control over the directivity patterns. For example, robustness to DOA offsets can be enhanced by forcing a flatter directivity patterns near the signal direction. This can be done using additional directional or derivative constraints [52]. If the desired source is in close proximity to the microphone array, near field regional constraints can be incorporated to provide distance discrimination [126].

Video Assisted Source Tracking

In many applications such as human-computer interaction and smart meeting rooms, it would be advantageous to use video data captured by a camera to estimate the location of the desired sources[25, 26].

Appendix A
Publications

Several methods and results presented in this thesis have been published in the following works:

Journal Article:

- M.A. Dmour, M.E. Davies; "A new framework for underdetermined speech extraction using mixture of beamformers", to appear in IEEE Transactions on Audio, Speech and Language Processing.

Conferences:

- M.A. Dmour, M.E. Davies; "An approach to under-determined speech separation based on a non-linear mixture of beamformers", European Conference on Signal Processing (EUSIPCO), 2009.
- M.A. Dmour, M.E. Davies; "Under-determined speech separation using GMM-based non-linear beamforming", European Conference on Signal Processing (EUSIPCO), 2008.

UNDER-DETERMINED SPEECH SEPARATION USING GMM-BASED NON-LINEAR BEAMFORMING

Mohammad A. Dmour and Michael E. Davies

Institute for Digital Communications and Joint Research Institute for Signal and Image Processing,
University of Edinburgh, Edinburgh, EH9 3JL, UK
{M.Dmour, Mike.Davies}@ed.ac.uk

ABSTRACT

This paper introduces a frequency-domain non-linear beamformer that can perform speech source separation of under-determined mixtures, is reasonably artifact-free and does not require prior knowledge of the number of speakers. This beamformer utilises a Gaussian mixture distribution to model the observation probability density in each frequency bin, which can be learnt using the expectation maximisation (EM) algorithm. A linear minimum-variance distortionless response (MVDR) beamformer is determined for each of the Gaussian components. The proposed non-linear beamformer is then a weighted sum of these linear MVDR beamformers and is therefore also distortionless. The relative contribution for each linear MVDR beamformer is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. Simulation results of the non-linear beamformer in under-determined mixtures with room reverberation confirm its ability to successfully separate speech sources with virtually no artifacts.

1. INTRODUCTION

Speech separation is the problem of extracting a target speech signal from observations corrupted by interfering signals such as other speech signals and background noise. Speech separation is used in a wide range of applications, such as hearing aids, human-computer interaction, surveillance, and hands-free telephony. In general, observations are obtained at the output of a set of microphones, each receiving different combinations of the source signals. The use of microphone arrays gives one the opportunity to exploit the fact that the desired source and the interfering sources originate at different points in space. The difficulty of the speech separation task depends on the way in which the signals are mixed within the acoustic environment. Speech separation is more difficult when the reverberation time of the acoustic environment is large, and when there are fewer microphones than sources.

Suppose that M source signals are mixed and observed at N microphones. The signal at microphone j can be modeled as:

$$x_j(t) = \sum_{i=1}^{M} \sum_{p=0}^{P-1} a_{ji}(p) s_i(t-p) \qquad (1)$$

where a_{ji} represents the impulse response from source i to microphone j, and P is the length of the impulse response between each source-microphone pair. A mixture is termed a determined mixture 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 when it is smaller.

One approach to speech separation is to use statistical modeling of source signals. Independent component analysis (ICA) is one of the major statistical tools for solving the problem of speech separation. In ICA, separation is performed using the assumption that the source signals are statistically independent with no information on the direction of arrival of source signals, or microphone array configuration. To perform source separation, we process the mixture channels by a set of time-invariant demixing filters and sum the filtered channels together. ICA implicitly estimates the source directions by maximising the independence of the sources, and acts as an adaptive null beamformer that reduces the undesired sources.

However, some aspects limit the application of ICA in real-world environments. Most ICA methods assume the number of sources is given a priori. In general, classical ICA techniques cannot perform source separation when spatially spread sources are involved, or in the under-determined mixtures case.

Another approach to speech separation is to use adaptive beamforming techniques. In adaptive beamforming, the microphone array is used to form a spatial filter which can extract a signal from a specific direction and reduce signals from other directions. For example, in minimum-variance distortionless response (MVDR) beamforming, the beamformer response is constrained so that signals from the direction of interest are passed with no distortion, while it suppresses noise and interference at the output of an array of microphones. In [2, 3], beamforming weights were calculated using time-domain recursive algorithms. It was shown recently in [4] that a frequency-domain MVDR (FMV) beamformer which performs sample matrix inversion using statistics estimated from a short sample support gives better performance than time-domain recursive algorithms in non-stationary acoustic environments. Compared to ICA, adaptive beamforming can utilise the available information about source signals and the microphone array configuration. In addition, there is no need to model the source signals or determine their number. Adaptive beamforming can attain excellent separation performance in determined or over-determined time-invariant mixtures involving point sources. However, when spatially spread sources are involved, or in under-determined mixtures, perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible. This, in turn, leads to performance degradation.

In the under-determined mixing case, the assumption of spatial diversity is insufficient to perform source separation, thereby necessitating additional assumptions. One increasingly popular and very useful assumption is that the sources have a sparse representation in a given basis. The advantage of sparse signal representation is that the probability of more than one active source is low. A sparse representation of a speech signal can be achieved by a short term Fourier transform (STFT). One popular approach to perform under-determined speech separation is time-frequency (t-f) masking. This approach is a special case of non-linear time-varying filtering that estimates the desired source signal by:

$$\hat{s}(n,f) = M(n,f)x_j(n,f) \qquad (2)$$

where $M(n,f)$ is a t-f mask containing positive gains which must be adapted to extract the desired source from the observed mixtures. A popular method used to perform speech separation of under-determined mixtures using only two microphones is the degenerate unmixing estimation technique (DUET) [8, 5]. In DUET, binary masks are determined from the spatial location information contained in the STFT coefficients of the mixture channels. DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. However, this method suffers from the so-called musical noise or burbling artifacts due to binary masking of t-f points where the sources overlap.

In this paper, we introduce a frequency-domain non-linear beamformer that can perform speech separation of under-determined mixtures and is distortionless. This beamformer utilises Gaussian mixture models (GMMs) to model the observation probability density in each frequency bin. This in turn can be learnt using the expectation maximisation (EM) algorithm. The signal estimator comprises of a set of MVDR beamformers, one for each component of the GMM. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM components. This approach results in a "soft decision" filter for the observed signal. The resulting non-linear beamformer has low computational costs, and does not need to know or estimate the number of sources. It combines the benefits of non-linear time-varying separation in t-f masking with the benefits of spatial filtering and distortionless response in the linear MVDR beamformer.

The organisation of this paper is as follows. Section 2 reviews the linear minimum mean square error (MMSE) beamformer, and then introduces the GMM-based non-linear beamformer. In Section 3, the EM algorithm is used to learn the GMM parameters. The experimental conditions and simulation results are presented in Section 4, followed by a discussion in Section 5.

2. OPTIMUM BEAMFORMERS

Consider a narrow band array signal $\mathbf{x} = [x_1, ..., x_N]^T$ that consists of the desired signal arriving at the array from a known direction, and an interference-plus-noise signal. That is,

$$\mathbf{x} = s\mathbf{e} + \mathbf{v} \tag{3}$$

where $\mathbf{e}$ is the known $N \times 1$ array response vector in the direction of the desired source signal (the array manifold), and $\mathbf{v}$ is the $N \times 1$ complex vector of interference-plus-noise snapshots. We assume that the signal and interference-plus-noise snapshots are uncorrelated. The interference has spatial correlation according to the angles of the contributing interferers. The ultimate goal is to combine the received signals in such as way that the interference-plus-noise signal is reduced while the desired signal is preserved.

2.1 Linear MMSE beamformer

We first consider the optimum estimator whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference and noise, assuming known desired signal direction. We assume that the desired source signal is a sample function from a zero-mean complex-valued Gaussian random process, $s \sim \mathbb{N}(0, \sigma_s^2)$. We also assume a zero-mean complex-valued Gaussian interference-plus-noise, $\mathbf{v} \sim \mathbb{N}(0, R_v)$. Additionally, it is assumed that the signal and interference-plus-noise snapshots are uncorrelated. Hence, $\mathbf{x} \sim \mathbb{N}(0, R_v + \sigma_s^2 \mathbf{e}\mathbf{e}^H)$, and $\mathbf{x}|s \sim \mathbb{N}(s\mathbf{e}, R_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}_{MMSE} = E\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x}).s\,ds \tag{4}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [6]:

$$E\left[s|\mathbf{x}\right] = \frac{\mathbf{e}^H R_v^{-1}\mathbf{x}}{\mathbf{e}^H R_v^{-1}\mathbf{e}} . \frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{e}^H R_v^{-1}\mathbf{e}\right)^{-1}} \tag{5}$$

The first term is an MVDR spatial filter, which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction. The second term is a single-channel Wiener post-filter. We see that the MMSE estimator is just a shrinkage of the MVDR beamformer. Unfortunately, the MMSE beamformer depends explicitly on σ_s^2 which is typically unknown. Therefore, we cannot implement the MMSE beamformer in practice. However, we can obtain a beamformer that does not depend on σ_s^2 by assuming a distortionless response in the specified direction. The result is the MVDR beamformer. However, since we have a distortionless response, we cannot exploit the sparsity of the desired source signal. The MVDR beamforming process can be written as:

$$\begin{aligned} \hat{s} &= \mathbf{w}^H \mathbf{x} \\ &= \frac{\mathbf{e}^H R_v^{-1}}{\mathbf{e}^H R_v^{-1}\mathbf{e}}\mathbf{x} \end{aligned} \tag{6}$$

In practice, the desired signal may either be present all the time, or it is difficult to estimate its activity periods. As a result of this, the estimation of the signal-free interference-plus-noise covariance matrix R_v is not possible. It can be shown, however, that if there is no mismatch between the vector $\mathbf{e}$ used in the MVDR beamformer and the true array manifold, then the estimator which uses the observed signal covariance matrix R_x is identical to the estimator which uses the signal-free interference-plus-noise covariance matrix R_v [6].

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear or the signals are Gaussian. Speech sources are generally non-stationary and non-Gaussian. This suggests extending the optimum beamformers to exploit the non-stationarity and non-Gaussianity of speech signals.

2.2 Frequency-domain MVDR (FMV) beamformer

Speech is a non-stationary process, but over short durations speech signals can be considered stationary. In the FMV algorithm [4], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could change over longer time spans. In the FMV algorithm [4], frequency-domain signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the 32 most recent STFT values . MVDR weights are then calculated using the correlation matrix. Therefore, in the FMV algorithm, new beamformer weights are calculated every small time interval in order to reduce the contribution to the extracted signal of interfering sources active during that time interval, while having a distortionless response in the desired source DOA. Only statistics gathered over a very short period of time are used in the calculation of weights.

The quick adaptation of the beamformer weights can substantially reduce a large number of non-stationary interferences while utilising few microphones [4]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. On the other hand, the spatial distribution of the sources does not change very quickly, and we can gather statistics for the desired signal estimator over a longer time span. Thus the FMV beamformer is forced to compromise between long intervals (good statistics) and short interval (rapid response).

2.3 GMM-based non-linear beamformer

In the frequency-domain, speech signals have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some t-f points, not all speech sources in a mixture are active at the same t-f points. It is therefore advantageous to exploit the sparsity property of speech signals in the frequency-domain in order to perform separation in under-determined environments. In order to model the speech non-Gaussianity, we propose to apply

GMMs, which are widely used for modeling highly complex probability densities.

In this section, we use a Gaussian mixture interference-plus-noise model and find the optimum estimator whose output is the MMSE estimate of the desired signal s assuming a known desired signal direction. We shall describe the density of the interference-plus-noise signal $\mathbf{v}$ as a mixture of k zero-mean Gaussians $q = 1,...,k$ with covariances $R_{v,q}$ and mixing proportions c_q:

$$p(\mathbf{v}|\theta) = \sum_{q=1}^{k} c_q \frac{1}{\pi^N |R_{v,q}|} \exp\{-\mathbf{v}^H R_{v,q}^{-1}\mathbf{v}\} \tag{7}$$

where $\theta = (c_1,...,c_k,R_{v,1},...,R_{v,k})$, and the mixing proportions c_q are constrained to sum to one. The number of components k controls the flexibility of the GMM. When dealing with mixture models, it is useful to consider that there exists a hidden random variable z, taking its values in a set $Z = [1,...,k]$ with probability $P(z=q) = c_q$, $1 \leq q \leq k$. Therefore we have $\mathbf{v}|z=q \sim \mathbb{N}(0,R_{v,q})$. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{MMSE} &= E[s|\mathbf{x}] \\
&= \int p(s|\mathbf{x}).s\,ds \\
&= \int \sum_q p(s,z=q|\mathbf{x}).s\,ds \\
&= \int \sum_q p(s|z=q,\mathbf{x}).p(z=q|\mathbf{x}).s\,ds \\
&= \sum_q p(z=q|\mathbf{x}) \int p(s|z=q,\mathbf{x}).s\,ds \\
&= \sum_q \tau_q \int p(s|z=q,\mathbf{x}).s\,ds \\
&= \sum_q \tau_q E[s|\mathbf{x},q] \qquad (8)
\end{aligned}$$

where $\tau_q = p(z=q|\mathbf{x})$ is the a posteriori probability that the component q is active in the Gaussian mixture, when observing $\mathbf{x}$.

We can see that the conditional mean $E[s|\mathbf{x},q]$ is the MMSE beamformer estimator derived in the previous section, with $R_v = R_{v,q}$. In practice, modelling the signal-free interference-plus-noise signal $\mathbf{v}$ is not possible, and therefore we model the observed signal $\mathbf{x}$ instead. The desired signal estimator in equation (8) is a weighted sum of linear beamformers $\mathbf{w}_q$ over all the GMM components, and the weighted coefficients are the a posteriori probabilities of the GMM components τ_q. The mixture of beamformers (MOB) is given by:

$$\mathbf{w} = \sum_{q=1}^{k} p(z=q|\mathbf{x})\mathbf{w}_q \tag{9}$$

The resulting MOB is a weighted sum of distortionless MVDR beamformers, where the weights sum to unity, therefore it is distortionless in the look-direction.

3. MODEL LEARNING

Using the EM algorithm, we can estimate the observation model density parameters $\theta = (c_1,...,c_k,R_1,...,R_k)$ from a set of observations $D = \{\mathbf{x}(n) : n = 1,...,\eta\}$. The EM algorithm is used to find a ML estimate of parameters in probabilistic models with latent variables. The EM algorithm is an iterative algorithm with two steps: (1) an expectation step (E-step), and (2) a maximisation step (M-step). In the E-step, we calculate the probability of the latent variables, given the observed variables and the current estimates of the parameters. In the M-step, the new estimates of the parameters are calculated to maximise the conditional expectation of the complete data likelihood $p(\mathbf{x},\mathbf{z}|\theta^l)$ given the observed data under the previous parameter value. For the estimation of the parameters of the observation model, the EM algorithm may be performed as follows: At each iteration l:

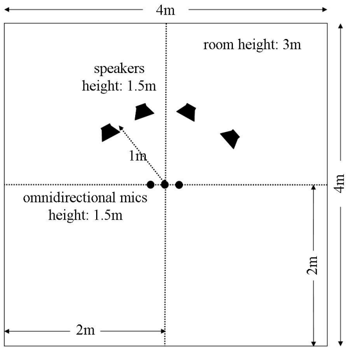

Figure 1: Layout of room used in simulations.

In the E-step, compute:

$$\tau_q^{(l)}(n) = \frac{c_q^{(l)}\mathbb{N}\left(\mathbf{x}(n)|R_q^{(l)}\right)}{\sum_{j=1}^{k} c_j^{(l)}\mathbb{N}\left(\mathbf{x}(n)|R_j^{(l)}\right)} \tag{10}$$

where $\mathbb{N}$ is the complex Gaussian distribution.

In the M-step, compute:

$$R_q^{(l+1)} = \frac{\sum_{n=1}^{\eta} \tau_q^{(l)}(n)\mathbf{x}(n)\mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_q^{(l)}(n)} \tag{11}$$

$$c_q^{(l+1)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \tau_q^{(l)}(n) \tag{12}$$

In order to perform frequency-domain beamforming, the signal received by each microphone is separated into narrow-band frequency bins using the STFT. The EM algorithm is then applied separately in each frequency bin. For each t-f point (n,f), the output of the non-linear beamformer is given by:

$$\hat{s}(n,f) = \sum_{q=1}^{k} \tau_{q,f}\mathbf{w}_{q,f}^H \mathbf{x}(n,f) \tag{13}$$

where:

$$\mathbf{w}_{q,f}^H = \frac{\mathbf{e}^H R_{q,f}^{-1}}{\mathbf{e}^H R_{q,f}^{-1}\mathbf{e}} \tag{14}$$

4. EXPERIMENTAL EVALUATION

In order to illustrate the performance of the non-linear beamformer, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [1] using the *rir.m*[1] function. The positions of the microphones and

[1] http://2pi.us/code/rir.m

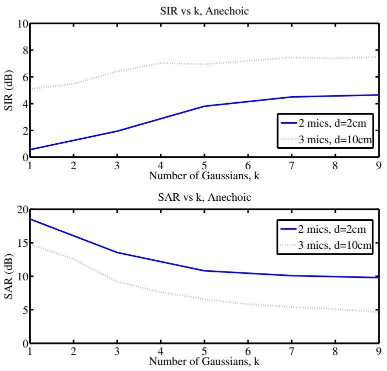

Figure 2: Average performance as a function of the number of Gaussian components k in the GMM model.

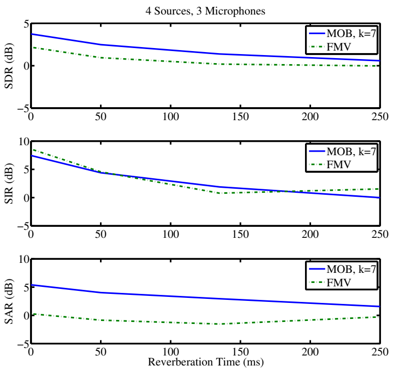

Figure 3: Separation using three microphones: average performance as a function of reverberation time.

the sources are illustrated in Figure 1. Two microphone arrays were used. The first has three microphones with a 10 cm spacing, and the second has two microphones with a 2 cm spacing. The number of the sources was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{45, 75, 100, 140\}^{\circ}$. We use the speech files used in the development data in [7], where eight speech files were grouped into two mixtures. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz.

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we use the signal to distortion ratio (SDR), source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [7]. Though we note that the SAR measure does not fully capture the nature of the distortion in the output and recommend that the reader also listens to the output signals. The speech files used in the simulations and the outputs can be found online [2].In our results, the SDR, SIR and SAR values were averaged over all the sources and mixtures.

Figure 2 shows the average performance at the output of the non-linear beamformer in the anechoic case as a function of the number of Gaussian components k in the GMM model. In this experiment, four sources were operating in an anechoic environment. The case of $k = 1$ is equivalent to a time-invariant MVDR beamformer. The SIR increases with k, and then stays constant when $k \geq 7$. The increase in the SIR is more pronounced in the two microphone case, where the separation using a time-invariant beamformer ($k = 1$) gives bad results. Although there is a unity-gain response in the direction of the desired source signal, the SAR decreases with k. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals. We stress that the MOB is by definition distortionless in the look-direction.

Figure 3 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has three microphones with a 10 cm microphone spacing. We compare the performance of a mixture of beamformers with the performance of the FMV algorithm. A STFT of frame size 1024 samples is used. In the FMV algorithm, the STFT step size is 16 samples, while a step size of 256 samples is used in the MOB algorithm. The MOB ($k = 7$) can attain an SIR of 7.5 dB in anechoic rooms.

Figure 4 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has two microphones with a 2 cm microphone spacing. We compare the performance of a mixture of beamformers with the performance of the DUET and FMV algorithms. The DUET algorithm gives a high SIR, but suffers from a low SAR. The low SAR can be attributed to the binary masking of t-f points where the sources overlap. In contrast to the MOB, this distorts the desired signal itself. In DUET, when the desired source is dominant, we attribute all the received signal to the source, and when it is not dominant, we null the output. This generates musical noise due to spectro-temporal discontinuities in the source estimates.

Figures 5 and 6 show the average performance as a function of the room reverberation time when 20 dB i.i.d. additive Gaussian noise is added at the microphones. Both the MOB and FMV were robust to the additive noise and achieved good separation performance.

5. CONCLUSION

A frequency-domain non-linear beamformer was introduced and applied to source separation for under-determined speech mixtures. The beamformer is derived assuming non-Gaussian interference-plus-noise signals modelled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers.

The non-linear beamformer has low computational costs, and does not need to know or estimate the number of interfering sources. The number of components in the mixture of Gaussians distribution controls the flexibility of the model and can be used to trade-off complexity with performance. The non-linear beamformer can be applied to microphone arrays with two or more microphones. The unity gain constraint on the direction of arrival of the desired source signal results in a clear desired signal output, and avoids any permutation ambiguities. Simulation results in under-determined mixtures with room reverberation confirmed the non-linear beamformer's ability to successfully separate speech sources.

In the future, we plan to investigate the use of an on-line EM algorithm - instead of the batch EM algorithm used herein - that allows for the observation model parameters to be updated in real-time. Furthermore, we would like to compare the MOB against other speech separation techniques.

[2] http://www.see.ed.ac.uk/~s0565920/EUSIPCO08/

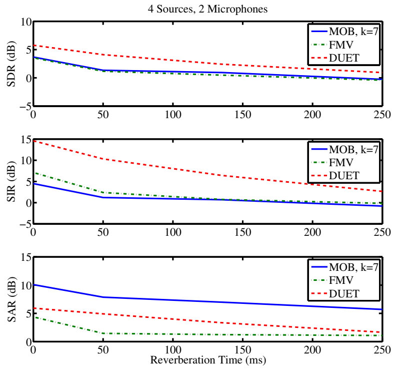

Figure 4: Separation using two microphones: average performance as a function of reverberation time.

Acknowledgements

The authors wish to thank Scott Rickard for providing the implementation of the DUET algorithm. Michael E. Davies acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership. Mohammad A. Dmour also acknowledges support of the Wolfson Microelectronics Scholarship.

REFERENCES

[1] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] O.L. Frost. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, 1972.

[3] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, 1982.

[4] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien Jr., B. Wheeler, and A. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *The Journal of the Acoustical Society of America*, 115(1):379–391, 2004.

[5] S. Rickard. The DUET blind source separation algorithm. In S. Makino, T.-W. Lee, and H. Sawada, editors, *Blind Speech Separation*. Springer Netherlands, 2007.

[6] H. L. Van Trees. *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[7] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. In *7th International Conference on Independent Component Analysis and Source Separation*, 2007.

[8] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

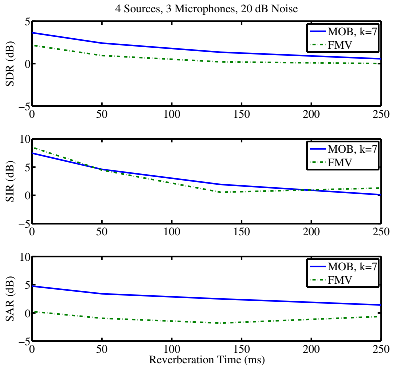

Figure 5: Separation using three microphones, with 20 dB noise: average performance as a function of reverberation time.

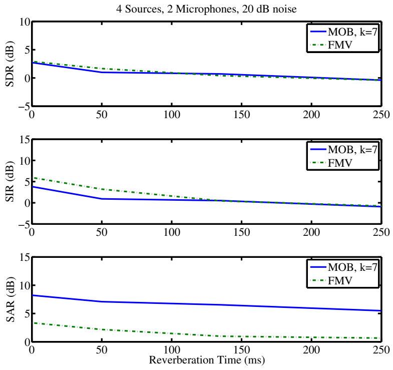

Figure 6: Separation using two microphones, with 20 dB noise: average performance as a function of reverberation time.

AN APPROACH TO UNDER-DETERMINED SPEECH SEPARATION BASED ON A NON-LINEAR MIXTURE OF BEAMFORMERS

Mohammad A. Dmour and Michael E. Davies

Institute for Digital Communications and Joint Research Institute for Signal and Image Processing,
University of Edinburgh, Edinburgh, EH9 3JL, UK
{M.Dmour, Mike.Davies}@ed.ac.uk

ABSTRACT

This paper describes frequency-domain non-linear beamformers that can extract a target speech source from among multiple interfering speech sources when there are fewer microphones than sources (the under-determined case). Our approach models the data in each frequency bin via Gaussian mixture distributions, which can be learnt using the expectation maximisation (EM) algorithm. A non-linear beamformer is then developed, based on this model. The proposed non-linear beamformer is a non-linear weighted sum of linear minimum mean square error (MMSE) or minimum variance distortionless response (MVDR) beamformers. The resulting beamformer requires the direction of arrival of the target speech source to be known in advance, but the number of interferers does not need to be known or estimated. Simulations of the non-linear beamformers in under-determined mixtures with room reverberation confirm its capability to successfully separate speech sources.

1. INTRODUCTION

Speech separation is the process of extracting a target speech source from observations corrupted by interfering sources and noise. Speech separation is used in a wide range of applications, such as hearing aids, human-computer interaction, surveillance, and hands-free telephony. The difficulty of the speech separation task depends on the way in which the signals are mixed within the acoustic environment. Speech separation is more difficult when the reverberation time of the acoustic environment is large, and when there are fewer microphones than sources (the under-determined case).

Various methods have been proposed for solving the speech separation problem. Linear multichannel filtering techniques such as independent component analysis (ICA) can attain excellent separation performance in determined mixtures. In under-determined mixtures, non-linear techniques which exploit the sparseness of speech sources and time-frequency (t-f) diversity play a vital role. One popular approach to perform under-determined speech separation is t-f masking. In the degenerate unmixing estimation technique (DUET) [9], binary masks are determined from the spatial location information contained in the short time Fourier transform (STFT) coefficients of a stereo mixture. DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. However, this method suffers from the so-called musical noise or burbling artifacts due to binary masking of t-f points where the sources overlap.

In independent factor analysis [2], it was proposed to learn the source densities from the observed data. The sources were modeled as independent random variables with Gaussian mixture models (GMMs). An expectation maximisation (EM) algorithm [4] was used to learn the parameters of the model, namely the mixing matrix, noise covariance, and source density parameters. In [3], approximations were used to overcome the problem that the number of mixtures in the observation density in [2] grows exponentially with the number of sources. The observation density is written as a summation of Gaussians with decaying weights, and then the number of Gaussians is truncated in order to retain only those with reasonable size weights.

In this paper, we describe frequency-domain non-linear beamformers that can perform speech separation of under-determined mixtures, and do not require knowledge of the number of speakers. This beamformer utilises GMMs to model the data in each frequency bin. This in turn can be learnt using the EM algorithm. The signal estimator comprises of a set of minimum mean square error (MMSE) or minimum variance distortionless response (MVDR) beamformers. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. This approach results in a "soft decision" filter for the observed signal. The resulting non-linear beamformer combines the benefits of non-linear time-varying separation in t-f masking with the benefits of spatial filtering in the linear beamformers.

The organisation of this paper is as follows. Section 2 reviews the linear MMSE beamformer, and then introduces the GMM-based non-linear beamformers. In Section 3, the EM algorithm is used to learn the GMM parameters. The experimental conditions and simulation results are presented in Section 4, followed by the conclusions in Section 5.

2. OPTIMUM BEAMFORMERS

Consider a narrow band array signal $\mathbf{x} = [x_1, ..., x_N]^T$ that consists of the desired signal arriving at the array from a known direction, and an interference signal. That is,

$$\mathbf{x} = s\mathbf{e} + \mathbf{v} \tag{1}$$

where $\mathbf{e}$ is the known $N \times 1$ array response vector in the direction of the desired source signal (the array manifold), and $\mathbf{v}$ is the $N \times 1$ complex vector of interference snapshots. We assume that the desired source and the interference are uncorrelated. The interference has spatial correlation according to the angles of the contributing interferers.

2.1 Linear MMSE beamformer

We first consider the optimum estimator whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference, assuming known desired signal direction. We assume that the desired source signal is a sample function from a zero-mean complex-valued Gaussian random process, $s \sim \mathbb{N}(0, \sigma_s^2)$. We also assume a zero-mean complex-valued Gaussian interference, $\mathbf{v} \sim \mathbb{N}(0, R_v)$. Additionally, it is assumed that the desired source and the interference are uncorrelated. Hence, $\mathbf{x} \sim \mathbb{N}(0, R_v + \sigma_s^2 \mathbf{e}\mathbf{e}^H)$, and $\mathbf{x}|s \sim \mathbb{N}(s\mathbf{e}, R_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}_{MMSE} = E[s|\mathbf{x}] = \int p(s|\mathbf{x}).s\,ds \tag{2}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [7]:

$$E\left[s|\mathbf{x}\right] = \frac{\mathbf{e}^H R_v^{-1}\mathbf{x}}{\mathbf{e}^H R_v^{-1}\mathbf{e}} . \frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{e}^H R_v^{-1}\mathbf{e}\right)^{-1}} \qquad (3)$$

The first term is an MVDR spatial filter, which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction. The second term is a single-channel Wiener post-filter. We see that the linear MMSE estimator is just a shrinkage of the MVDR beamformer.

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear or the signals are Gaussian. Speech sources are generally non-stationary and non-Gaussian. This suggests extending the optimum beamformers to exploit the non-stationarity and non-Gaussianity of speech signals.

2.2 Frequency-domain MVDR (FMV) beamformer

Speech is a non-stationary process, but over short durations speech signals can be considered stationary. In the FMV algorithm [6], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could change over longer time spans. In the FMV algorithm [6], frequency-domain signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the 32 most recent STFT values. MVDR weights are then calculated using the correlation matrix. Therefore, in the FMV algorithm, new beamformer weights are calculated every small time interval in order to reduce the contribution to the extracted signal of interfering sources active during that time interval, while having a distortionless response in the desired source DOA. Only statistics gathered over a very short period of time are used in the calculation of weights.

The quick adaptation of the beamformer weights can substantially reduce a large number of non-stationary interferences while utilising few microphones [6]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. On the other hand, the spatial distribution of the sources does not change very quickly, and we can gather statistics for the desired signal estimator over a longer time span. Thus the FMV beamformer is forced to compromise between long intervals (good statistics) and short intervals (rapid response).

2.3 GMM-based non-linear beamformers

In the frequency-domain, speech signals have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some t-f points, not all speech sources in a mixture are active at the same t-f points. It is therefore advantageous to exploit the sparsity property of speech signals in the frequency-domain in order to perform separation in under-determined environments. In order to model the speech non-Gaussianity, we propose to apply GMMs, which are widely used for modeling highly complex probability densities.

In a previous paper [5], a non-linear beamformer was developed assuming a distortionless response in the direction of the desired source, and a mixture of k zero-mean Gaussians $q = 1, \ldots, k$ with covariances $R_{x,q}$ and mixing proportions c_q were used to model the observed mixture $\mathbf{x}$ (the desired source and interference together). This leads to a simple learning algorithm and the desired signal can be estimated using this mixture of MVDR beamformers:

$$\mathbf{w}_1^H = \sum_{q=1}^{k} \tau_q \frac{\mathbf{e}^H R_{x,q}^{-1}}{\mathbf{e}^H R_{x,q}^{-1}\mathbf{e}} \qquad (4)$$

where τ_q is the relative contribution for each linear MVDR beamformer, and is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. This beamformer is a non-linear weighted sum of distortionless MVDR beamformers, where the weights sum to unity, therefore it is distortionless in the look-direction. However, since we have a distortionless constraint, we cannot exploit the sparsity of the desired source signal.

In this section, we shall describe the density of the desired source signal s as a mixture of k_s zero-mean complex-valued 1-dimensional Gaussians $q_s = 1, \ldots, k_s$ with variances σ^2_{s,q_s} and mixing proportions c_{s,q_s}:

$$p(s|\theta_s) = \sum_{q_s=1}^{k_s} c_{s,q_s} \frac{1}{\pi\sigma^2_{s,q_s}} \exp\left(\frac{-|s|^2}{\sigma^2_{s,q_s}}\right) \qquad (5)$$

where $\theta_s = (c_{s,1}, \ldots, c_{s,k_s}, \sigma^2_{s,1}, \ldots, \sigma^2_{s,k_s})$, and the mixing proportions $c_{s,q_s} = p(q_s)$ are constrained to sum to one. In addition, we shall describe the density of the interference signal $\mathbf{v}$ as a mixture of k_v zero-mean complex-valued N-dimensional Gaussians $q_v = 1, \ldots, k_v$ with covariances R_{v,q_v} and mixing proportions c_{v,q_v}:

$$p(\mathbf{v}|\theta_v) = \sum_{q_v=1}^{k_v} c_{v,q_v} \frac{1}{\pi^N \left|R_{v,q_v}\right|} \exp\left(-\mathbf{v}^H R_{v,q_v}^{-1}\mathbf{v}\right) \qquad (6)$$

where $\theta_v = (c_{v,1}, \ldots, c_{v,k_v}, R_{v,1}, \ldots, R_{v,k_v})$, and the mixing proportions $c_{v,q_v} = p(q_v)$ are constrained to sum to one. The number of components k_s and k_v controls the flexibility of the model.

The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{MMSE} &= E\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x}).s\,ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(s, q_s, q_v|\mathbf{x}).s\,ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}).p(s|\mathbf{x}, q_s, q_v).s\,ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \int p(s|\mathbf{x}, q_s, q_v).s\,ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} E\left[s|\mathbf{x}, q_s, q_v\right] \qquad (7)
\end{aligned}$$

where

$$\begin{aligned}
\tau_{q_s,q_v} &= p(q_s, q_v|\mathbf{x}) \\
&= \frac{p(\mathbf{x}|q_s, q_v).p(q_s).p(q_v)}{\sum_{q'_s=1}^{k_s}\sum_{q'_v=1}^{k_v} p(\mathbf{x}|q'_s, q'_v).p(q'_s).p(q'_v)} \qquad (8)
\end{aligned}$$

is the a posteriori probability that the components q_s and q_v are active in each respective GMM when observing $\mathbf{x}$, with $\sum_{q_s}\sum_{q_v}\tau_{q_s,q_v} = 1$.

We can see that the conditional mean $E\left[s|\mathbf{x}, q_s, q_v\right]$ is the linear MMSE beamformer estimator in equation (3), with $R_v = R_{v,q_v}$ and $\sigma_s^2 = \sigma^2_{s,q_s}$. The desired signal estimator in equation (7) is a non-linear weighted sum of linear MMSE beamformers over all the GMM components, and the weighting coefficients are the a posteriori probabilities of the GMM components τ_{q_s,q_v}. The mixture of MMSE beamformers is given by:

$$\mathbf{w}_2^H = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\sigma^2_{s,q_s}}{\sigma^2_{s,q_s} + \left(\mathbf{e}^H R_{v,q_v}^{-1}\mathbf{e}\right)^{-1}} . \frac{\mathbf{e}^H R_{v,q_v}^{-1}}{\mathbf{e}^H R_{v,q_v}^{-1}\mathbf{e}} \qquad (9)$$

In comparison to independent factor analysis [2], where all sources were modeled with a mixture of Gaussians, the mixture of MMSE beamformers models all the interfering sources using one mixture of Gaussians in the observation (microphones) domain. Consequently, the number of sources in the mixture is not required. This also avoids the exponential growth of the number of Gaussian components in the observation density with the number of sources.

In Section 4, we compare the performance of these two beamformers. Also, we use the interference Gaussian mixture model to implement a distortionless response mixture of beamformers, which uses the interference model covariances R_{v,q_v} instead of $R_{x,q}$:

$$\mathbf{w}_3^H = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\mathbf{e}^H R_{v,q_v}^{-1}}{\mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}} \tag{10}$$

3. MODEL LEARNING

Using the EM algorithm, we can estimate the model density parameters $\theta = (\theta_s, \theta_v) = (c_{s,1}, ..., c_{s,k_s}, \sigma_{s,1}^2, ..., \sigma_{s,k_s}^2, c_{v,1}, ..., c_{v,k_v}, R_{v,1}, ..., R_{v,k_v})$ from a set of observations $D = \{\mathbf{x}(n) : n = 1, ..., \eta\}$. The EM algorithm is an iterative algorithm with two steps: (1) an expectation step (E-step), and (2) a maximisation step (M-step).

In the E-step, evaluate for $q_v = 1, ..., k_v$, $q_s = 1, ..., k_s$ and every received vector $\mathbf{x}(n)$:

$$p\left(q_s, q_v | \mathbf{x}(n)\right) = \tau_{q_s,q_v}(n) = \frac{c_{s,q_s} c_{v,q_v} p\left(\mathbf{x}(n) | q_s, q_v\right)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} c_{s,q_s'} c_{v,q_v'} p\left(\mathbf{x}(n) | q_s', q_v'\right)} \tag{11}$$

where

$$\begin{aligned} p(\mathbf{x}|q_s, q_v) &= \int p(\mathbf{x}, s|q_s, q_v) ds \\ &= \int p(\mathbf{x}|s, q_v).p(s|q_s) ds \\ &= \int \mathbb{N}\left(\mathbf{x} - \mathbf{e}s, R_{v,q_v}\right).\mathbb{N}\left(s, \sigma_{s,q_s}^2\right) ds \\ &= \mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{e}\mathbf{e}^H\right) \end{aligned} \tag{12}$$

and evaluate the conditional mean and variance of the desired source given both the observed mixture and the hidden states, which are denoted by $\langle s|\mathbf{x}(n), q_s, q_v\rangle$ and $\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle$ respectively. Given the hidden states and the mixture, the likelihood of s is Gaussian:

$$\begin{aligned} p(s|\mathbf{x}, q_s, q_v) &= \frac{p(\mathbf{x}, s, q_s, q_v)}{p(\mathbf{x}, q_s, q_v)} \\ &= \frac{p(s|q_s).p(\mathbf{x}|s, q_v).p(q_s).p(q_v)}{p(\mathbf{x}|q_s, q_v).p(q_s).p(q_v)} \\ &= \frac{\mathbb{N}(s, \sigma_{s,q_s}^2).\mathbb{N}(\mathbf{x} - \mathbf{e}s, R_{v,q_v})}{\mathbb{N}\left(\mathbf{x}, R_{v,q_v} + \sigma_{s,q_s}^2 \mathbf{e}\mathbf{e}^H\right)} \\ &= \mathbb{N}\left(s - \alpha_{q_s,q_v}, \beta_{q_s,q_v}\right) \end{aligned} \tag{13}$$

where

$$\beta_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}\right)^{-1} \tag{14}$$

$$\alpha_{q_s,q_v} = \left(\sigma_{s,q_s}^{-2} + \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{e}\right)^{-1} \mathbf{e}^H R_{v,q_v}^{-1} \mathbf{x} \tag{15}$$

In the M-step, evaluate for $q_v = 1, ..., k_v$ and $q_s = 1, ..., k_s$:

$$c_{v,q_v} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p\left(q_s, q_v | \mathbf{x}(n)\right) \tag{16}$$

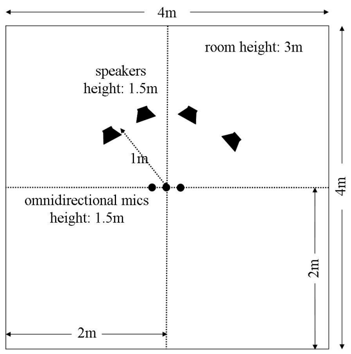

Figure 1: Layout of room used in simulations.

$$c_{s,q_s} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p\left(q_s, q_v | \mathbf{x}(n)\right) \tag{17}$$

$$\sigma_{s,q_s}^2 = \frac{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p\left(q_s, q_v | \mathbf{x}(n)\right) \langle ss^* | \mathbf{x}(n), q_s, q_v\rangle}{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} p\left(q_s, q_v | \mathbf{x}(n)\right)} \tag{18}$$

$$R_{v,q_v} = \frac{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p\left(q_s, q_v | \mathbf{x}(n)\right) \Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} p\left(q_s, q_v | \mathbf{x}(n)\right)} \tag{19}$$

where

$$\begin{aligned} \Lambda_{q_s,q_v}(n) = {} & \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n) \langle s^* | \mathbf{x}(n), q_s, q_v\rangle \mathbf{e}^H \\ & -\mathbf{e} \langle s | \mathbf{x}(n), q_s, q_v\rangle \mathbf{x}(n)^H \\ & +\mathbf{e} \langle ss^* | \mathbf{x}(n), q_s, q_v\rangle \mathbf{e}^H \end{aligned} \tag{20}$$

In this model, there is an ambiguity in associating variance between the desired source and the interference. It is possible to incorporate some of the source signal in the interference. To avoid this, updating the desired source component variances is not performed in the first few iterations. This prevents the source components shrinking to zero variance.

In order to perform frequency-domain beamforming, the signal received by each microphone is separated into narrow-band frequency bins using the STFT. The EM algorithm is then applied separately in each frequency bin. For each t-f point (n, f), the output of the non-linear beamformer is given by:

$$\hat{s}_f(n) = \mathbf{w}_f^H(n) \mathbf{x}_f(n) \tag{21}$$

4. EXPERIMENTAL EVALUATION

In order to illustrate the performance of the non-linear beamformer, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [1]. The positions of the microphones and the sources are illustrated in Figure 1. Two microphone arrays were used. The first has three microphones with a 10 cm spacing, and the second has two microphones with a 2 cm spacing. . We use speech files taken from the TIMIT speech corpus to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of

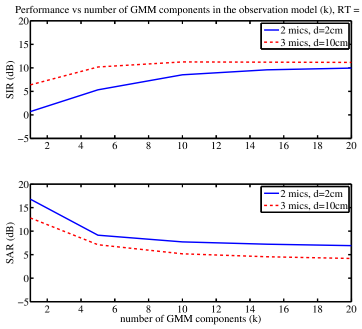

Figure 2: Average performance of the non-linear beamformer $\mathbf{w}_1$ in equation (4) as a function of the number of Gaussian components k in the GMM model.

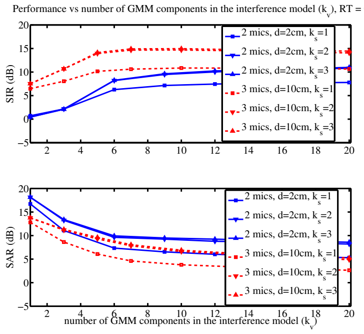

Figure 3: Average performance of the non-linear beamformer $\mathbf{w}_3$ in equation (10) as a function of the number of Gaussian components k_v and k_s in the GMM model.

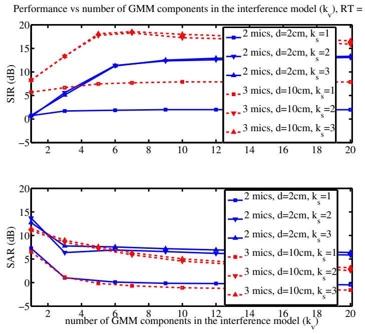

Figure 4: Average performance of the non-linear beamformer $\mathbf{w}_2$ in equation (9) as a function of the number of Gaussian components k_v and k_s in the GMM model.

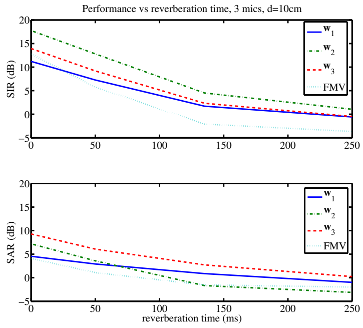

Figure 5: Separation using three microphones: average performance as a function of reverberation time.

the sources in each mixture was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{45, 75, 100, 140\}°$.

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we use the source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [8]. In our results, the SIR and SAR values were averaged over all the sources and mixtures.

Figure 2 shows the average performance at the output of the non-linear beamformer of equation (4) in the anechoic case as a function of the number of Gaussian components k in the GMM model. In this experiment, four sources were operating in an anechoic environment. The case of $k = 1$ is equivalent to a time-invariant MVDR beamformer. The SIR increases with k, but the improvement is insignificant at $k > 10$. The increase in the SIR is more pronounced in the two microphone case, where the separation using a time-invariant beamformer ($k = 1$) gives bad results. Although there is a unity-gain response in the direction of the desired source signal, the SAR decreases with k. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals. We stress that the mixture of MVDR beamformers is by definition distortionless in the look-direction.

Figure 3 shows the average performance at the output of the non-linear beamformer of equation (10) in the anechoic case as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. We can see that there is little gain for increasing the number of source Gaussian components k_s to more than two. In the two microphones case, The SIR increases with k_v, but the improvement is insignificant at $k_v > 10$. In the three microphones case, The SIR peaks around $k_v = 7$, and then levels off at higher k_v. The non-linear beamformer can attain a SIR of 10 dB in the two microphones case, and 15 dB using three microphones.

Figure 4 shows the average performance at the output of the mixture of MMSE beamformers of equation (9) in the anechoic case as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. The non-linear beamformer can attain a SIR of 13 dB in the two microphones case, and 18 dB using three microphones. However, the SAR was decreased in comparison to Figure 3 because the distortionless constraint is no longer held.

Figure 5 shows the average performance as a function of the

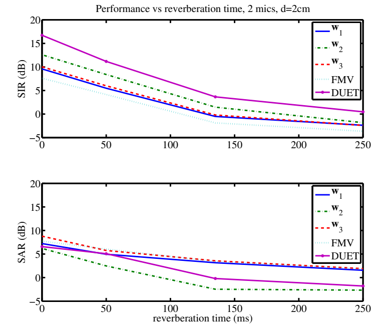

Figure 6: Separation using two microphones: average performance as a function of reverberation time.

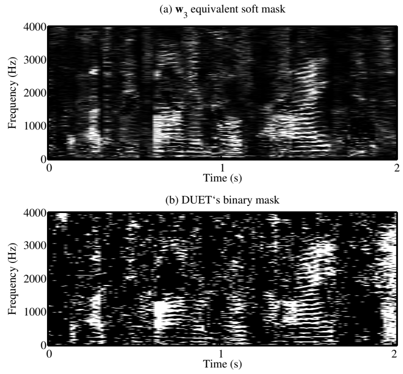

Figure 7: (a) $\mathbf{w}_3$ equivalent soft mask (b) DUET's binary masks

room reverberation time when four sources are operating, and the microphone array used has three microphones with a 10 cm microphone spacing. We compare the performance of the three non-linear beamformers (equations (4), (10), and (9)) with the performance of the FMV algorithm. $k = 15$ was used in the beamformer of equation (4), and $k_s = 2, k_v = 5$ was used in the two other beamformers. A STFT of frame size 1024 samples is used. In the FMV algorithm, a small step size of 16 samples is required, while a step size of 256 samples is sufficient in the non-linear beamformers.

Figure 6 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has two microphones with a 2 cm microphone spacing. $k = 15$ was used in the beamformer of equation (4), and $k_s = 2, k_v = 12$ was used in the two other beamformers. We compare the performance of the three non-linear beamformers with the performance of the DUET and FMV algorithms. The DUET algorithm and the mixture of MMSE beamformers ($\mathbf{w}_2$) gives a high SIR, but suffers from a very low SAR at higher reverberation times. The non-linear beamformers of equations (4) and (10) have significantly lower artifacts in higher reverberation times.

Figure 7 compares the equivalent mask of the non-linear beamformer $\mathbf{w}_3$ with the t-f mask of DUET. The equivalent mask is computed at each t-f point as the ratio of the energy of the desired signal estimate to the energy of the observed mixture. The non-linear beamformers approach results in a "soft decision" mask for the observed signal.

5. CONCLUSION

A frequency-domain non-linear beamformer was introduced and applied to source separation for under-determined speech mixtures. The beamformer is derived assuming non-Gaussian interference signals modelled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers.

The non-linear beamformer does not need to know or estimate the number of interfering sources. The number of components in the mixture of Gaussians distributions controls the flexibility of the model and can be used to trade-off complexity with performance. The non-linear beamformer can be applied to microphone arrays with two or more microphones. Simulation results in under-determined mixtures with room reverberation confirmed the non-linear beamformer's ability to successfully separate speech sources.

In the future, we would like to investigate the use of other linear constrained minimum variance (LCMV) beamformers and the use of auditory filter banks instead of the STFT. Through this, we aim to improve the performance of the beamformers in higher reverberation times.

Acknowledgements

The authors wish to thank Scott Rickard for providing the implementation of the DUET algorithm. Michael E. Davies acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership. Mohammad A. Dmour also acknowledges support of the Wolfson Microelectronics Scholarship.

REFERENCES

[1] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[2] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[3] M. Davies and N. Mitianoudis. Simple mixture model for sparse overcomplete ICA. In *Vision, Image and Signal Processing, IEE Proceedings*, volume 151, pages 35–43, feb 2004.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[5] M. Dmour and M.E. Davies. Under-determined speech separation using gmm-based non-linear beamforming. In *Proceedings of the sixteenth European Conference on Signal Processing (EUSIPCO 2008)*, 2008.

[6] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien Jr., B. Wheeler, and A. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *The Journal of the Acoustical Society of America*, 115(1):379–391, 2004.

[7] H. L. Van Trees. *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[8] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, jul 2006.

[9] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, jul 2004.

A New Framework for Underdetermined Speech Extraction Using Mixture of Beamformers

Mohammad A. Dmour *, *Student Member, IEEE,* and Mike Davies, *Member, IEEE*

***Abstract*—This paper describes frequency-domain non-linear mixture of beamformers that can extract a speech source from a known direction when there are fewer microphones than sources (the underdetermined case). Our approach models the data in each frequency bin via Gaussian mixture distributions, which can be learned using the expectation maximization algorithm. The model learning is performed using the observed mixture signals only, and no prior training is required. Non-linear beamformers are then developed based on this model. The proposed estimators are a non-linear weighted sum of linear minimum mean square error or minimum variance distortionless response beamformers. The resulting non-linear beamformers do not need to know or estimate the number of sources, and can be applied to microphone arrays with two or more microphones. We test and evaluate the described methods on underdetermined speech mixtures.**

I. Introduction

MOST audio signals result from the mixing of several sound sources. In many applications, there is a need to separate the multiple sources or extract a source of interest while reducing undesired interfering signals and noise. The estimated signals may then be either directly listened to or further processed, giving rise to a wide range of applications such as hearing aids, human-computer interaction, surveillance, and hands-free telephony.

There has been a lot of research on the speech enhancement problem, where the focus is on attenuating the background noise. Speech denoising algorithms are well established and have been used for many years [1], [2]. The extension of the speech enhancement problem to deal with mixtures of speech sources is a topic of intense research. Source mixing can occur in a wide variety of situations under different environments. The difficulty of source extraction depends on the way the source signals are mixed within the environment and on the a priori knowledge of the sources, microphones, and mixing parameters. Blind methods do not assume a priori knowledge of sources, microphones, or mixing parameters. By contrast, informed methods exploit some a priori information about the sources and microphones (for example, the location of a desired source). In general, the problem is more difficult when the reverberation time (RT) of the acoustic environment is large, and in the underdetermined case (fewer microphones than sources).

In general, observations are obtained at the output of a set of microphones, each receiving different combinations of the source signals. The use of microphone arrays gives one the opportunity to exploit the fact that the desired source and the interfering sources originate at different points in space. Suppose that M simultaneously active source signals are mixed and observed at N microphones. The signal recorded at the j^{th} microphone at time t can be modeled as:

$$x_j(t) = \sum_{i=1}^{M} \sum_{p=0}^{P-1} a_{ji}(p) s_i(t-p) \tag{1}$$

where a_{ji} represents the impulse response of the acoustic path from source i to microphone j and P is the length of the impulse response between each source-microphone pair. A mixture is termed a determined mixture when the number of microphones is equal to the number of sources, overdetermined when the number of microphones is larger than the number of sources, and underdetermined when it is smaller. In certain applications, source separation methods are used to estimate separated signals $\hat{s}_1, ..., \hat{s}_M$, which correspond to each of the original source signals $s_1, ..., s_M$. In many practical applications, however, prior information about a desired source, such as source location or identity, might be available and exploited to extract only one source of interest.

Various methods have been proposed for solving the speech separation problem. One approach is to use statistical modeling of source signals. Independent component analysis (ICA) is one of the major statistical tools used. In ICA, separation is performed on the assumption that the source signals are statistically independent, and does not require information on microphone array configuration or the direction of arrival (DOA) of the source signals to be available. To perform source separation, we process the mixture channels by a set of linear time-invariant demixing filters. ICA implicitly estimates the source directions by maximizing the independence of the sources, and acts as an adaptive null beamformer that reduces the undesired sources. However, some aspects limit the application of ICA to real-world environments. Most ICA methods assume the number of sources is given a priori. In general, classical ICA techniques cannot perform source separation in the underdetermined mixtures case. For some excellent reviews of convolutive ICA methods for speech separation, see [3], [4]. Another strategy suited to speech mixtures is to incorporate speaker-independent models that can be learned from large speech datasets. For example, in

Mohammad A. Dmour would like to acknowledge the support of the Wolfson Microelectronics Scholarship. Mike Davies acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the HeriotWatt University as a component part of the Edinburgh Research Partnership.

The authors are with IDCOM and the Joint Research Institute for Signal and Image Processing, Edinburgh University, Edinburgh EH9 3JL, U.K. (e-mail: m.dmour@ed.ac.uk; mike.davies@ed.ac.uk)

[5], [6], Gaussian mixture models (GMMs) are used to model the sources, and the parameters of the mixing channel and noise were inferred using variational expectation maximization (EM) techniques [7]. The sources are then estimated using a minimum mean square error (MMSE) estimator. However, this method requires the number of sources to be known, and it cannot perform source separation in the underdetermined mixtures case.

A popular approach to speech extraction is to use adaptive beamforming techniques. With adaptive beamforming, the microphone array is also used to form a spatial filter which can extract a signal from a specific direction and reduce signals from other directions. For example, in minimum-variance distortionless response (MVDR) beamforming, the beamformer response is constrained so that signals from the direction of interest are passed with no distortion, while it suppresses noise and interference. In [8], [9], beamforming weights were calculated using time-domain recursive algorithms. It was shown recently in [10] that a frequency-domain MVDR (FMV) beamformer which performs sample matrix inversion using statistics estimated from a short sample support gives better performance than time-domain recursive algorithms in non-stationary acoustic environments. Unlike the ICA approach, adaptive beamforming requires information about the microphone array configuration and the sources (for example, the direction of the desired source). However, adaptive beamforming techniques can attenuate spatially spread and reverberant interferers, and there is no need to determine their number. In general, linear adaptive beamforming can attain excellent separation performance in determined or over-determined time-invariant mixtures. However, in underdetermined mixtures, perfect attenuation of all interferers becomes impossible and only partial interference attenuation is possible. This, in turn, leads to performance degradation.

In the underdetermined mixtures case, the assumption of spatial diversity alone is insufficient to perform source separation/extraction, thereby necessitating additional assumptions. The assumption that the sources have a sparse representation in a given basis is an increasingly popular addition. Sparseness of a signal means that only a few instances have a value significantly different from zero. A sparse representation of a speech signal can be achieved by a short term Fourier transform (STFT). One popular approach to sparsity-based separation is time-frequency masking [11]–[17]. This approach is a special case of non-linear time-varying filtering that estimates the desired source s_i from a mixture signal x_j by:

$$\hat{s}_i(n,f) = M_i(n,f)x_j(n,f) \tag{2}$$

where $s_i(n,f)$ and $x_j(n,f)$ are the STFT coefficients of $s_i(t)$ and $x_j(t)$ respectively in the time frame n and frequency bin f, and M_i is a time-frequency mask containing positive gains which must be adapted to extract the desired source s_i from the observed mixture. A popular method to estimate the time-frequency masks using only two microphones is the degenerate unmixing estimation technique (DUET) [11], [12]. It is assumed that the time-frequency representation of speech signals are approximately disjoint (i.e., sources do not overlap too much):

$$s_i(n,f)s_j(n,f) \simeq 0, \qquad \forall i \neq j, \forall f \tag{3}$$

This assumption is not fully met in practice. In DUET, the source directions and the active source indices are alternately optimized by partitioning the mixture STFT coefficients based on the inter-channel level/phase difference (ILD/IPD). DUET is capable of performing separation of two or more sources using just two channels, and without significant computational complexity. However, this method suffers from so-called musical noise or burbling artifacts due to binary masking of time-frequency points where the sources overlap. An extension to the DUET algorithm for more than two microphones in was proposed in in [18], [19]. This method, Multiple sENsor dUET (MENUET), can be applied to non-linear microphone arrangements with 2- or 3-dimensional arrays. In [13]–[17], probabilistic models are used to model the IPD/ILD, and after estimating its parameters with an EM algorithm [20], soft masks can be derived. All of these methods require the number of sources to be given a priori, and it is difficult to expand these methods to more than two microphones. Furthermore, separation methods based on time-frequency masking suffer from the fact that clustering becomes difficult in reverberation, as ILD/IPD resulting from each sound source then tend to spread and overlap.

When only one microphone is available, source separation/extraction becomes significantly more challenging, as spatial cues are absent in this case. In this situation, the assumptions of independence and time-frequency sparsity becomes insufficient, and more advanced source models relying on spectro-temporal models are needed. Different strategies have been employed using these models [21]–[23]. However, they require prior training and some knowledge about the identity of the speech or music sources in the mixture.

In this paper, we deal with the problem of extracting a speech source of interest from a known direction. we present a framework which extends the use of beamforming techniques to underdetermined speech mixtures. We describe frequency-domain non-linear mixture of beamformers that can extract a desired speech source from a known direction when there are fewer microphones than sources, and do not require knowledge of the number of speakers. These beamformers utilize Gaussian mixture models (GMMs) to model the observation data in each frequency bin. In contrast to other speech enhancement and separation methods which use GMMs such as [6], [23], [24], our approach do not couple the Gaussian states across frequency, and the covariance matrices of each Gaussian state represent a spatial covariance matrix. The model learning is performed using the observed mixture signals only, and no prior training is required. The signal estimator comprises of a set of MMSE or MVDR beamformers. In order to estimate the signal, all beamformers are concurrently applied to the observed signal, and the weighted sum of the beamformers' outputs is used as the signal estimator, where the weights are the posterior probabilities of the GMM states. These weights are specific to each time-frequency point. This approach results in a soft decision filter for the observed signal. The resulting non-linear beamformer combines the benefits of non-linear time-varying separation in time-frequency masking with the benefits of spatial filtering in the linear beamformers.

The remainder of the paper is structured as follows. Section II presents the signal mixing model. Section III reviews the linear MMSE, MVDR, and FMV beamformers. Then, in Section IV, the proposed GMM-based non-linear beamformers are described. The experimental conditions and simulation results are presented in Section V, followed by the conclusions in Section VI.

II. Mixing Model

Consider the convolved mixing model in (1). The time-domain observed signals $x_j(t)$ may be mapped to the time-frequency domain using the STFT. Denoting the STFT coefficients of $x_j(t)$ and $s_i(t)$ as $x_j(n,f)$ and $s_i(n,f)$, in the time frame n and frequency bin f, and approximating the mixing filters by complex mixing scalars $a_{ji}(f)$, we get:

$$x_j(n,f) = \sum_{i=1}^{M} a_{ji}(f) s_i(n,f) \tag{4}$$

Assuming we are only interested in extracting source i', $i' \in \{1,2,...,M\}$, the mixing model in (4) can be reformulated as:

$$\begin{aligned} x_j(n,f) &= a_{ji'}(f)s_{i'}(n,f) + \sum_{\substack{i=1 \\ i \neq i'}}^{M} a_{ji}(f)s_i(n,f) \\ &= a_{ji'}(f)s_{i'}(n,f) + v_j(n,f) \end{aligned} \tag{5}$$

where v_j represents the contribution of the interferers to the mixture signal x_j. In vector form, the mixing model can be written as:

$$\mathbf{x}(n,f) = \mathbf{a}(f)s(n,f) + \mathbf{v}(n,f) \tag{6}$$

where $\mathbf{x}(n,f) = [x_1(n,f),...,x_N(n,f)]^T$ is the observed multichannel mixture signal, $\mathbf{a}(f)$ is the $N \times 1$ array response vector in the direction of the desired source signal s (also called the propagation vector or steering vector), and $\mathbf{v}(n,f) = [v_1(n,f),...,v_N(n,f)]^T$ is the $N \times 1$ vector of the interferers' contribution to the mixture signal. We assume that the direction of the desired source signal is known. In this model, no assumptions are made about the interferers. The interferers are not restricted to point sources in low reverberation conditions, but can also be of any nature such as spatial extended sources, diffuse sources, or a combination of them. The array response vector $\mathbf{a}(f)$ is the representation of the delays and the attenuation in the frequency domain, and depends on the array geometry and the direction of the desired source signal. If d were to represent the microphone spacing, c the sound velocity, ϕ the DOA relative to broadside, and assuming far-field conditions, we have for a uniform linear array:

$$\mathbf{a}(f) \;=\; [e^{-\iota 2\pi f \Delta_1},...,e^{-\iota 2\pi f \Delta_N}]^T \tag{7}$$

where $\iota = \sqrt{-1}$, and $\Delta_j = (j-1)(d/c)\sin\phi$. Note that $\mathbf{x}$, $\mathbf{a}$, s, and $\mathbf{v}$ are complex valued, and depend on frequency f, but for readability and simplicity, we will omit this variable in the rest of paper. From now on, we implicitly work in a given frequency band.

III. Optimum Beamformers

A. Linear MMSE Beamformer

We first consider the optimum estimator whose output is the MMSE estimate of the desired signal s in the presence of Gaussian interference, assuming a known desired signal direction. We assume that the desired source signal is a sample function from a zero-mean complex-valued Gaussian random process, $s \sim \mathbb{N}(0,\sigma_s^2)$. We also assume a zero-mean complex-valued Gaussian interference, $\mathbf{v} \sim \mathbb{N}(0,\mathbf{R}_v)$. Additionally, it is assumed that the signal and interference snapshots are uncorrelated. Hence, $\mathbf{x} \sim \mathbb{N}(0, \mathbf{R}_v + \sigma_s^2\mathbf{a}\mathbf{a}^H)$, and $\mathbf{x}|s \sim \mathbb{N}(\mathbf{a}s, \mathbf{R}_v)$, where $(.)^H$ denotes the Hermitian transpose operator. The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\hat{s}_{\text{MMSE}} = \mathrm{E}\left[s|\mathbf{x}\right] = \int s \; p(s|\mathbf{x}) \; ds \tag{8}$$

This mean is referred to as the conditional mean. It can be shown that the conditional mean can be expressed as [25]:

$$\hat{s}_{\text{MMSE}} \;=\; \underbrace{\frac{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{x}}{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}}}_{\text{MVDR}} \; \underbrace{\frac{\sigma_s^2}{\sigma_s^2 + \left(\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}\right)^{-1}}}_{\text{Wiener post filter}} \tag{9}$$

The first term is an MVDR spatial filter, which suppresses the interfering signals and noise without distorting the signal propagating along the desired source direction. The second term is a single-channel Wiener post-filter. We see that the MMSE estimator is just a shrinkage of the MVDR beamformer. Unfortunately, the MMSE beamformer depends explicitly on σ_s^2 which is typically unknown. However, we can obtain a beamformer that does not depend on σ_s^2 by imposing a distortionless response in the specified direction. The result is the MVDR beamformer, and the estimate of the desired source signal can be written as:

$$\hat{s}_{\text{MVDR}} = \frac{\mathbf{a}^H\mathbf{R}_v^{-1}}{\mathbf{a}^H\mathbf{R}_v^{-1}\mathbf{a}}\mathbf{x} \tag{10}$$

In general, the conditional mean estimator is not linear. The MMSE estimator is linear if either the estimator is constrained to be linear, or all the signals are Gaussian. However, speech sources are generally non-stationary and non-Gaussian. This suggests extending the optimum beamformers to exploit the non-stationarity and non-Gaussianity of speech signals.

B. Frequency-Domain MVDR (FMV) Beamformer

Speech is a non-stationary process, but over short durations speech signals can be considered stationary. In the FMV algorithm [10], it is assumed that source activity patterns are constant over small time intervals of speech signals in each frequency band, but could vary over longer time spans. In the FMV algorithm, time-frequency representations of the mixture signals are stored in a buffer, and a correlation matrix is calculated for each frequency bin using the 32 most recent mixture signal STFT values. MVDR weights are then calculated using the correlation matrix. Therefore, in the

FMV algorithm, new beamformer weights are calculated for every small time interval, in order to reduce the contribution of interfering sources active during that time interval to the extracted signal while maintaining a distortionless response in the desired source DOA. Only statistics gathered over a very short period of time are used in the calculation of weights.

The quick adaptation of the beamformer weights can substantially reduce a large number of non-stationary interferences while utilizing few microphones [10]. But the computational load is high due to recurrent matrix inversions in each frequency band and the need to have a very small step size in the STFT. In practice, however, source activity patterns can change abruptly between samples, and the FMV will perform spatial filtering based on the average power of the interfering sources active in the time interval during which the beamformer weights are calculated. On the other hand, the spatial distribution of the sources does not change very quickly, and we can gather statistics for the desired signal estimator over a longer time span. Thus the FMV beamformer is forced to compromise between long intervals (good statistics) and short intervals (rapid response).

IV. Proposed Method: Mixture of Beamformers

In the time-frequency domain, speech signals typically have a super-Gaussian (sparse) distribution, due to a combination of the non-stationarity and harmonic content of speech. Therefore, even if sources might overlap at some time-frequency points, not all speech sources in a mixture are active at the same time-frequency points. It is therefore advantageous to exploit the sparsity property of speech signals in the time-frequency domain in order to perform separation in underdetermined environments. In this paper, we use GMMs to model the speech non-Gaussianity and the spatial distribution of the sources.

In this section, we present three non-linear beamformers that can perform underdetermined speech separation. The first two non-linear beamformers are based on modeling the desired source signal s and the interference $\mathbf{v}$ separately. The desired source signal is modeled using a 1-dimensional GMM, and the observed interference is modeled using an N-dimensional GMM, where N is the number of mixture channels. The third non-linear beamformer is based on modeling the observed mixture signal $\mathbf{x}$ (the desired source and interference together) using an N-dimensional GMM.

We describe the density of the interference signal $\mathbf{v}$ in each frequency bin as a mixture of k_v zero-mean, complex-valued, N-dimensional Gaussians with indices $q_v = 1, ..., k_v$, covariances $\mathbf{R}_{v,q_v}$ and mixing proportions c_{v,q_v}:

$$p(\mathbf{v}|\theta_v) = \sum_{q_v=1}^{k_v} c_{v,q_v} \frac{1}{\pi^N |\mathbf{R}_{v,q_v}|} \exp\left(-\mathbf{v}^H \mathbf{R}_{v,q_v}^{-1} \mathbf{v}\right) \tag{11}$$

where $\theta_v = \{c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_v \leq k_v\}$, and the mixing proportions $c_{v,q_v} = p(q_v)$ (prior probabilities of the Gaussian states) are constrained to sum to one. In addition, we shall describe the density of the desired source signal s in each frequency bin as a mixture of k_s zero-mean complex-valued 1-dimensional Gaussians with indices $q_s = 1, ..., k_s$, variances σ^2_{s,q_s} and mixing proportions c_{s,q_s}:

$$p(s|\theta_s) = \sum_{q_s=1}^{k_s} c_{s,q_s} \frac{1}{\pi \sigma^2_{s,q_s}} \exp\left(\frac{-|s|^2}{\sigma^2_{s,q_s}}\right) \tag{12}$$

where $\theta_s = \{c_{s,q_s}, \sigma^2_{s,q_s} : 1 \leq q_s \leq k_s\}$, and the mixing proportions $c_{s,q_s} = p(q_s)$ (prior probabilities of the Gaussian states) are constrained to sum to one. The number of components k_s and k_v control the flexibility of the model. In our model, the Gaussian states are not coupled across frequency, and the parameters $\{\theta_s, \theta_v\}$ are frequency dependent.

The MMSE estimate of the desired signal s is the mean of the a posteriori probability density of s given $\mathbf{x}$:

$$\begin{aligned}
\hat{s}_{\text{MMSE}} &= \text{E}\left[s|\mathbf{x}\right] = \int p(s|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(s, q_s, q_v|\mathbf{x})\ s\ ds \\
&= \int \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x})\ p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} p(q_s, q_v|\mathbf{x}) \int p(s|\mathbf{x}, q_s, q_v)\ s\ ds \\
&= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \text{E}\left[s|\mathbf{x}, q_s, q_v\right] \qquad (13)
\end{aligned}$$

where

$$\begin{aligned}
\tau_{q_s,q_v} &= p(q_s, q_v|\mathbf{x}) \\
&= \frac{p(\mathbf{x}|q_s, q_v)\ p(q_s)\ p(q_v)}{\sum_{q_s'=1}^{k_s} \sum_{q_v'=1}^{k_v} p(\mathbf{x}|q_s', q_v')\ p(q_s')\ p(q_v')} \qquad (14)
\end{aligned}$$

is the a posteriori probability that the components q_s and q_v are active in their respective GMMs when observing $\mathbf{x}$, with $\sum_{q_s} \sum_{q_v} \tau_{q_s,q_v} = 1$. The posteriori probability is specific to each time frequency point, and has a non-linear dependency on the observed data.

We can see that the conditional mean $\text{E}\left[s|\mathbf{x}, q_s, q_v\right]$ is the linear MMSE beamformer estimator in (9), with $\mathbf{R}_v = \mathbf{R}_{v,q_v}$ and $\sigma_s^2 = \sigma^2_{s,q_s}$. The desired signal estimator in (13) is a non-linear weighted sum of linear MMSE beamformers over all the GMM components, and the weighting coefficients are the a posteriori probabilities of the GMM components τ_{q_s,q_v} (specific to each time-frequency point). This mixture of MMSE beamformers will be denoted by $\mathbf{w}_1$ and is given by [26]:

$$\mathbf{w}_1 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\sigma^2_{s,q_s}}{\sigma^2_{s,q_s} + \left(\mathbf{a}^H \mathbf{R}^{-1}_{v,q_v} \mathbf{a}\right)^{-1}} \frac{\mathbf{a}^H \mathbf{R}^{-1}_{v,q_v}}{\mathbf{a}^H \mathbf{R}^{-1}_{v,q_v} \mathbf{a}} \tag{15}$$

In comparison to independent factor analysis [27], where sources were also modeled with GMMs, the mixture of MMSE beamformers models all the interfering sources using one N-dimensional mixture of Gaussians in the observation (microphones) domain. Consequently, the number of interferers in

Algorithm 1 Separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$

```
1: Compute the STFT of the mixture $\mathbf{x}$.
2: Apply the EM algorithm (see Appendix A) separately in each frequency bin to compute $\{\tau_{q_s,q_v}(n,f), \sigma^2_{s,q_s}(f), \mathbf{R}_{v,q_v}(f) : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$.
3: For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:
       $$\hat{s}(n,f) = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}(n,f)\, \mathbf{w}_{q_s,q_v}(f)\, \mathbf{x}(n,f) \quad (17)$$
   where $\mathbf{w}_{q_s,q_v}(f)$ can be either a linear MMSE or a linear MVDR beamformer:
       $$\mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) = \frac{\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f)}{\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f)\mathbf{a}(f)} \quad (18)$$
       $$\mathbf{w}^{\text{MMSE}}_{q_s,q_v}(f) = H^{\text{Wiener}}_{q_s,q_v}(f)\, \mathbf{w}^{\text{MVDR}}_{q_s,q_v}(f) \quad (19)$$
   where the scalar, single channel Wiener post filter is given by:
       $$H^{\text{Wiener}}_{q_s,q_v}(f) = \frac{\sigma^2_{s,q_s}(f)}{\sigma^2_{s,q_s}(f) + \left(\mathbf{a}(f)^H \mathbf{R}^{-1}_{v,q_v}(f)\mathbf{a}(f)\right)^{-1}} \quad (20)$$
4: The corresponding time-domain signal $\hat{s}$ is derived by an STFT inversion.
```

Algorithm 2 Separation procedure using $\mathbf{w}_3$

```
1: Compute the STFT of the mixture $\mathbf{x}$.
2: Apply the EM algorithm (see Appendix B) separately in each frequency bin to compute $\{\tau_{q_x}(n,f), \mathbf{R}_{x,q_x}(f) : 1 \leq q_x \leq k_x\}$.
3: For each time-frequency point $(n,\ f)$, the output of the non-linear beamformer is given by:
       $$\hat{s}(n,f) = \sum_{q_x=1}^{k_x} \tau_{q_x}(n,f)\, \mathbf{w}_{q_x}(f)\, \mathbf{x}(n,f) \quad (23)$$
   where:
       $$\mathbf{w}_{q_x}(f) = \frac{\mathbf{a}(f)^H \mathbf{R}^{-1}_{x,q_x}(f)}{\mathbf{a}(f)^H \mathbf{R}^{-1}_{x,q_x}(f)\, \mathbf{a}(f)} \quad (24)$$
4: The corresponding time-domain signal $\hat{s}$ is derived by an STFT inversion.
```

the mixture is not required to be known or have a unique mixing structure. This also avoids the exponential growth of the number of Gaussian components in the observation density with the number of sources.

If a distortionless response in the direction of the desired source is required, a distortionless response mixture of MVDR beamformers can be used [26]:

$$\mathbf{w}_2 = \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v} \frac{\mathbf{a}^H \mathbf{R}^{-1}_{v,q_v}}{\mathbf{a}^H \mathbf{R}^{-1}_{v,q_v} \mathbf{a}} \quad (16)$$

This mixture of MVDR beamformers is a non-linear weighted sum of linear distortionless MVDR beamformers, where the weights sum to unity. As a result, it is constrained to a distortionless response in the look-direction. By distortionless we mean it has a unity gain in the look-direction at all time-frequency points.

In Appendix A, we develop an EM algorithm to learn the model density parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. The model learning is applied separately in each frequency bin.

We briefly summarize the main steps in the separation procedure using $\mathbf{w}_1$ or $\mathbf{w}_2$ in Algorithm 1. Note that the model learning step is applied separately in each frequency bin, and that the Gaussian states' posterior probabilities are specific to each time-frequency point (no coupling across all frequencies).

In a previous paper [28], a non-linear beamformer was developed assuming a distortionless response in the direction of the desired source. A mixture of k_x zero-mean, complex-valued, N-dimensional Gaussians with indices $q_x = 1, ..., k_x$, covariances $\mathbf{R}_{x,q_x}$ and mixing proportions c_{x,q_x} was used to model the observed mixture $\mathbf{x}$ (the desired source and interference together) in each frequency bin:

$$p(\mathbf{x}|\theta_x) = \sum_{q_x=1}^{k_x} c_{x,q_x} \frac{1}{\pi^N |\mathbf{R}_{v,q_v}|} \exp\left(-\mathbf{x}^H \mathbf{R}^{-1}_{x,q_x} \mathbf{x}\right) \quad (21)$$

where $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x\}$, and the mixing proportions $c_{x,q_x} = p(q_x)$ (prior probabilities of the Gaussian states) are constrained to sum to one. This leads to a simple learning algorithm, and the learning of model parameters is independent on the desired source direction. The desired signal can be estimated using this mixture of MVDR beamformers [28]:

$$\mathbf{w}_3 = \sum_{q_x=1}^{k_x} \tau_{q_x} \frac{\mathbf{a}^H \mathbf{R}^{-1}_{x,q}}{\mathbf{a}^H \mathbf{R}^{-1}_{x,q} \mathbf{a}} \quad (22)$$

where $\tau_{q_x} = p(q_x|\mathbf{x})$ is the relative contribution for each linear MVDR beamformer, and is calculated as the posterior probability (specific to each time-frequency point) of its corresponding Gaussian component. The resulting beamformer has a unity gain in the look-direction at all time-frequency points.

In Appendix B, we develop an EM algorithm to learn the observation model density parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$. The model learning is applied separately in each frequency bin.

The main steps in the separation procedure using $\mathbf{w}_3$ are summarized in Algorithm 2.

V. Experimental Evaluation

A. Setup

In order to illustrate the performance of the non-linear beamformers, multichannel recordings of several speech sources were simulated using impulse responses determined by the room image method [29]. The positions of the microphones and the sources were as illustrated in Fig. 1. Two microphone arrays were used. The first has three microphones with a spacing d = 2.5 cm, and the second has two microphones

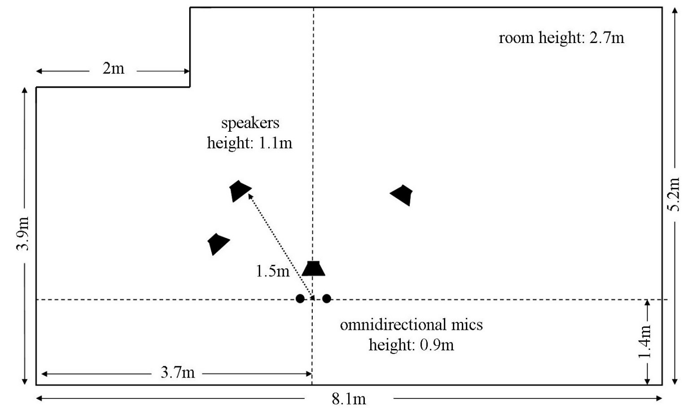

Fig. 2. Layout of room used in recordings.

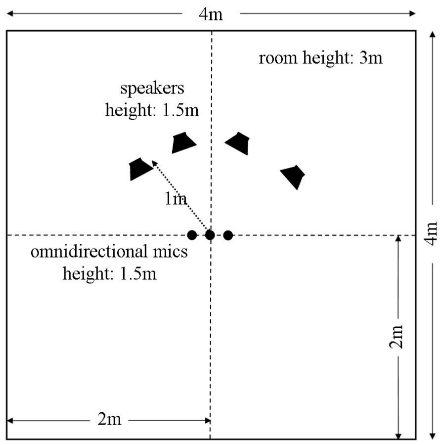

Fig. 1. Layout of room used in simulations.

with a spacing $d = 5$ cm. In section V-H, live recordings in the room illustrated in Fig. 2 were used.

We used speech files taken from the TIMIT speech corpus [30] to create five mixtures of male sources, and five mixtures of female sources. The speech signals were of a duration equal to 10 s, and were sampled at 16 kHz. The number of the sources in each mixture was four. The sources were placed in a semi-circle of radius 1 m around the microphone arrays at angles $\phi = \{-45, -15, 10, 50\}°$.

B. Evaluation Measures

To measure the quality of the signal estimate $\hat{s}$ with respect to the original signal s, we used the source to distortion ratio (SDR), source to interference ratio (SIR) and the sources to artifacts ratio (SAR) calculated as defined in [31]. The computation of the evaluation measures involves two steps. First, the estimated signal $\hat{s}$ is decomposed as

$$\hat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}} \tag{25}$$

where s_{target} is a version of the desired source s modified by an allowed distortion, and where e_{interf} and e_{artif} are respectively the interferences and artifacts error terms. In a second step, we compute energy ratios to evaluate the relative amount of each of these terms as follows:

$$\text{SDR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \tag{26}$$

$$\text{SIR} = 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \tag{27}$$

$$\text{SAR} = 10 \log_{10} \frac{\|s_{\text{target}} + e_{\text{interf}}\|^2}{\|e_{\text{artif}}\|^2} \tag{28}$$

In our results, the SDR, SIR and SAR values were averaged over all the sources and mixtures.

C. Algorithm Parameters

Unless mentioned otherwise, we use the values listed in Table I for the STFT frame size, STFT step size, number of GMM components, and number of iterations.

D. Effect Of Design Parameters

We first investigate the effect of various parameters on the performance of the non-linear beamformers. We study the effect of the number of Gaussian components in the GMM model, the required number of EM iterations, and the effect of the learning block size.

TABLE I
ALGORITHM PARAMETERS.

		$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$
STFT frame		1024	1024	1024
STFT step		256	256	256
GMM components	(2 mics)	$k_s = 2$, $k_v = 15$	$k_s = 2$, $k_v = 15$	$k_x = 15$
	(3 mics)	$k_s = 2$, $k_v = 5$	$k_s = 2$, $k_v = 5$	$k_x = 5$
EM Iterations	(2 mics)	100	100	50
	(3 mics)	100	100	20

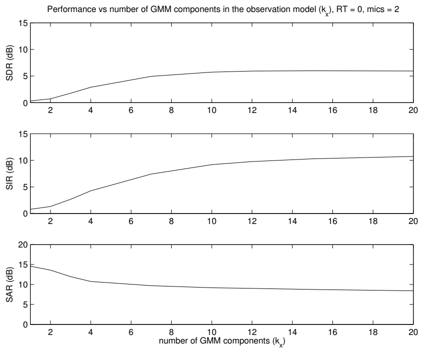

Fig. 3. Average performance of the non-linear beamformer $\mathbf{w}_3$ in equation (22) as a function of the number of Gaussian components k_x in the GMM model.

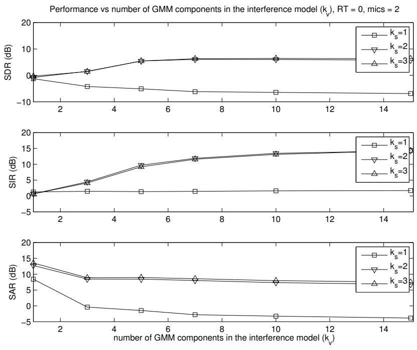

Fig. 5. Average performance of the non-linear beamformer $\mathbf{w}_1$ in equation (15) as a function of the number of Gaussian components k_v and k_s in the GMM model.

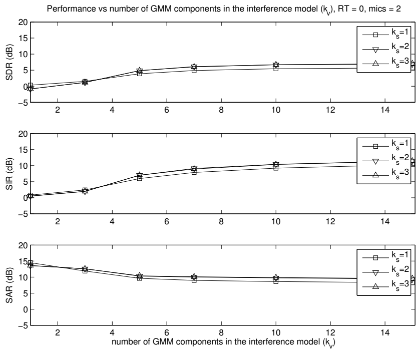

Fig. 4. Average performance of the non-linear beamformer $\mathbf{w}_2$ in equation (16) as a function of the number of Gaussian components k_v and k_s in the GMM model.

1) Effect of the Number of Gaussian Components: In this experiment, four sources were operating in an anechoic environment (RT = 0), and the microphone array used has two microphones with a 5 cm microphone spacing. Fig. 3 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_3$ defined in (22) as a function of the number of Gaussian components k_x in the GMM model. The case of $k_x = 1$ is equivalent to a time-invariant MVDR beamformer. The SIR increases with k_x, but the improvement is insignificant at $k_x > 10$. Although there is a unity-gain response in the direction of the desired source signal, the SAR decreases with k_x. The decrease in the SAR can be attributed to the non-linear attenuation of the interfering sources. These artifacts therefore introduce distortion only into the residual interfering signals. We stress that the mixture of MVDR beamformers is by definition distortionless in the look-direction.

Fig. 4 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_2$ defined in (16) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. We can see that there is little to be gained in increasing the number of source Gaussian components k_s to more than two. The SIR increases with k_v, but the improvement again is insignificant for $k_v > 10$. The non-linear beamformer can attain a SIR of 10 dB in the two microphones case.

Fig. 5 shows the average performance at the output of the mixture of MMSE beamformers $\mathbf{w}_1$ defined in (15) as a function of the number of Gaussian components in the interference model k_v and the number of Gaussian components in the source model k_s. The non-linear beamformer can attain an SIR of 13 dB. However, the SAR was reduced in comparison to Fig. 4 because the distortionless constraint is no longer held.

2) Effect of the Number of Iterations: Fig. 6 shows the average performance at the output of the non-linear beamform-

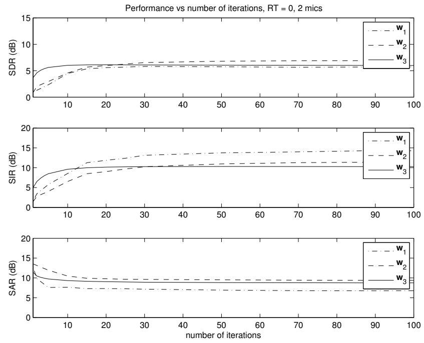

Fig. 6. Separation using two microphones: average performance as a function of the number of EM iterations.

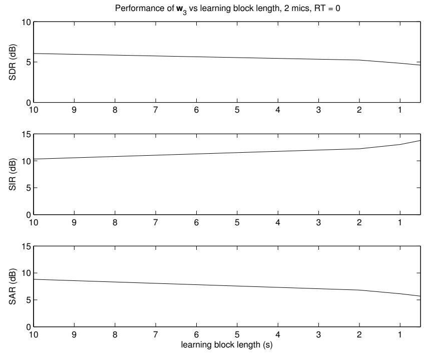

Fig. 7. Average performance of $\mathbf{w}_3$ vs learning block length in seconds.

ers in the anechoic case as a function of the number of EM iterations. The microphone array used has two microphones with a 5 cm microphone spacing. The non-linear beamformer $\mathbf{w}_3$ defined in (22) require less than 20 iterations to converge, whereas the other two non-linear beamformers require more iterations to converge (about 100 iterations).

3) Effect of the Learning Block Size: The EM algorithm used in our experiments is a batch learning algorithm. We studied the effect of varying the size of learning data on the performance of the non-linear beamformers. Fig. 7 shows the average performance at the output of the non-linear beamformer $\mathbf{w}_3$ defined in (22) in the anechoic case as a function of the EM learning block length. The performance is fairly consistent even when using shorter learning blocks. Note that the FMV algorithm can be considered as a special case of the non-linear beamformer $\mathbf{w}_3$, with $k_x = 1$ and very short learning blocks (≈ 100 ms).

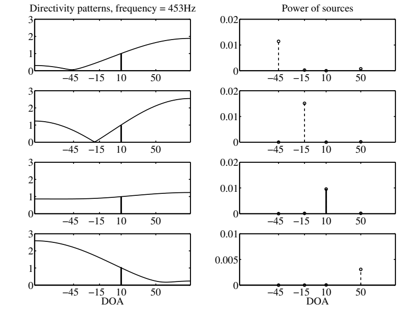

Fig. 8. Examples of directivity patterns at 453 Hz, when the desired source is at angle 10°, and the interfering sources are at $\{-45, -15, 50\}°$. Left column: directivity patterns. Right column: power of sources.

E. Directivity Patterns

Fig. 8 shows four examples of directivity patterns for the non-linear beamformer $\mathbf{w}_3$ of equation (22) in the anechoic case. The directivity patterns are defined as the magnitude of the response of the beamformer at frequency f for a far-field signal coming from direction Φ:

$$\mathrm{D}(f,\Phi) = \left| \sum_{j=1}^{N} w_j(f).e^{\iota 2\pi f(j-1)dc^{-1}\sin\Phi} \right| \qquad (29)$$

In this experiment, the desired source was at an angle of 10°, and the interfering sources at $\{-45, -15, 50\}°$. The microphone array used has two microphones with a 5 cm microphone spacing. The four examples are at four different time frames at the frequency of 453 Hz. In the first example (first row), the desired source and the interferer at angle $-45°$ were active. In the second example (second row), the interferer at angle $-15°$ was active. In the third example (third row), the desired source was active, and in the fourth example (fourth row), the interferer at angle 50° was active. The non-linear beamformer effectively nullifies the active interferer while having a distortionless response in the direction of the desired source.

F. Effect Of DOA Offset

In a typical application, the DOA of the desired source is scanned across a region of interest in space. The desired signal can arrive from a different direction than that assumed. We tested the effect of the mismatch between the assumed DOA of the desired source and the true one. Fig. 9 shows the average performance at the output of the non-linear beamformers in the anechoic case as a function of the DOA offset. The non-linear beamformers appear to be robust to small DOA offsets.

G. Effect Of Reverberation

Fig. 10 shows the average performance as a function of the room reverberation time when four sources are operating,

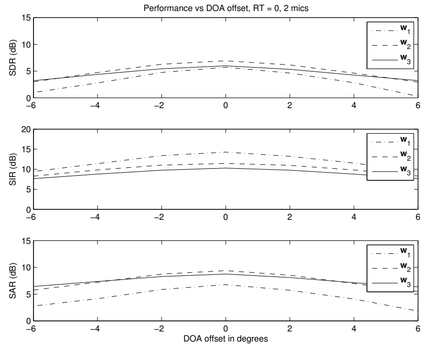

Fig. 9. Separation using two microphones: average performance as a function of DOA offset.

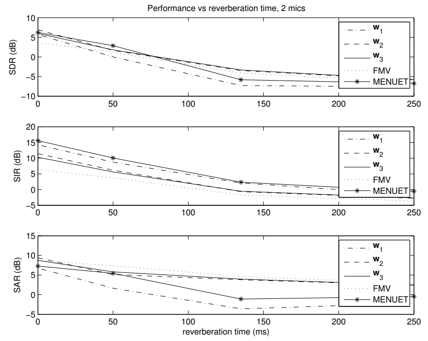

Fig. 10. Separation using two microphones: average performance as a function of reverberation time.

and the microphone array used has two microphones with a 5 cm microphone spacing. $k_x = 15$ was used in the beamformer defined in (22), and $k_s = 2, k_v = 15$ was used in the two other beamformers. We compared the performance of the three non-linear beamformers with the performance of the MENUET and FMV algorithms. A STFT of frame size 1024 samples is used. In the FMV algorithm, a small step size of 16 samples is required, while a step size of 256 samples is sufficient in the non-linear beamformers. The MENUET algorithm and the mixture of MMSE beamformers ($\mathbf{w}_1$) gives a high SIR, but suffers from a very low SAR at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ of equations (16) and (22) respectively have significantly lower artifacts at higher reverberation times.

Fig. 11 shows the average performance as a function of the room reverberation time when four sources are operating, and the microphone array used has three microphones with a 2.5 cm microphone spacing. We compared the performance of the three non-linear beamformers with the performance of the MENUET and FMV algorithms. $k_x = 5$ was used in the beamformer of equation (22), and $k_s = 2, k_v = 5$ was used in the two other beamformers. The performance of non-linear beamformers improved with the addition of the third microphone.

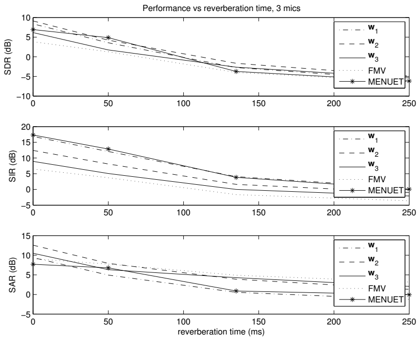

Fig. 11. Separation using three microphones: average performance as a function of reverberation time.

TABLE II
AVERAGE PERFORMANCE USING REAL LIFE RECORDINGS IN A ROOM WITH 810 MS REVERBERATION TIME.

	SDR	SIR	SAR
$\mathbf{w}_1$	0.3	3.8	4.5
$\mathbf{w}_2$	-1.7	-0.5	7.9
$\mathbf{w}_3$	-1.8	-0.3	7.0
FMV	-2.1	-0.3	5.7
MENUET	-0.4	3.3	3.6

H. Live Recordings

In order to illustrate the performance of the non-linear beamformers in real life recordings, multichannel recordings of several speech sources were recorded in a room with a reverberation time of 810 ms. The dimensions of the room and the positions of the microphones and the sources are illustrated in Figure 2. The microphone array has two microphones with spacing $d = 7$ cm. We use the same speech files used in the simulations (five mixtures of male sources, and five mixtures of female sources). The number of the sources in each mixture was four. The desired source was placed 30 cm away from the microphone array, while the interferers were placed in a semi-circle of radius 1.5 m around the microphone arrays at angles $\phi = \{-60, -30, 50\}^\circ$. We compared the three non-linear beamformers with the FMV and MENUET algorithms. The SIR and SAR values were averaged over all the mixtures. Table II shows the results.

Due to the high reverberation times, all of the methods suffer from low SIR values, but they all afford SIR improvements over the input mixture (the mixture SIR is –4.7 dB). The non-linear beamformer $\mathbf{w}_2$ shows the highest SAR. The non-linear beamformer $\mathbf{w}_1$ has the highest SIR and SDR performance, and also achieves better SAR than MENUET which had the

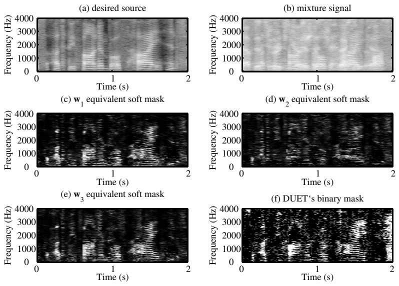

Fig. 12. (a) Spectrogram of a desired signal. (b) Spectrogram of a mixture of 4 sources. (c-f) t-f masks.

TABLE III
COMPUTATIONAL TIME FOR THE PROPOSED METHODS.

	time (s)	parameters
$\mathbf{w}_1$ (2 mics)	498	$k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_2$ (2 mics)	498	$k_s = 2$, $k_v = 15$, 100 iterations
$\mathbf{w}_3$ (2 mics)	87	$k_x = 15$, 50 iterations
$\mathbf{w}_1$ (3 mics)	193	$k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_2$ (3 mics)	193	$k_s = 2$, $k_v = 5$, 100 iterations
$\mathbf{w}_3$ (3 mics)	14	$k_x = 5$, 20 iterations

second best SIR.

I. *Time-Frequency Masks*

To understand how the various beamformers are achieving their signal enhancement, we can look at equivalent time-frequency masks for each algorithm. Fig. 12 compares the equivalent mask of the three non-linear beamformers with the time-frequency mask of DUET on an example mixture. The equivalent mask was computed at each time-frequency point as the ratio of the energy of the desired signal estimate to the energy of the observed mixture. The non-linear beamformer's approach results in a soft decision mask for the observed signal.

J. *Computational Time*

In this subsection, we report the time it took for our Matlab implementation of the non-linear beamformers to run on 2.5 GHz CPU. The time reported is for the extraction of one 10 s speech source. We used the same design parameters used in subsection V-G. Table III shows the results.

We can see that the beamformer $\mathbf{w}_3$ took less time than the other two beamformers. Beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ have similar computational time because they use the same learning algorithm (Appendix A). The computational time for the three microphones case is lower than that of the two microphones case. This is because the performance of the beamformers using the three microphones array peak using fewer Gaussian components than the two microphone case.

VI. CONCLUSION

Frequency-domain non-linear mixture of beamformers were introduced and applied to the extraction of a desired speech source from a known direction in underdetermined speech mixtures. The system model assumes an anechoic desired source signal, but no assumptions are made about the interferers, which can be of any nature such as point sources, spatial extended sources, diffuse sources, or a combination of them. The beamformers are derived assuming non-Gaussian interference signals modeled using a mixture of Gaussians distribution. This estimator introduces additional degrees of freedom to the beamformer by exploiting the super-Gaussianity (sparsity) of the interferers and dynamically finds suitable directivity patterns in order to reduce active interfering signals.

The non-linear beamformers require the location of the target speech source to be known or estimated in advance, but they have the following advantages:

- No need to know - or estimate - the number of interfering sources.
- Can be applied to underdetermined speech mixtures.
- The number of components in the GMM model controls the flexibility of the model. We did not incur overfitting in our experiments, therefore the number of GMM components can be used to trade-off complexity with performance. When using a larger number of microphones, the performance peaks with a small number of GMM components.
- Can be applied to microphone arrays with two or more microphones.
- Robust to small errors in the desired source DOA.

While one could impose models that are coupled across frequency to represent spectral patterns, we want to avoid that to keep the model as general as possible. This allows a close match to the actual properties of the observed signals and avoids the effect of microphone and channel variability which can cause a mismatch with the prior training conditions [23]. With our GMM model, we are aiming to impose as little structure on the source and interference models as possible. However, in future work, we would like to investigate the effect of source specific models.

The non-linear beamformers have been tested and evaluated on underdetermined speech mixtures. It was shown that the non-linear beamformer $\mathbf{w}_1$ defined in (15) gives better interference rejection at the expense of higher artifacts, especially at higher reverberation times. The non-linear beamformers $\mathbf{w}_2$ and $\mathbf{w}_3$ defined in (22) and (16) are distortionless beamformers (constant gain in the look-direction), and have significantly lower artifacts at higher reverberation times.

In terms of computational complexity, non-linear beamformer $\mathbf{w}_3$ employs a simpler learning algorithm and requires fewer iterations than non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$. Furthermore, the model learning for non-linear beamformer $\mathbf{w}_3$ is independent of the location of the desired source, which makes this non-linear beamformer suitable in applications where scanning for the source direction is needed.

In our current implementation, the EM algorithm used is in a batch learning mode. In section V.D.3, we studied the effect of using short blocks of data. The batch mode with short blocks of data can be used in applications where short delays are permissible, such as in human-computer interaction or surveillance. However, it is not appropriate for real-time applications. In these applications, online model learning is essential [32]. The online model learning should have a forgetting factor, and a mechanism for adding, deleting, and reassigning Gaussians to handle changes in the environment [33].

In the future, we would like to investigate the use of other linearly constrained minimum variance (LCMV) beamformers and Bayesian beamformers that are robust to DOA uncertainty [34] in the mixture of beamformers framework. We would also like to investigate the use of other filter banks instead of the STFT, such as auditory or constant-Q filter banks [35]. Through this, we aim to improve the performance of the beamformers at higher reverberation times.

Appendix
Derivation Of The EM Algorithm

Using the EM algorithm, we can estimate the model density parameters from a set of observations $D = \{\mathbf{x}(n) : 1 \leq n \leq \eta\}$. The EM algorithm is used to find a maximum likelihood estimate of parameters in probabilistic models with latent variables (incomplete data problems). In our case, $\mathbf{x}$ is the observed (or incomplete) data, and the latent variables are the state sequence of the Gaussian mixtures that indicate which Gaussian components are responsible for $\mathbf{x}(n)$. In EM terminology, the complete data is composed of both the observed data and the latent variables. The EM algorithm is an iterative algorithm with two steps: (1) an expectation step (E-step), and (2) a maximization step (M-step). In the E-step, we calculate the conditional expectation of the complete data log likelihood. The expectation is taken with respect to the conditional probability of the hidden states, given the observed data and the parameter values obtained in the previous iteration. In the M-step, the new estimates of the parameters are calculated to maximize the conditional expectation of the complete data log likelihood.

A. Learning Interference And Desired Source Parameters

In this section, the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$ of the interference $\mathbf{v}$ and desired source s are estimated using the EM algorithm. These parameters are required for the non-linear beamformers $\mathbf{w}_1$ and $\mathbf{w}_2$ of equations (15) and (16). Let us define a complete data set $D_c = \{\mathbf{x}, s, q_s, q_v\}$ composed of both the observed and the latent data. If we were to actually have such a complete data set, we could define its log likelihood as:

$$\begin{aligned} l_c(\theta|D_c) &= \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), s(n), q_s(n), q_v(n)|\theta) \\ &= \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), s(n), q_s(n), q_v(n)|\theta) \quad (30) \end{aligned}$$

Given an initial value θ^0, the EM algorithm performs the following steps at each iteration l:

E-step:: In the E-step, we compute the expectation of the complete data log likelihood:

$$\begin{aligned} Q(\theta, \theta^{l-1}) &= \sum_{q_s=1}^{k_s} \sum_{q_v=1}^{k_v} \int ds\, p\left(s, q_s, q_v|\mathbf{x}, \theta^{l-1}\right) . \\ & \ln p(\mathbf{x}, s, q_s, q_v|\theta) \quad (31) \end{aligned}$$

In the Gaussian mixture problem, this simply reduces to calculating $p\left(q_s, q_v|\mathbf{x}, \theta^{l-1}\right)$, the posterior probability of the latent variables, given the observed data and the parameters obtained in the previous iteration:

$$\begin{aligned} \tau^{(l)}_{q_s,q_v} &= p\left(q_s, q_v|\mathbf{x}, \theta^{(l-1)}\right) \\ &= \frac{p\left(q_s, q_v, \mathbf{x}|\theta^{(l-1)}\right)}{p\left(\mathbf{x}|\theta^{(l-1)}\right)} \\ &= \frac{c^{(l-1)}_{s,q_s} c^{(l-1)}_{v,q_v} p\left(\mathbf{x}|q_s, q_v, \theta^{(l-1)}\right)}{\sum_{q'_s=1}^{k_s} \sum_{q'_v=1}^{k_v} c^{(l-1)}_{s,q'_s} c^{(l-1)}_{v,q'_v} p\left(\mathbf{x}|q'_s, q'_v, \theta^{(l-1)}\right)} \end{aligned} \quad (32)$$

where

$$\begin{aligned} p(\mathbf{x}|q_s, q_v) &= \int p(\mathbf{x}, s|q_s, q_v)\ ds \\ &= \int p(\mathbf{x}|s, q_v)\ p(s|q_s)\ ds \\ &= \int \mathbb{N}\left(\mathbf{x} - \mathbf{a}s, \mathbf{R}_{v,q_v}\right)\ \mathbb{N}\left(s, \sigma^2_{s,q_s}\right)\ ds \\ &= \mathbb{N}\left(\mathbf{x}, \mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right) \quad (33) \end{aligned}$$

Moreover, we evaluate the conditional mean and variance of the desired source given both the observed mixture and the hidden states, which are denoted by $\langle s|\mathbf{x}(n), q_s, q_v\rangle$ and $\langle ss^*|\mathbf{x}(n), q_s, q_v\rangle$ respectively. Given the hidden states and the mixture, the conditional probability of s is Gaussian:

$$\begin{aligned} p(s|\mathbf{x}, q_s, q_v) &= \frac{p(\mathbf{x}, s, q_s, q_v)}{p(\mathbf{x}, q_s, q_v)} \\ &= \frac{p(s|q_s)\, p(\mathbf{x}|s, q_v)\, p(q_s)\, p(q_v)}{p(\mathbf{x}|q_s, q_v)\, p(q_s)\, p(q_v)} \\ &= \frac{\mathbb{N}(s, \sigma^2_{s,q_s})\ \mathbb{N}(\mathbf{x} - \mathbf{a}s, \mathbf{R}_{v,q_v})}{\mathbb{N}\left(\mathbf{x}, \mathbf{R}_{v,q_v} + \sigma^2_{s,q_s}\mathbf{a}\mathbf{a}^H\right)} \\ &= \mathbb{N}\left(s - \alpha_{q_s,q_v},\ \beta_{q_s,q_v}\right) \quad (34) \end{aligned}$$

where

$$\alpha_{q_s,q_v} = \left(\sigma^{-2}_{s,q_s} + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1} \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{x} \quad (35)$$

$$\beta_{q_s,q_v} = \left(\sigma^{-2}_{s,q_s} + \mathbf{a}^H\mathbf{R}^{-1}_{v,q_v}\mathbf{a}\right)^{-1} \quad (36)$$

M-step:: In the M-step, we maximize the expected complete log likelihood with respect to the parameters $\theta = \{\theta_s, \theta_v\} = \{c_{s,q_s}, \sigma^2_{s,q_s}, c_{v,q_v}, \mathbf{R}_{v,q_v} : 1 \leq q_s \leq k_s, 1 \leq q_v \leq k_v\}$. This can be done by taking derivatives with respect to θ and setting them to be equal to zero (under the constraints $\sum_{q_s=1}^{k_s} c_{s,q_s} = 1$ and $\sum_{q_v=1}^{k_v} c_{v,q_v} = 1$). This results in the following update rules:

$$c^{(l)}_{v,q_v} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau^{(l)}_{q_s,q_v}(n) \quad (37)$$

$$c_{s,q_s}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n) \tag{38}$$

$$\sigma_{s,q_s}^{2^{(l)}} = \frac{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n) \, \langle ss^* | \mathbf{x}(n), q_s, q_v \rangle}{\sum_{n=1}^{\eta} \sum_{q_v=1}^{k_v} \tau_{q_s,q_v}^{(l)}(n)} \tag{39}$$

$$\mathbf{R}_{v,q_v}^{(l)} = \frac{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n) \, \Lambda_{q_s,q_v}(n)}{\sum_{n=1}^{\eta} \sum_{q_s=1}^{k_s} \tau_{q_s,q_v}^{(l)}(n)} \tag{40}$$

where

$$\begin{aligned} \Lambda_{q_s,q_v}(n) \;&=\; \mathbf{x}(n)\mathbf{x}(n)^H - \mathbf{x}(n) \langle s^* | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{a}^H \\ &\quad -\mathbf{a} \langle s | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{x}(n)^H \\ &\quad +\mathbf{a} \langle ss^* | \mathbf{x}(n), q_s, q_v \rangle \, \mathbf{a}^H \end{aligned} \tag{41}$$

In this model, there is an ambiguity in associating variance between the desired source and the interference. It is possible to incorporate some of the source signal into the interference. To avoid this, updating the desired source component variances is not performed in the first few iterations. This prevents the source components shrinking to zero variance.

B. Learning Observed Mixture Parameters

In this section, the parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$ of the observed mixture $\mathbf{x}$ are estimated using the EM algorithm. These parameters are required for the non-linear beamformer $\mathbf{w}_3$ of equation (22). Let us define a complete data set $D_c = \{\mathbf{x}, q_x\}$ composed of both the observed and the latent data. If we were to actually have such a complete data set, we define its log likelihood as:

$$\begin{aligned} l_c(\theta_x | D_c) \;&=\; \ln \prod_{n=1}^{\eta} p(\mathbf{x}(n), q_x(n) | \theta_x) \\ &=\; \sum_{n=1}^{\eta} \ln p(\mathbf{x}(n), q_x(n) | \theta_x) \end{aligned} \tag{42}$$

The EM algorithm may be executed as follows:

E-step:: In the E-step, we compute the expectation of the complete data log likelihood:

$$Q(\theta_x, \theta_x^{l-1}) = \sum_{q_x=1}^{k_x} p\left(q_x | \mathbf{x}, \theta_x^{l-1}\right) \ln p(\mathbf{x}, q_x | \theta_x) \tag{43}$$

This reduces to calculating $p\left(q_x | \mathbf{x}, \theta_x^{l-1}\right)$, the posterior probability of the latent variables given the observed data and the current estimates of the parameters.

$$\begin{aligned} \tau_{q_x}^{(l)} \;&=\; p\left(q_x | \mathbf{x}, \theta_x^{(l-1)}\right) \\ &=\; \frac{p\left(q_x, \mathbf{x} | \theta_x^{(l-1)}\right)}{p\left(\mathbf{x} | \theta_x^{(l-1)}\right)} \\ &=\; \frac{p\left(q_x | \theta_x^{(l-1)}\right) p\left(\mathbf{x} | q_x, \theta_x^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} p\left(q_x' | \theta_x^{(l-1)}\right) p\left(\mathbf{x} | q_x', \theta_x^{(l-1)}\right)} \\ &=\; \frac{c_{q_x}^{(l-1)} \, \mathbb{N}\left(\mathbf{x} | \mathbf{R}_{x,q_x}^{(l-1)}\right)}{\sum_{q_x'=1}^{k_x} c_{q_x'}^{(l-1)} \, \mathbb{N}\left(\mathbf{x} | \mathbf{R}_{x,q_x'}^{(l-1)}\right)} \end{aligned} \tag{44}$$

M-step:: In the M-step, we maximize the expected complete log likelihood with respect to the parameters $\theta_x = \{c_{x,q_x}, \mathbf{R}_{x,q_x} : 1 \leq q_x \leq k_x)$. This can be done by taking derivatives with respect to θ_x and setting them to be equal to zero, while also including a Lagrangian term to account for the constraint that $\sum_{q_x=1}^{k_x} c_{q_x} = 1$. This results in the following update rules:

$$\mathbf{R}_{x,q_x}^{(l)} = \frac{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \, \mathbf{x}(n) \, \mathbf{x}(n)^H}{\sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n)} \tag{45}$$

$$c_{x,q_x}^{(l)} = \frac{1}{\eta} \sum_{n=1}^{\eta} \tau_{q_x}^{(l)}(n) \tag{46}$$

Acknowledgment

The authors wish to thank S. Rickard for providing the implementation of the DUET algorithm.

References

[1] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

[2] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, ser. Signals and Communication Technology. Springer, 2005.

[3] S. Douglas and M. Gupta, "Convolutive blind source separation for audio signals," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[4] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[5] H. Attias, "Source separation with a sensor array using graphical models and subband filtering," in *Advances in Neural Inf. Process. Syst. (NIPS'02)*, 2002, pp. 1229–1236.

[6] ——, "New EM algorithms for source separation and deconvolution with a microphone array," in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP'03)*, vol. 5, April 2003, pp. V–297–300 vol.5.

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.

[8] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[9] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[10] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O'Brien, Jr., B. Wheeler, and A. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *J. Acoust. Soc. Am.*, vol. 115, no. 1, pp. 379–391, 2004.

[11] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul 2004.

[12] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T.-W. Lee, and H. Sawada, Eds. Springer Netherlands, 2007.

[13] Z. El-Chami, A. D. Pham, C. Servière, and A. Guerin, "A new model-based underdetermined speech separation," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC'08)*, Seattle, USA, Sep. 2008.

[14] D.-T. Pham, Z. El-Chami, A. Gurin, and C. Servire, "Modeling the short time Fourier transform ratio and application to underdetermined audio source separation." in *Proc. Int. Conf. Ind. Compon. Anal. Signal Separation (ICA'09).* Paraty, Brazil: Springer, 2009, pp. 98–105.

[15] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Inf. Process. Syst. (NIPS 19)*, B. Schlkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2006, pp. 953–960.

[16] M. Mandel and D. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, NY, USA, Oct. 2007, pp. 275–278.

[17] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch BSS using the EM algorithm in reverberant environment," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'07)*, NY, USA, Oct. 2007, pp. 147–150.

[18] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. Intl. Workshop Acoust. Echo and Noise Control (IWAENC'05)*, Sep. 2005, pp. 117–120.

[19] ——, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.

[20] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[21] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[22] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA'05)*, 2005, pp. 90–93.

[23] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1564–1578, Jul. 2007.

[24] H. Attias and L. Deng, "A new approach to speech enhancement by a microphone array using em and mixture models," in *Int.l Conf. on Spoken Lang. Process., Denver CO,*, 2002.

[25] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[26] M. A. Dmour and M. E. Davies, "An approach to under-determined speech separation based on a non-linear mixture of beamformers," in *Proc. Eur. Signal Process. Conf. (EUSIPCO'09)*, Glasgow, UK, Aug. 2009.

[27] H. Attias, "Independent factor analysis," *Neural Comput.*, vol. 11, no. 4, pp. 803–851, 1999.

[28] M. A. Dmour and M. E. Davies, "Under-determined speech separation using GMM-based non-linear beamforming," in *Proc. Eur. Signal Process. Conf. (EUSIPCO'08)*, Lausanne, Switzerland, Aug. 2008.

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

[30] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *DARPA Workshop on Speech Recognition*, 1986.

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[32] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. Roy. Stat. Soc.*, vol. 46, no. 2, pp. 257–267, 1984.

[33] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 27, no. 5, pp. 827–832, May 2005.

[34] K. Bell, Y. Ephraim, and H. Van Trees, "A Bayesian approach to robust adaptive beamforming," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 386–398, Feb. 2000.

[35] J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proc. Int. Conf. Inform., Commun, Signal Process. (ICICS'05)*, Bangkok, Thailand, Dec. 2005, pp. 1466–1470.

Mohammad A. Dmour (S'04) received the B.Sc. in electrical engineering from the University of Jordan, Amman, Jordan, in 2005 and the M.Sc. degree in signal processing and communications (with distinction) from the University of Edinburgh, Edinburgh, U.K. in 2006. He is currently working towards the Ph.D. degree at the University of Edinburgh. His current research interests include audio source separation and speech enhancement. Mr. Dmour was awarded the Wolfson Microelectronics scholarship for the pursuit of his PhD at the University of Edinburgh, and was presented with the class medal upon the culmination of his M.Sc. degree in Signal Processing and Communications.

Mike Davies (M'00) received the B.A. (Hons.) degree in engineering from Cambridge University, Cambridge, U.K., in 1989 and the Ph.D. degree in nonlinear dynamics and signal processing from University College London, London (UCL), U.K., in 1993. Mike Davies was awarded a Royal Society Research Fellowship in 1993 and was an Associate Editor for IEEE Transactions in Speech, Language and Audio Processing, 2003-2007. He currently holds the Jeffrey Collins SHEFC funded chair in Signal and Image Processing at the University of Edinburgh. His current research interests include: sparse approximation, compressed sensing and their applications.

References

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

[2] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[3] J. Benesty, S. Makino, and J. Chen, eds., *Speech Enhancement*. Signals and Communication Technology, Springer, 2005.

[4] S. Makino, T.-W. Lee, and H. Sawada, eds., *Blind Speech Separation*. Signals and Communication Technology, Springer, 2007.

[5] E. Vincent and Y. Deville, “Audio applications,” in *Handbook of Blind Source Separation* (P. Comon and C. Jutten, eds.), Elsevier, 2010.

[6] A. Bregman, *Auditory Scene Analysis*. MIT Press, 2nd ed., 1990.

[7] G. Brown and D. Wang, “Separation of speech by computational auditory scene analysis,” in *Speech Enhancement*, pp. 371–402, Springer Berlin Heidelberg, 2005.

[8] M. Weintraub, *A theory and computational model of auditory monaural sound separation (stream, speech enhancement, selective attention, pitch perception, noise cancellation)*. PhD thesis, Stanford, CA, USA, 1985.

[9] A. Van der Kouwe, D. Wang, and G. Brown, “A comparison of auditory and blind separation techniques for speech segregation,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 189–195, Mar. 2001.

[10] G. Brown and M. Cooke, “Computational auditory scene analysis,” *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[11] T. Irino and R. Patterson, “A dynamic compressive gammachirp auditory filterbank,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 2222–2232, Nov. 2006.

[12] R. Patterson, “Auditory images: How complex sounds are represented in the auditory system,” *The Journal of the Acoustical Society of Japan*, vol. (E) 21, pp. 183–190, 2000.

[13] B. Van Veen and K. Buckley, “Beamforming: a versatile approach to spatial filtering,” *IEEE ASSP Magazine*, vol. 5, pp. 4–24, Apr. 1988.

[14] M. Brandstein and D. Ward, eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Digital Signal Processing series, Springer, 2001.

[15] P. Comon, “Independent component analysis, a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[16] J.-F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE, special issue on blind identification and estimation*, vol. 86, pp. 2009–2025, Oct. 1998.

[17] S. Douglas and M. Gupta, "Convolutive blind source separation for audio signals," in *Blind Speech Separation* (S. Makino, T.-W. Lee, and H. Sawada, eds.), Springer Netherlands, 2007.

[18] H. Sawada, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Blind Speech Separation* (S. Makino, T.-W. Lee, and H. Sawada, eds.), Springer Netherlands, 2007.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, Jul. 2004.

[20] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation* (S. Makino, T.-W. Lee, and H. Sawada, eds.), Springer Netherlands, 2007.

[21] Z. El-Chami, D.-T. Pham, C. Servière, and A. Guerin, "A new model-based underdetermined speech separation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC'08)*, (Seattle, USA), Sep. 2008.

[22] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.

[23] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, pp. 90–93, 2005.

[24] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1564–1578, Jul. 2007.

[25] S. Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pp. 232–237, 1998.

[26] M. Beal, H. Attias, and N. Jojic., "Audio-video sensor fusion with probabilistic graphical models," in *European Conference on Computer Vision (ECCV'02)*, pp. 736–752, 2002.

[27] N. Mitianoudis and M. Davies, "Audio source separation: Solutions and problems," *International Journal of Adaptive Control and Signal Processing*, vol. 18, pp. 299–314, Apr. 2004.

[28] H. Attias, "New EM algorithms for source separation and deconvolution with a microphone array," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 5, pp. 297–300, Apr. 2003.

[29] H. Attias and L. Deng, "A new approach to speech enhancement by a microphone array using EM and mixture models," in *International Conference on Spoken Language Processing (ICSLP'02)*, (Denver CO), 2002.

[30] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[31] P. Loizou, *Speech Enhancement: Theory and Practice*. Taylor and Francis Group, 2007.

[32] S. Mallat, *A Wavelet Tour of Signal Processing, (Wavelet Analysis & Its Applications)*. Academic Press, 2nd ed., Sep. 1999.

[33] R. Patterson, M. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform," *The Journal of the Acoustical Society of America*, vol. 98, pp. 1890–1894, Oct. 1995.

[34] J. Burred and T. Sikora, "Comparison of frequency-warped representations for source separation of stereo mixtures," in *121st AES Convention*, Oct. 2006.

[35] J. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, pp. 425–434, 1991.

[36] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," tech. rep., Apple Computer Co., 1993.

[37] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, Oct. 2001.

[38] M. Davies, "Audio source separation," in *Mathematics in Signal Processing V*, Oxford University Press, 2002.

[39] W. Kellermann, "Beamforming for speech and audio signals," in *Handbook of Signal Processing in Acoustics* (D. Havelock, S. Kuwano, and M. Vorländer, eds.), pp. 691–702, Springer New York, 2008.

[40] I. Mccowan, D. Moore, and S. Sridharan, "Near-field adaptive beamformer for robust speech recognition," *Digital Signal Processing*, vol. 12, no. 1, pp. 87–106, 2001.

[41] E. Vincent, M. Jafari, S. Abdallah, M. Plumbley, and M. Davies, "Model-based audio source separation," Tech. Rep. C4DM-TR-05-01, Queen Mary University of London, 2006.

[42] L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*. John Wiley & Sons, Inc, 4th ed., 2000.

[43] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *108th AES Convention*, pp. 18–22, 2000.

[44] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *The Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, 2001.

[45] M. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6, pp. 1187–1188, 1965.

[46] I. 3382:1997, *Acoustics. Measurement of the reverberation time of rooms with reference to other acoustical parameters*. International Organization for Standardization, Geneva, Switzerland, 1997.

[47] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *International Conference on Language Resources & Evaluation (LREC'00)*, (Athen), 2000.

[48] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, 2007.

[49] E. De Geest and R. Garcea, "Simulation of room transmission functions using a triangular beam tracing computer model," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'95)*, pp. 253 –256, Oct. 1995.

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

[51] J. Borish, "Extension of the image model to arbitrary polyhedra," *The Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–183, 1984.

[52] H. Van Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[53] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1462–1469, Jul. 2006.

[54] C. Fevotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide, revision 2.0," Tech. Rep. 1706, IRISA, Rennes, France, Apr. 2005.

[55] T. Thiede, W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ-the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3–29, 2000.

[56] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," tech. rep., McGill University, 2003.

[57] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *International Conference on Independent Component Analysis and Signal Separation (ICA'07)*, pp. 454–461, 2007.

[58] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'07)*, 4892, pp. 295–305, 2008.

[59] J. Paulus and T. Virtanen, "Drum transcription with nonnegative spectrogram factorization," in *European Signal Processing Conference (EUSIPCO'05)*, 2005.

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

[61] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[62] M. Lockwood, D. Jones, R. Bilger, C. Lansing, W. O’Brien, Jr., B. Wheeler, and A. Feng, “Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms,” *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 379–391, 2004.

[63] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proceedings of the IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[64] A. Bell and T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[65] J.-F. Cardoso, “Infomax and maximum likelihood for blind source separation,” *IEEE Signal Processing Letters*, vol. 4, pp. 112–114, Apr. 1997.

[66] S. Amari, A. Cichocki, and H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems (NIPS’95)*, pp. 757–763, 1995.

[67] A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[68] E. Bingham and A. Hyvärinen, “A fast fixed-point algorithm for independent component analysis of complex valued signals.,” *International Journal of Neural Systems*, vol. 10, pp. 1–8, Feb. 2000.

[69] A. Belouchrani, K. Abed-Meraim, J. Cardoso, and E. Moulines, “A blind source separation technique using second-order statistics,” *IEEE Transactions on Signal Processing*, vol. 45, pp. 434–444, 1997.

[70] K. Matsuoka, M. Ohya, and M. Kawamoto, “A neural net for blind separation of nonstationary signals,” *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[71] J. Eriksson and V. Koivunen, “Identifiability and separability of linear ICA models revisited,” in *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA’03)*, (Nara, Japan), pp. 23–27, 2003.

[72] M. Davies, “Identifiability issues in noisy ICA,” *IEEE Signal Processing Letters*, vol. 11, pp. 470–473, May 2004.

[73] A. Hyvärinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[74] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[75] T.-W. Lee, A. Bell, and R. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems (NIPS'97)* (M. Mozer, M. Jordan, and T. Petsche, eds.), vol. 9, pp. 758–764, The MIT Press, 1997.

[76] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[77] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 352–362, Sep. 2002.

[78] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*, vol. 5, pp. 3140–3143, Jun. 2000.

[79] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, vol. 1, 2002.

[80] N. Mitianoudis and M. Davies, "Using beamforming in the audio source separation problem," in *International Symposium on Signal Processing and Its Applications (ISSPA'03)*, vol. 2, pp. 89–92, Jul. 2003.

[81] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust approach to the permutation problem of frequency-domain blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, vol. 5, Apr. 2003.

[82] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, Sep. 2004.

[83] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 86, no. 4, pp. 846–858, 2003.

[84] A. Hyvärinen, "Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood," *Neural Computing Surveys*, vol. 22, pp. 49–67, Nov. 1998.

[85] M. Zibulevsky, P. Kisilev, Y. Zeevi, and B. Pearlmutter, "Blind source separation via multinode sparse representation," in *Advances in Neural Information Processing Systems (NIPS'02)*, vol. 14, pp. 1049–1056, MIT Press, 2002.

[86] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[87] T.-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, pp. 87–90, Apr. 1999.

[88] H. Attias, “Independent factor analysis,” *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[89] M. Davies and N. Mitianoudis, “Simple mixture model for sparse overcomplete ICA,” in *IEE Proceedings Vision, Image & Signal Processing*, vol. 151, pp. 35–43, Feb. 2004.

[90] M. Zibulevsky and B. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[91] P. Bofill and M. Zibulevsky, “Underdetermined blind source separation using sparse representations,” *Signal Processing*, vol. 81, pp. 2353–2362, 2001.

[92] D.-T. Pham, Z. El-Chami, A. Guérin, and C. Servière, “Modeling the short time Fourier transform ratio and application to underdetermined audio source separation.,” in *International Conference on Independent Component Analysis and Signal Separation (ICA’09)*, (Paraty, Brazil), pp. 98–105, Springer, 2009.

[93] M. Mandel, D. Ellis, and T. Jebara, “An EM algorithm for localizing multiple sound sources in reverberant environments,” in *Advances in Neural Information Processing Systems (NIPS’06)* (B. Schölkopf, J. Platt, and T. Hoffman, eds.), pp. 953–960, MIT Press, 2006.

[94] M. Mandel and D. Ellis, “EM localization and separation using interaural level and phase cues,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’07)*, (NY, USA), pp. 275–278, Oct. 2007.

[95] Y. Izumi, N. Ono, and S. Sagayama, “Sparseness-based 2ch BSS using the EM algorithm in reverberant environment,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’07)*, (NY, USA), pp. 147–150, Oct. 2007.

[96] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’00)*, vol. 5, pp. 2985–2988, Jun. 2000.

[97] S. Rickard and Z. Yilmaz, “On the approximate W-disjoint orthogonality of speech,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’02)*, vol. 1, pp. 529–532, 2002.

[98] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *International Workshop on Acoustic Echo and Noise Control (IWAENC’05)*, pp. 117–120, Sep. 2005.

[99] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors,” *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[100] R. Duda, P. Hart, and D. Stork, *Pattern Classification.* Wiley-Interscience, 2nd ed., 2000.

[101] G. Hu and D.-L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’02)*, vol. 1, pp. I–553–I–556 vol.1, 2002.

[102] G. Hu and D.-L. Wang, “Separation of fricatives and affricates,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’05)*, vol. 1, pp. 1101–1104, 2005.

[103] L. Atlas and C. Janssen, “Coherent modulation spectral filtering for single-channel music source separation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’05)*, vol. 4, pp. 461–464, Mar. 2005.

[104] N. Roman, D. Wang, and G. Brown, “Speech segregation based on sound localization,” *The Journal of the Acoustical Society of America*, vol. 114, pp. 2236–2252, 2003.

[105] D.-L. Wang and G. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, May 1999.

[106] S. Roweis, “One microphone source separation,” in *Advances in Neural Information Processing Systems (NIPS’00)*, pp. 793–799, MIT Press, 2000.

[107] D. Kolossa and R. Orglmeister, “Nonlinear postprocessing for blind speech separation,” in *International Conference on Independent Component Analysis and Blind Signal Separation (ICA’04)*, pp. 832–839, 2004.

[108] M. Cobos and J. Lopez, “Improving isolation of blindly separated sources using time-frequency masking,” *IEEE Signal Processing Letters*, vol. 15, pp. 617–620, 2008.

[109] M. Dmour and M. Davies, “An approach to under-determined speech separation based on a non-linear mixture of beamformers,” in *European Signal Processing Conference (EUSIPCO’09)*, (Glasgow, UK), Aug. 2009.

[110] K. Harmanci, J. Tabrikian, and J. Krolik, “Relationships between adaptive minimum variance beamforming and optimal source localization,” *IEEE Transactions on Signal Processing*, vol. 48, pp. 1–12, Jan. 2000.

[111] R. Lorenz and S. Boyd, “Robust minimum variance beamforming,” *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1684–1696, 2005.

[112] M. Dmour and M. Davies, “Under-determined speech separation using GMM-based non-linear beamforming,” in *European Signal Processing Conference (EUSIPCO’08)*, (Lausanne, Switzerland), Aug. 2008.

[113] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, “Darpa timit acoustic-phonetic continuous speech corpus cdrom.” prepared at the National Institute of Standards and Technology (NIST), 1993.

[114] D. Titterington, “Recursive parameter estimation using incomplete data,” *Journal of the Royal Statistical Society*, vol. 46, no. 2, pp. 257–267, 1984.

[115] D.-S. Lee, “Effective Gaussian mixture learning for video background subtraction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 827–832, May 2005.

[116] S. Lang, *Linear Algebra*. New York: Springer-Verlag, 3rd ed., 1987.

[117] D. Harville, *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, 1997.

[118] G. Golub and C. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 3rd ed., 1996.

[119] J. Dattorro, *Convex Optimization and Euclidean Distance Geometry*. Meboo Publishing, 2005.

[120] J. Cermak, S. Araki, H. Sawada, and S. Makino, "Blind source separation based on beamformer array and time-frequency binary masking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, pp. 145–148, 2007.

[121] M. Baeck and U. Zölzer, "Real-time implementation of a source separation algorithm," in *International Conference on Digital Audio Effects (DAFx'03)*, 2003.

[122] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," in *IEEE Workshop on Neural Networks for Signal Processing (NNSP'02)*, pp. 465–474, Sep. 2002.

[123] J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *International Conference on Information, Communications and Signal Processing (ICICS'05)*, (Bangkok, Thailand), pp. 1466–1470, Dec. 2005.

[124] B. Moore and B. Glasberg, "A revision of Zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, pp. 335–345, 1996.

[125] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *The Journal of the Acoustical Society of America*, vol. 33, p. 248, 1961.

[126] Y. Zheng, R. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 478 – 488, 2004.