Online Source Separation in Reverberant Environments Exploiting Known Speaker Locations

Séparation de sources en ligne dans des environnements réverbérants en exploitant la localisation des sources

Thesis submitted to Loughborough University and Université de Grenoble in candidature for degrees of Doctor of Philosophy.

Jack Duncan Harris

2015

Loughborough
University

Advanced Signal
Processing Group
Loughborough Unversity

UNIVERSITÉ DE
GRENOBLE

GIPSA-Lab
Université de Grenoble

Certificate of Originality

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

.... (Signed)

.... (Candidate)

ABSTRACT

This thesis concerns blind source separation techniques using second order statistics and higher order statistics for reverberant environments. A focus of the thesis is algorithmic simplicity with a view to the algorithms being implemented in their online forms. The main challenge of blind source separation applications is to handle reverberant acoustic environments; a further complication is changes in the acoustic environment such as when human speakers physically move.

A novel time-domain method which utilises a pair of finite impulse response filters is proposed. The method of principle angles is defined which exploits a singular value decomposition for their design. The pair of filters are implemented within a generalised sidelobe canceller structure, thus the method can be considered as a beamforming method which cancels one source. An adaptive filtering stage is then employed to recover the remaining source, by exploiting the output of the beamforming stage as a noise reference.

A common approach to blind source separation is to use methods that use higher order statistics such as independent component analysis. When dealing with realistic convolutive audio and speech mixtures, processing in the frequency domain at each frequency bin is required. As a result this introduces the permutation problem, inherent in independent component analysis, across the frequency bins. Independent vector analysis directly addresses this issue by modeling the dependencies between frequency bins, namely making use of a source vector prior. An alternative source prior for real-time (online) natural gradient independent vector analysis is proposed. A Student's t probability density function is known to be more suited for speech sources, due to its heavier tails, and is incorporated into a real-time version of natural gradient independent vector analysis. The final algorithm is realised as a real-time embedded application on a floating point Texas Instruments digital signal processor platform.

Moving sources, along with reverberant environments, cause significant problems in realistic source separation systems as mixing filters become time variant. A method which employs the pair of cancellation filters, is proposed to cancel one source coupled with an online natural gradient independent vector analysis technique to improve average separation performance in the context of step-wise moving sources. This addresses 'dips' in performance when sources move. Results show the average convergence time of the performance parameters is improved.

Online methods introduced in thesis are tested using impulse responses measured in reverberant environments, demonstrating their robustness and are shown to perform better than established methods in a variety of situations.

RÉSUMÉ

Cette thèse porte sur les techniques de séparation de sources en aveugle en utilisant des statistiques de second ordre et statistiques d'ordre supérieur pour les environnements réverbérants. Un objectif de la thèse est la simplicité algorithmique en vue de l'implantation en ligne des algorithmes. Le principal défi des applications de séparation de sources aveugles est de s'occuper des environnements acoustiques de réverbération; une complication supplémentaire concerne les changements dans l'environnement acoustique lorsque les sources humaines se déplacent physiquement.

Une nouvelle méthode dans le domaine temporel qui utilise une paire de filtres à réponse impulsionnelle finie est proposée. Cette méthode, dite des angles principaux, sur une décomposition en valeurs singulières. Une paire de filtres, jouant le rôle de formation de voie, est estimée de facon à annuler une des sources. Une étape de filtrage adaptatif est ensuite utilisée pour récupérer la source restante, en exploitant la sortie de l'étage de beamforming en tant que référence de bruit.

Une approche commune de la séparation de sources aveugle est d'utiliser des méthodes fondées sur les statistiques d'ordre supérieur comme l'analyse en composantes indépendantes. Cependant, pour des mélanges convolutifs audio et vocaux de parole réalistes, la transformation dans le domaine fréquentiel pour chaque fréquence de calcul est nécessaire. Ceci introduit le problème de permutations, inhérentes à l'analyse en composantes indépendantes, pour toutes les fréquences. L'analyse en vecteurs indépendants résout directement cette question par la modélisation des dépendances entre les fréquences de calcul, à partir d'a priori sur les sources. Un algorithme de gradient naturel en temps réel est également proposé avec un autre a priori sur les sources. Cette méthode exploite la fonction de densité de probabilité de Student, connue pour être bien adaptée pour les sources de parole, en raison de queues de distribution plus lourdes. L'algorithme final est implanté en temps réel sur un processeur numérique de signal à virgule flottante de Texas Instruments.

Les sources mobiles, avec des environnements réverbérants, causent des problèmes significatifs dans les systèmes de séparation de sources réalistes car les filtres de mélange deviennent variants dans le temps. Dans ce cadre, une méthode qui utilise conjointement le principe de la paire de filtres d'annulation et le principe de l'analyse en vecteurs indépendants est proposée. Cette approche permet de limiter les baisses de performances lorsque les sources sont mobiles. Les résultats montrent également que les temps moyens de convergence des divers paramètres sont diminués.

Les méthodes en ligne qui sont introduites dans la thèse, sont testées en utilisant des réponses impulsionnelles mesurées dans des environnements réverbérants. Les résultats montrent leur robustesse et d'excellentes performances par rapport à d'autres méthodes classiques, dans plusieurs situations expérimentales.

To Mum,

who sadly wasn't here to see this
thesis come to fruition.

ACKNOWLEDGEMENTS

I consider myself very fortunate to have decided to take a module in “Fundamentals of Digital Signal Processing”, during my undergraduate degree at Loughborough University, otherwise I would never have met my future PhD supervisor, Prof. Jonathon Chambers. He made it a very easy decision to do a PhD, his guidance and input were invaluable. Without his help I don’t think I would have finished my PhD, and I can’t thank him enough for his support for the past few years. Through Jonathon I have had some amazing opportunities, including starting my PhD in Grenoble, and I met some wonderful researchers, three of whom became a part of my ‘team’ of supervisors.

I am very thankful for Prof. Christian Jutten for inviting me to the Gipsa Lab in Grenoble, a truly remarkable experience, and one for which I will always be grateful for and will remember for the rest of my life. I’d also like to thank Dr. Bertrand Rivet for his support and help especially in the first couple of years bringing me up to speed on many technical aspects, he was always very patient and very thorough in his explanations.

Lastly, but by no means least, of my supervision team, Dr. Mohsen Naqvi, I don’t think I ever exited his office without gaining a new insight to my work, and he has often ‘pushed’ my work just at the right times during the PhD.

I would also like to thank my technical partner from Dstl, Dr. Josef Kornycky, for the funding for my fourth year of study and also lending his understanding ear whenever we met.

There is countless assortment of people who I've met in both Grenoble and Loughborough, and I can't possibly name everyone without causing offence ("you know who you are" seems a bit cheesy, but true!) though I would like to mention the PhD students at Gipsa-Lab, especially members of the 'ViBs' team for their kindness and putting up with many "questions stupide", whilst learning French. I also want to thank Anna for her support (and eagle-eyed proof reading abilities) these last few months.

I think it possibly goes without saying, I'd like to thank my Mum and Dad for their many years of support. Frequently putting me before themselves in many circumstances so I could continue my education. I once asked my Mum how long I would have to go to school for after just starting primary school. She didn't have the heart to tell me it would be many years, so instead said "until you've learnt enough", I think I could say that I've truly reached that point, this thesis is dedicated to her.

STATEMENT OF ORIGINALITY

The contributions of this thesis are mainly associated with online blind source separation techniques, including so-called target cancellation techniques and independent vector analysis (IVA). The novelty of the contributions is supported by the following international journal and conference papers.

In Chapter 3 a time domain method is proposed for finding a pair of finite impulse response filters for target cancellation in the context of blind source separation, when the location of one speaker is known. The work was published in:

1. J. Harris, B. Rivet, S.M. Naqvi, J.A. Chambers and C. Jutten, ‘Principal angles approach to time-domain filter design for target cancelation’, 2014 19th International Conference on Digital Signal Processing (DSP), Hong Kong, pp.184,189, 20-23 Aug. 2014.

In Chapter 4, a real-time source separation method, online IVA, is enhanced with a new source prior with heavier tails which is more suited to speech signals:

2. J. Harris, B. Rivet, S.M. Naqvi, J.A. Chambers and C. Jutten, ‘Real-time independent vector analysis with Student’s t source

prior for convolutive speech mixtures', 2015 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, 19-24 April. 2015.

3. Y. Liang, J. Harris, G. Chen, S. M. Naqvi, C. Jutten and J.A. Chambers, 'Auxiliary function based IVA using a source prior exploiting fourth order relationships', 2013 European Signal Processing Conference (EUSIPCO), Marrakech, Morocco, 2013.

In Chapter 5, a pair of finite impulse response filters are combined with the independent analysis algorithm to aid the online separation of a step-wise moving source. A journal article is in preparation:

4. J. Harris, S.M. Naqvi, B. Rivet, J.A. Chambers and C. Jutten, 'Enhanced Independent Vector Analysis for Moving Sources Exploiting Known Speaker Locations', Journal article in preparation, 2015.

Other published work includes work into target cancellation in the time domain:

5. J. Harris, B. Rivet, S.M. Naqvi, J.A. Chambers and C. Jutten, 'Video-informed approach for enhancing audio source separation through noise source suppression', 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Southampton, pp.1-6, 22-25 Sept. 2013.

6. J. Harris, S.M. Naqvi, B. Rivet, J.A. Chambers and C. Jutten, 'Visual Based Reference for Enhanced Audio-Video Source Extraction', IMA Conference on Mathematics in Signal Processing 2012, Birmingham, Dec. 2012.

CONTENTS

Acronyms

ADC	Analogue to digital converter
BRIR	Binaural room impulse response
BSS	Blind source separation
CASA	Computational auditory scene analysis
CPP	Cocktail party problem
DFT	Discrete Fourier transform
FD-BSS	Frequency domain blind source separation
GD	Gradient descent
GSC	Generalised sidelobe canceller
HOS	Higher order statistics
ICA	Independent Component Analysis
IM	Image method for small room acoustics
IR	Impulse response
IVA	Independent vector analysis
MAC	Multiply accumulate

MSS	Mean-squared sum
NG-ICA	Natural gradient ICA
NG-IVA	Natural gradient IVA
NLMS	Normalised least mean squares
PA	Principal angles
PCA	Principal component analysis
pdf	Probability density function
RIR	Real impulse response
RMS	Root mean square
SDR	Signal-to-distortion ratio
SIR	Signal-to-interference ratio
SOS	Second order statistics
STFT	Short-time Fourier transform
SVD	Singular value decomposition

LIST OF SYMBOLS

Some frequently used notations are as follows:

Functions and operators

$\lvert\cdot\rvert$	Absolute value
$\lvert\lvert\cdot\rvert\rvert_2$	Euclidean norm
$(\cdot)^T$	Transpose operator
$(\cdot)^H$	Hermitian transpose operator
$(\cdot)^{-1}$	Inverse operator
$(\cdot)^*$	Complex conjugate operator
$det(\cdot)$	Matrix determinant operator
$\mathcal{H}(\cdot)$	Differential entropy
$\mathcal{I}(\cdot)$	Mutual information
$E[\cdot]$	Mathematical expectation
$\mathrm{proj}(\cdot)$	Matrix projection
$\varphi(\cdot)$	Non-linear score function

$*$	Convolution
$\circledast$	Circular convolution

Latin characters

G	Unmixing matrix
H	Mixing matrix
I	Identity matrix
J	Cost function value (scalar)
K	FFT length (scalar)
L	Order of an FIR filter (scalar)
M	Number of sources
N	Number of mixtures
P	Permutation matrix
Q_w	Whitening matrix
Q	Matrix with an orthonormal basis
R	Upper triangular matrix
S	Source matrix
X	Observation matrix
k	Frequency bin index
n	Time block index of a short-time Fourier transform

$\mathbf{s}$	Source signal vector
$\hat{\mathbf{s}}$	Estimated source signal vector
t	Discrete time index
$\mathbf{w}$	Cancellation filter vector
$\mathbf{x}$	Mixture signal vector

Greek characters

β	Smoothing factor
δ	Kronecker delta
ϵ	Error value
$\boldsymbol{\epsilon}$	Error vector
ζ	Noise term
η	Learning rate
Λ	Diagonal matrix
μ	Mean value
Σ	Covariance matrix
σ	Standard deviation
υ	Degrees of freedom parameter
ω	Weighting term

Other Symbols

$\mathcal{D}$	Distance between two adjacent frequency bins
ℓ	Non-time dependent iteration index
RT_{60}	Reverberation time (60dB)

List of Figures

List of Tables

Chapter 1

INTRODUCTION

> "One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may call it 'the cocktail party problem.' No machine has been constructed to do just this, to filter out one conversation from a number jumbled together?"
>
> *Colin Cherry - 1954*

1.1 Cocktail party problem

The cocktail party problem (CPP) was first proposed by Colin Cherry [1], [2] and describes a problem where there are multiple human speakers talking simultaneously within an enclosed environment where it is required that each speaker's voice is isolated (separated) from the other present voices, similar to the manner in which the human sensory system can identify, and listen to, individual speakers in a situation such as a crowded party, Figure 1.1; hence the name of the problem. An overview of the concept and review of methods addressing the problem can be found in [3], [4].

Despite many years of research, a full solution to the problem is lacking, and there are many facets of the problem still to be investigated and addressed, such as the situation where there are more speakers than sensors ('ears') and also how a human exploits a priori knowledge of a speaker and/or environment to aid the separation. This thesis will address some of these issues.

Engineers look to tackle this signal processing problem by using a machine, or an algorithm implemented on an embedded system (such as a digital signal processor system) to mimic the ability of the human sensory system to separate speech sources. One aspect that a human uses is eyesight, which provides useful information to the human brain such as the location of the speaker, amongst other pieces of information that could be considered a priori knowledge in a source separation algorithm. Indeed, as the McGurk effect demonstrates [5], there is in an inherent link between human eyesight and hearing. An interesting perspective of an engineering challenge such as this is how this link can be replicated in an automated digital system to enhance the solution.

The neurobiological perspective of how the human brain approaches the cocktail party problem with multiple speakers is beyond the scope of this engineering-based problem [6].

1.2 Blind source separation

The cocktail party problem is a typical blind source separation (BSS) problem. BSS problems are characterised by an unknown mixing pro-

Figure 1.1. The cocktail party problem (Image from *Telegraph.co.uk*).

cess and unknown signal sources, where only sensor (e.g. microphone) observations are available. Such problems are commonly addressed with independent component analysis (ICA) [7], [8].

The concept of blind source separation was first examined in 1982 by the authors of [9] and Jean-Pierre Rolls within the context of decoding muscle motion (motion decoding) at the end of muscle fibers in the field of neuroscience, as described in Section 1.1.1 of [8]. It was later in 1994 that Comon went on to formalise the concept of independent component analysis [10]. ICA has many applications in biomedical signal processing including: electroencephalography (EEG), electrocardiography (ECG) signals, electromyography (EMG) and magnetoencephlaography (MEG) signals [11], [12].

In addition, blind source separation and independent component analysis has applications in many fields including communications in the form of multiple-input multiple-output (MIMO) equalisers, RADAR systems, SONAR systems and image processing applications, including remote sensing. BSS also has interesting applications within finance, potentially revealing underlying trends in markets.

Given the growing tendency for voice-automated technology such as Apple's 'Siri' [13] and Google's 'Google Now' [14] services on smartphones (and now even 'smart watches'), blind source separation is an important component in the field of natural language processing which improves the intelligibility of speech signals, and can be considered as a preprocessing stage before a speech recognition algorithm. Smartphones are also often equipped with video cameras, offering potential for audio-video processing applications exploiting the speaker location. Teleconferencing is also a significant area where blind source separation for speech mixtures can be used, as well as surveillance and security applications.

Over the years, different methods to address blind source separation problems have emerged. Computational auditory scene analysis (CASA) [15], non-negative matrix factorisation (NMF) [16] and deep learning for neural networks (DNN) [17] have all been applied to BSS. None of these methods are considered in this thesis, due to the computational power necessary to perform these methods, causing algorithms to be unsuitable for online or in real-time implementation. Furthermore, deep learning for neural networks would require a training phase to initially train the neural network.

The methods described in this thesis come under either methods exploiting the geometry between two microphones and use only second order statistics (SOS) or methods which use higher order statistics (HOS) such as independent vector analysis, or in the last full chapter a combination of the two.

1.2.1 Audio-visual source separation

The major problem for using blind source separation are the multiple paths a speech signal can travel within a real environment such as a room. Illustrated in Figure 1.2 is an example of how an acoustic signal such as speech can propagate in a room environment. Several paths are depicted over which the speech signal could travel. As each of these paths in an enclosed acoustic environment may be modelled as a filter this is described as a convolutive model. The reverberant nature of the room environment is the main challenge to overcome in blind source separation algorithms, especially for physically moving speech sources, as the mixing filters would be time variant. Known speaker locations from video cues can help overcome some of the difficulty encountered. This convolutive model could be thought of as multichannel blind deconvolution, however within the context of this thesis this term is not used as filtered versions of the original sources would be accepted as outputs to a BSS algorithm, thus this is not strictly deconvolution. Throughout the thesis the terms 'blind source separation' and 'source separation' are used inter-changeably.

Colin Cherry outlined in his original paper how visual information could be used to aid the source separation process [1]. Moreover, many emerging technologies are likely to be equipped with cameras (e.g. smartphones, wearable technology and robotic human machine interfaces). The content of this thesis looks to exploit video information, by employing a priori knowledge of the speaker location, in a similar manner in which a human speaker uses eyesight and hearing to focus attention on one speaker; as such audio-visual source separation increasingly becoming an important aspect of the CPP. A recent review

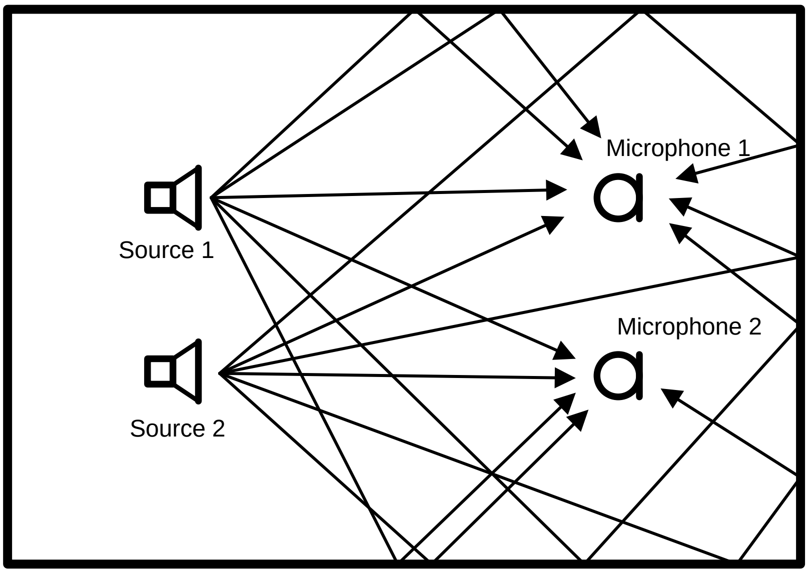

Figure 1.2. A schematic of a selection of paths a sound pressure wave could take between two sources and two microphones.

of the main techniques [18] highlights several areas of the field.

1.2.2 Batch, online and real-time

The term ‘online’ is generally used within the field of signal processing to mean a process that contains a sequence of instructions that is performed iteratively as signal data are provided, unlike a batch algorithm that would wait until all signal data (or a time block of sufficient size) are received before processing the signal data with the sequence of instructions.

A real-time method is one that can process signal data at a constant rate with a constant time delay between input and output signal data, however no such constraint is placed on an online method.

Typically, an online algorithm is derived with real-time implementation in mind, and therefore often the two terms are muddled. For clarity, in this thesis the distinction is made that online methods have the potential to be implemented in real-time due to their algorithmic formulation, however, an online method may not have been implemented in real-time and results may not have been gathered from the real-time version.

1.3 Local scientific context and support

Work in the thesis is the culmination of several though highly related research areas. Previous research has investigated the source separation of moving sources by exploiting video cues. Along the same lines, the work in this thesis has benefited from additional funding opportunities related to audio-visual source separation, primarily a joint UK-France PhD scholarship funding by Direction Generale de l'Armement (DGA) and the Defence Science and Technology Laboratory (Dstl). Dstl also provided extra funding in the fourth year of PhD study. The PhD research project came under a larger European research project Challenges for Extraction and Separation of Sources (CHESS). Support from this project primarily came in the form of addressing multimodality and source extraction issues.

1.4 Aims and objectives of the thesis

The main aim of the thesis is to develop effective, low complexity online methods for source separation, avoiding higher order statistics if possible in an attempt to reduce computational complexity. The contents of the thesis address the following objectives:

- Objective 1: To develop methods which avoid the need for higher order statistics, by exploiting video cues in the form of known speaker locations to aid the process. By doing so, it is hoped that algorithm complexity can be reduced to an extent where the developed methods can be considered online.

- Objective 2: Investigate potential time domain methods using only SOS which attempt to solve the circularity problem.

Chapter 3 deals with Objectives 1 and 2, detailing a time domain solution which employs the singular value decomposition to find a pair of finite impulse response filters. An adaptive filtering stage follows to recover all remaining sources.

- Objective 3: Provide evidence in the form of an embedded system that blind source separation can be performed in real-time by developing a demonstration of an online algorithm.

- Objective 4: Improve existing online techniques so that they are suitable for speech separation in real-time.

Chapter 4 addresses Objectives 3 and 4 by utilising the online independent vector analysis algorithm. A real-time demonstration on a digital signal processor was implemented, and a new source prior for independent vector analysis is suggested to better suit speech signals.

- Objective 5: Address convergence and separation performance of online independent vector analysis in the moving source case with the aid of known speaker locations.

Chapter 5 examines independent vector analysis in the context of physically moving source signals to address Objective 5. Previous methods from the thesis are employed to improve the convergence and performance of online independent vector analysis for moving sources.

1.5 Outline of the thesis

The thesis is organised as follows:

- Chapter 2, introduction of blind source separation within the independent component analysis framework. An overview of independent component analysis is included and how this leads to the permutation problem in the frequency domain blind source separation. The independent vector analysis algorithm is introduced to address the permutation problem. Frequency domain ICA and IVA are compared for convolutive mixtures. Also introduced are datasets, relevant details on acoustics and performance parameters used throughout the thesis.

- Chapter 3, the design of a pair of time-domain filters is considered to achieve target signal cancellation in a multi-source environment. The problem is formulated as a minimisation of a sum squared error cost function with respect to the pair of finite impulse response cancelation filters. Two methods are compared; direct minimisation, which is achieved through an alternating gradient descent based method and a novel method based on the method of principal angles is proposed, which exploits the singular value decomposition which is the main focus of the chapter. Simulation studies show that the gradient descent method suffers from slow convergence, but this is overcome by the method based on principal angles which also achieves a lower cost than the gradient descent approach. The cancellation filters are combined with an adaptive filtering scheme to address a video-informed audio source separation problem.

- Chapter 4, independent vector analysis is employed to directly address the permutation problem by modeling the dependencies between frequency bins, namely making use of a source prior. An alternative source prior for online natural gradient independent vector analysis is proposed. A Student's t probability density function is known to be more suited for speech sources, due to its heavier tails, and is incorporated into a real-time version of online natural gradient independent vector analysis. The importance of the degrees of freedom parameter within the Student's t distribution is highlighted. The final algorithm is realised as a real-time embedded application on a Texas Instruments digital signal processor platform.

- Chapter 5, describes a combined technique based on work from the previous two chapters which separates one source from a mixture using the pair of cancellation filters used in Chapter 3. Online independent vector analysis is then employed to recover the original sources. It is shown that this approach improves convergence times of online independent vector analysis in the case of physically moving source signals.

- Chapter 6 concludes the thesis and suggests future work.

In addition, an appendix is also included describing a potential online method which exploits a technique for creating artificial impulse responses in a room.

Chapter 2

CONVOLUTIVE SOURCE SEPARATION TECHNIQUES AND RELEVANT LITERATURE REVIEW

2.1 Independent component analysis

The cocktail party problem is a typical blind source separation application. Such problems are characterised by an unknown mixing process and unknown original signal sources, where only sensor (e.g. microphone) observations are available. They are typically addressed with independent component analysis (ICA) [7], [8], [19].

Early interest in blind source separation originated from France, Hérault and Jutten being pioneers in the field [9], [20] who proposed a method with foundations in neural network theory to recover unknown source signals. The concept of independent component analysis was formalised by Comon in [10]. In the subsequent years several flavours of the ICA algorithm emerged, including: the introduction of the Infomax algorithm [21], [22] and the popular FastICA algorithm described in [23]. A tensoral method can also be applied to ICA, for example in [24] the joint approximate diagonalisation of eigenmatrices (JADE) algorithm is introduced which is a member of the family of methods known as fourth-order blind identification (FOBI) algorithms. A recent review of ICA techniques can be found in [25].

Realistic audio signals measured at microphones are generally generated by a convolutive model due to the reverberant nature of real world environments; thus ICA algorithms which address the audio BSS problem are commonly implemented in the frequency domain [26]–[29]. A drawback of frequency domain ICA and other frequency domain blind source separation techniques is that the calculated unmixing filters may permute the sources at each frequency bin (known as the permutation problem), due to the permutation ambiguity inherent in ICA. The other ambiguity present in ICA, the scaling ambiguity, is easily overcome, normally by appropriately scaling the outputs.

Various methods have been suggested to mitigate the effect of the permutation problem, in [30] smoothing over adjacent frequency bins is suggested as a way of addressing the issue. In addition, [31] suggests limiting the length of the filter in the time domain. Also, in [32] video tracking of sources is suggested as an approach to address the permutation problem. Other methods to address convolutive mixtures can be found in [30], [33], and methods focusing on speech separation can be found in [34].

2.1.1 Linear mixing and unmixing model

2.1.1.1 The instantaneous model

In the context of acoustic sources such as speech, the instantaneous mixing model describes only the direct path between speakers and microphones. This is given in the two-microphone and two-source case (2×2)) as:

$$x_1(t) = h_{11}s_1(t) + h_{12}s_2(t) \tag{2.1.1}$$

$$x_2(t) = h_{21}s_1(t) + h_{22}s_2(t), \tag{2.1.2}$$

this model is at a given instant in time, where t denotes the discrete time index, and assumes no time delays are present. The h_{ji} terms (for the j-th microphone and the i-th source) are scaling parameters which describe a simplistic acoustic path between a source and a microphone, a diagram of the simple instantaneous two-microphone two-speaker situation is depicted in Figure 2.1.

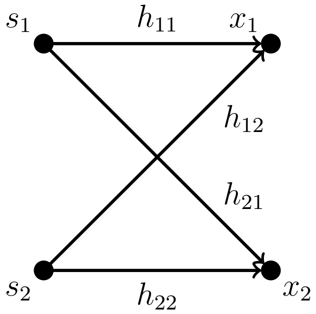

Figure 2.1. A schematic of the instantaneous 2×2 mixing model.

Equations (2.1.1) and (2.1.2) can be written generally for any number of sources, with additive noise, at each microphone as:

$$x_j(t) = \sum_{i=1}^{N} h_{ji} s_i(t) + \zeta_j(t), \tag{2.1.3}$$

where $x_j(t)$ is the measured signal at the j-th microphone, $s_i(t)$ is the speech signal generated by the i-th source, h_{ji} is the scaling parameter that models the effect of the environment between the i-th source and the j-th microphone, $\zeta_j(t)$ is additive zero mean noise uncorrelated with the speech signals and N is the number of sources. The noise, $\zeta_j(t)$, can be considered as noise from extra spatially separated sources, however, the situation where additive noise is caused by non-spatial factors (such as quantisation noise within an analogue to digital converter) is considered beyond the scope of this thesis. Nevertheless, as in many works in the field [35], [36], the noise term $\zeta_j(t)$ is dropped for brevity for the remainder of the thesis. More generally the instantaneous model in Equation (2.1.3), without noise, can be written in vector form as:

$$\mathbf{x} = H\mathbf{s}, \tag{2.1.4}$$

where $H \in \mathbb{R}^{M \times N}$, $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{s} \in \mathbb{R}^N$. The parameters h_{11}, h_{21}, h_{12}, and h_{22}, for the 2×2 case, can be grouped together in the matrix H, therefore[1]:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \tag{2.1.5}$$

where $\mathbf{x}$ is decomposed as $\mathbf{x} = [x_1, x_2]^T$, where $(\cdot)^T$ denotes a matrix transpose, and $\mathbf{s}$ is decomposed as $\mathbf{s} = [s_1, s_2]^T$. In fact, Equation (2.1.4) could represent any number of sources (N) and microphones (M). Methods to address the over-determined case ($M > N$) and

[1] For notational simplicity the discrete time index t is dropped.

the under-determined case ($M < N$) have been developed, such as via dimensionality reduction using principal component analysis (PCA) for the over-determined case [37]. However, for brevity in this thesis these situations are not considered. An assumption throughout the thesis is that the number of original sources is equal to the number of microphone observations ($M = N$). Therefore the matrix H is assumed to be square, and agrees with the standard noise free ICA model.

The unmixing model for the instantaneous case is then given by:

$$\hat{\mathbf{s}} = G\mathbf{x}, \tag{2.1.6}$$

where $G \in \mathbb{R}^{N\times M}$ and $\hat{\mathbf{s}} \in \mathbb{R}^{N}$. As H is constrained to be square, theoretically, if the mixing matrix H is known it would be possible to find a matrix that perfectly reconstructs the original sources by taking the inverse of H, i.e. $H^{-1} = G$. The goal of ICA is to find an unmixing matrix G given only the observed signals, hence the term *blind* source separation. A schematic of the full mixing and unmixing model, when $M = N = 2$, is illustrated in Figure 2.2.

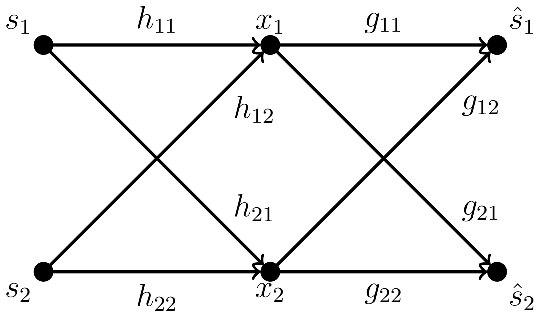

Figure 2.2. A schematic of the instantaneous mixing and unmixing model when $M = N = 2$.

2.1.2 Statistical independence

Within the ICA model the original sources are modelled as random variables; within this model it is assumed that a variable (source) s_1 gives no information about s_2, i.e. the mutual information is zero. This is a major assumption of the ICA model, which is encapsulated by the concept of statistical independence, which can be formally defined in the real signal case as:

$$p(\mathbf{s}) = \prod_{i=1}^{N} p(s_i), \tag{2.1.7}$$

In other words variables are independent when the joint probability density function (pdf) can be factorised into the product of its marginal probability density functions. For the two component (source) case the probability density functions can be factorised as:

$$p(s_1, s_2) = p(s_1)p(s_2). \tag{2.1.8}$$

Statistical independence is not the same as uncorrelatedness, which can be considered to be a weaker form of independence. Independence implies uncorrelatedness, however uncorrelatedness does not imply independence, except for Gaussian sources.

2.1.3 Ambiguities of independent component analysis

ICA is able to recover independent components (estimates of the original sources), however two inherent ambiguities arise as a result.

1. Scaling ambiguity - the signal power of the estimated source signals do not generally match those of the source signals. This is not normally a problem because it is easy to set the variances of the

estimated zero mean sources to unit variance (i.e. $E[|\hat{s}_i|^2] = 1$), assuming that the signal in question has zero mean. This would still have a sign ambiguity though for many types of signal, including speech signals, this is usually trivial.

2. Permutation ambiguity - ICA cannot determine the order of the estimated sources. To be precise, the unmixing model could be written as $\hat{\mathbf{s}} = PGx$, where P is a permutation matrix to be determined. With an instantaneous model this problem is not of great impact in a variety of situations (so long as the sources have been recovered adequately). However, when dealing with convolutive mixtures, which necessitates operating in the frequency domain, this becomes a major problem, as discussed later in the chapter.

The permutation problem has been an active area of research for several years, one of the most promising frequency domain methods to address the problem is introduced later in this chapter.

2.1.4 Derivation of natural gradient ICA

By using the concept of mutual information [7] which gives a measure of the independence of two random variables, $\mathcal{I} = \sum_i \mathcal{H}(s_i) - \mathcal{H}(\mathbf{s})$, where $\mathcal{H}(\cdot)$ denotes the differential entropy, it is possible to write:

$$\begin{aligned} J_{ICA} &= \mathcal{KL}\Big(p(\hat{\mathbf{s}})||\prod_{i=1}^{N} q(\hat{s}_i)\Big) && \text{(2.1.9a)} \\ &= \int p(\hat{\mathbf{s}}) \log \frac{p(\hat{\mathbf{s}})}{\prod_{i=1}^{N} q(\hat{s}_i)} d\hat{\mathbf{s}} && \text{(2.1.9b)} \\ &= \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} - \log|\det G| - \sum_{i=1}^{N} \int p(\hat{\mathbf{s}}_i) \log q(\hat{s}_i) d\hat{\mathbf{s}}_i && \text{(2.1.9c)} \\ &= \text{const.} - \log|\det G| - \sum_{i=1}^{N} E[\log q(\hat{s}_i)], && \text{(2.1.9d)} \end{aligned}$$

where $E[\cdot]$ denotes the mathematical expectation, $\mathcal{KL}(\cdot)$ denotes the Kullback-Liebler divergence and $q(\cdot)$ is an approximated pdf of the original sources. Between equations (2.1.9b) and (2.1.9c) the Jacobian expression:

$$p(\hat{\mathbf{s}}) = p(G^{-1}\hat{\mathbf{s}})|\det G|^{-1} = p(\mathbf{x})|\det G|^{-1}, \quad (2.1.10)$$

is used to derive the differential entropy of the observations in first term of Equation (2.1.9c), which becomes constant in Equation (2.1.9d), where $(\cdot)^{-1}$ denotes the inverse of a matrix. This is exactly the same as for ICA derived by mutual information and can be shown to be the same as the Kullback-Liebler divergence between the joint pdf and the product of the marginal pdfs:

$$\mathcal{KL}\Big(p(\hat{\mathbf{s}})||\prod_{i=1}^{N} q(\hat{s}_i)\Big) = \sum_{i=1}^{N} \mathcal{H}(\hat{s}_i) - \mathcal{H}(\hat{\mathbf{s}}). \quad (2.1.11)$$

By taking the partial derivatives of Equation (2.1.9d), the gradient of the cost function can be calculated and is given as:

$$\Delta G = -\frac{\partial J_{ICA}}{\partial G} = G^{-T} - E[\varphi_{ICA}(\hat{\mathbf{s}})]\mathbf{x}^T, \tag{2.1.12}$$

where $(\cdot)^{-T}$ denotes the inverse of a matrix combined with a matrix transpose and $-\frac{\partial \log q(\cdot)}{\partial \mathbf{s}_i} = \varphi_{ICA}(\cdot)$ is the nonlinear score function for ICA in its general form.

The natural gradient [38] is then calculated by right multiplying through by $G^T G$:

$$\Delta G \propto (I - E[\varphi_{ICA}(\hat{\mathbf{s}})\hat{\mathbf{s}}^T])G, \tag{2.1.13}$$

thus the update rule for natural gradient ICA (NG-ICA):

$$G(\ell + 1) = G(\ell) + \eta(I - E[\varphi_{ICA}(\hat{\mathbf{s}})\hat{\mathbf{s}}^T])G(\ell), \tag{2.1.14}$$

where η is a learning rate, and an iteration index (ℓ) has been added. The non-linear score function is based on a pdf which is chosen to model the statistics of the original sources. Often for speech, a Laplacian pdf is chosen, for example:

$$q(s_i) \propto \exp\left(- \frac{|s_i - \mu_i|}{\sigma_i} \right), \tag{2.1.15}$$

where σ_i is the standard deviation of each source. The non-linear score function becomes:

$$\varphi_{ICA}(\hat{s}) = \frac{\hat{s}_i}{|\hat{s}_i|}. \tag{2.1.16}$$

Various source priors can be chosen, each yielding a different result.

Here the above Laplacian source prior is chosen for the example later in the chapter.

2.1.4.1 The convolutive model

In the previous model Equation (2.1.4) time delays were not considered. Realistically, this model would not be applicable in a real reverberant environment such as a room. So that the model is more realistic, time delays are introduced in the acoustic path between a speaker and microphone, modelling possible acoustic reflections in an environment. This is described as the 'convolutive model'. The observation at each sensor of a microphone array can be modelled in the general case in the frequency domain as a multiplicative mixture from each source of the form:

$$x_j^{(k)}[n] = \sum_{i=1}^{N} h_{ji}^{(k)} s_i^{(k)}[n], \tag{2.1.17}$$

where the variables are now complex valued and the superscript $(\cdot)^{(k)}$ has been added to denote an operation at frequency bin k and omits the noise term $(\zeta_j^{(k)})$ as discussed previously. The short-time Fourier transform (STFT) time block index (n) is also added, however for brevity is dropped for this derivation.

The time-domain convolutive form can thereby be written in a frequency-domain matrix form in a similar manner to Equation (2.1.4):

$$\mathbf{x}^{(k)} = H^{(k)}\mathbf{s}^{(k)}, \tag{2.1.18}$$

where the variables are redefined for the complex case as: $H^{(k)} \in \mathbb{C}^{M\times N}$, $\mathbf{x}^{(k)} \in \mathbb{C}^M$ and $\mathbf{s}^{(k)} \in \mathbb{C}^N$. Note that in this thesis it is assumed that there is an equal number of microphones and speakers, therefore H is assumed to be square at each frequency bin, k. By writing the mixing model in this manner it can be seen that this can be thought of an instantaneous mixture at each frequency bin. For the remainder of this chapter (and thesis) it is assumed that all mixtures are convolutive in the time domain and that calculations are carried out in the frequency domain except where explicitly mentioned, hence the addition of the frequency bin index, k, in superscript.

The goal of a frequency domain BSS (FD-BSS) algorithm is to find an unmixing matrix $G^{(k)}$ at each frequency bin, k, so that:

$$\hat{\mathbf{s}}^{(k)} = G^{(k)}\mathbf{x}^{(k)}, \tag{2.1.19}$$

where the unmixing variables are redefined for the complex case as: $G^{(k)} \in \mathbb{C}^{N\times M}$, $\hat{\mathbf{s}}^{(k)} \in \mathbb{C}^{N}$ and $\mathbf{x}^{(k)} \in \mathbb{C}^{M}$. Similar to each mixing matrix $H^{(k)}$, $G^{(k)}$ is assumed to be square at each frequency bin.

This can be written in its decomposed form as:

$$\hat{s}_i^{(k)} = \sum_{j=1}^{M} g_{ij}^{(k)} x_j^{(k)}, \tag{2.1.20}$$

where $\hat{s}_i$ is the estimated signal for the i-th source and g_{ij} is the frequency domain unmixing filter to find the estimation of the i-th source from the j-th observation.

If the mixing system was known the inverse of $H^{(k)}$ could be found, so that $H^{(k)-1} = G^{(k)}$, which would recover the original sources exactly, however it is assumed the mixing system is unknown in ICA and IVA. As the mixing matrices are unknown the goal now becomes to find an estimate of the unmixing matrices $G^{(1,\dots,K)}$, only using the observed signals $\mathbf{x}^{(1,\ldots,K)}$; the full frequency domain mixing and unmixing system at frequency bin k, is illustrated in Figure 2.3 when $M = N = 2$. Separation can be potentially achieved by assuming that the original sources are statistically independent from each other at each frequency bin.

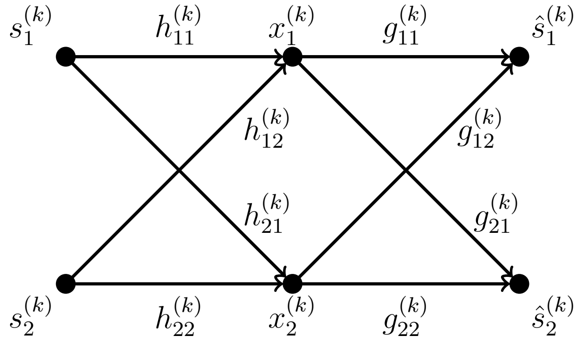

Figure 2.3. A schematic of the convolutive mixing and unmixing model when $M = N = 2$.

2.1.5 The permutation problem in frequency domain independent component analysis

An approach to solving the case of convolutive mixtures is to operate in the frequency domain. A naïve approach would be to estimate the unmixing matrix at each frequency bin by treating each frequency bin as a separate instantaneous problem. This seems an attractive proposition at first, however it soon becomes apparent that one of the ambiguities of ICA, the permutation ambiguity, has a major effect on processing. As the instantaneous problem is effectively solved at each frequency bin it is highly improbable that the estimated mixtures at each frequency bin are in a consistent order across frequency bins. Figure 2.4 illustrates this problem, in this example of the problem for one frequency bin, source s_2 is in the position of source s_1, source s_N is in the position

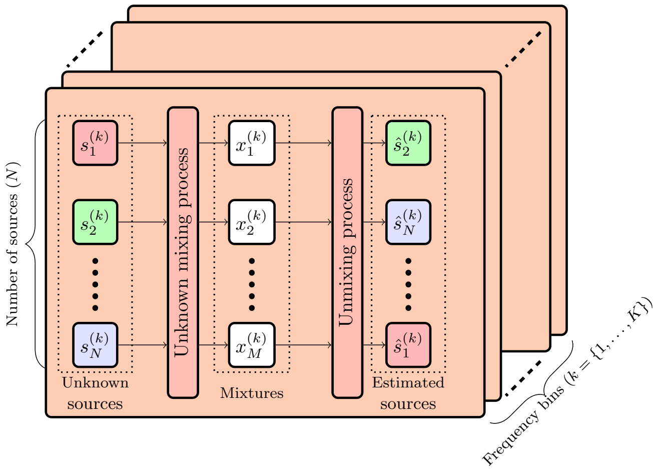

Figure 2.4. An illustration of the permutation problem of FD-BSS.

of s_2 and s_1 is in the position of s_N. The order of the estimated sources order would generally differ in each frequency bin. This is a problem inherent to ICA and the permutation is not known at each frequency bin (illustrated by the orange 'slices' in Figure 2.4). The instantaneous model is not suited to realistic mixing environments due to time delays in the convolutive mixing model. An early attempt to address this in time domain is described in [39], which cancels 4-th order cross cumulants to find estimates. However convolutive mixtures often motivate operating in the frequency domain, see Equation (2.1.17) for the associated mixing model. Various early attempts to address convolutive mixtures include [40] and [41] which introduces a feedback network based on [21].

As previously mentioned, a method which was introduced in 2000 is Parra and Spence's method [31]. By restricting the length of a filter in the time domain the effect of this forces 'smoothness' across frequency bins. In 2004 another robust approach was presented based on a combination of direction of arrival and an interfrequency correlation

[42].

Previous attempts to model multidimensional ICA includes [43]. This method models the dependencies between frequency bins and includes a technique called independent subspace analysis (ISA) [44], which does not require independence between sources, though does require independence on the projections on a set of subspaces which allows the method to model dependencies (such as those found in a frequency domain speech signal).

However, [35] introduces the idea of independent vector analysis (IVA), which explicitly models the dependencies between frequency bins in the algorithmic formulation, by modelling dependencies within the vector sources and independence between vector sources by using a multivariate pdf. This is the most promising method to date within the ICA-style framework which addresses the permutation problem in FD-BSS and is described in the following section.

2.2 Independent vector analysis

Independent vector analysis, first proposed by Kim in [35], whilst modelling the statistical independence between sources, also models the statistical relationships across frequency bins; thus within a source, dependency between frequency bins is maintained. This approach directly addresses the permutation problem inherent in FD-ICA.

This intra-frequency bin dependency is built into the algorithmic formulation of the algorithm and does not attempt to solve the instantaneous problem at each frequency bin, and no kind of ‘post’ or ‘pre’ processing is required. The way in which the joint statistics across frequency bins are captured is by using a multi-variate source prior in the

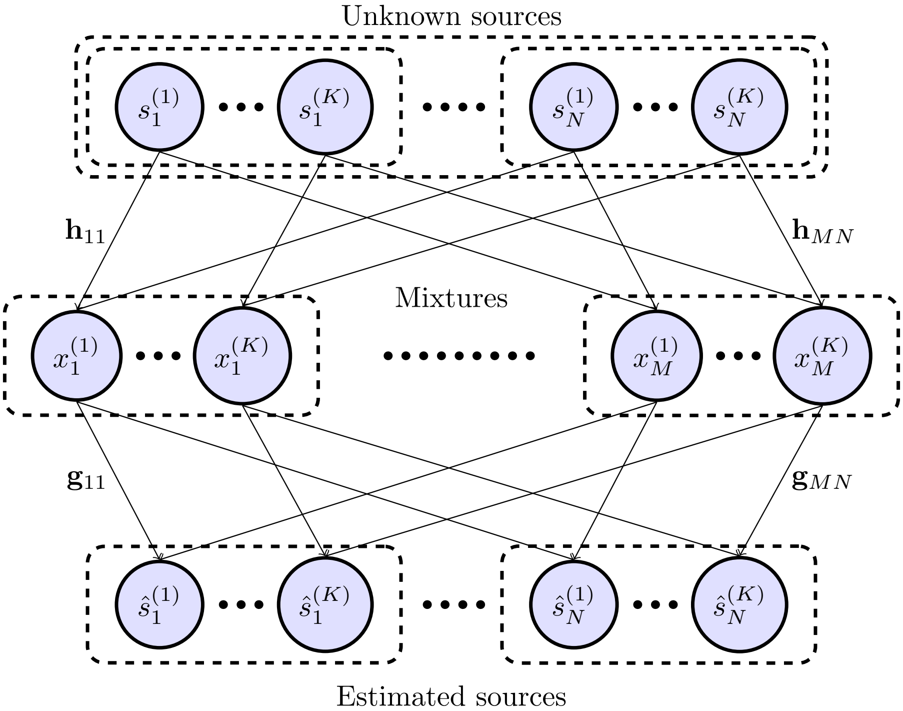

Figure 2.5. Independent vector analysis model, showing independence between sources and dependence within individual sources.

derivation.

Originally proposed for acoustic mixtures, IVA also has the potential for being used with other types of data and was reviewed and given a more general context in [36]. An online version with a mixture of Gaussian source priors is described in [45]. Potential use of IVA includes smartphone applications [46] in the form of auxiliary function IVA [47]. Similar to ICA a 'fast' version of the algorithm can be found in [48].

As well as assuming independence between individual sources (components), dependencies are assumed within the sources in the mathematical model, for a schematic representation of this see Figure 2.5.

In an audio source separation context these dependencies are used to model the higher-order dependencies between frequency bins in an

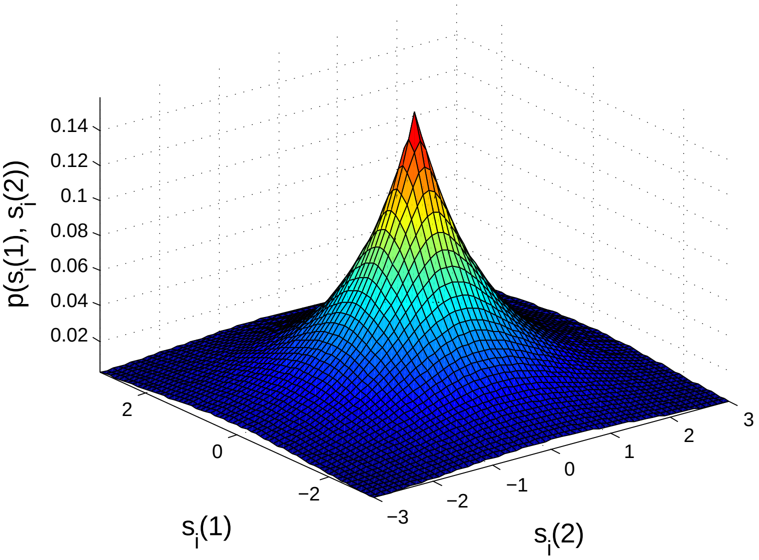

Figure 2.6. Bivariate example of a super-Gaussian distribution, which models the inter source higher order frequency dependencies. As the frequency domain data is complex this graph is valid for the real or imaginary part of the frequency domain data.

audio or speech signal, whilst maintaining inter-source independence. The way in which these intra-source dependencies are modelled is by introducing a multi-variate probability density function. The original article on IVA [35] proposes the use of a super-Gaussian multivariate probability density function, a bivariate version is shown in Figure 2.6.

As for ICA, independence between sources is modelled by the Kullback-

Leibler divergence, thus a cost function (J_{IVA}) is derived as:

$$\begin{aligned} J_{IVA} &= \mathcal{KL}(p(\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N)||\prod_{i=1}^{N} q(\hat{\mathbf{s}}_i)) && \text{(2.2.1a)}\\ &= \int p(\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N) \log \frac{p(\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N)}{\prod_{i=1}^{N} q(\hat{\mathbf{s}}_i)} d\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N && \text{(2.2.1b)}\\ &= \int p(\mathbf{x}_1 \ldots \mathbf{x}_M) \log p(\mathbf{x}_1 \ldots \mathbf{x}_M) d\mathbf{x}_1 \ldots \mathbf{x}_M && \text{(2.2.1c)}\\ &- \sum_{k=1}^{K} \log|\det G^{(k)}| - \sum_{i=1}^{N} \int p(\hat{\mathbf{s}}_i) \log q(\hat{\mathbf{s}}_i) d\hat{\mathbf{s}}_i \\ &= \text{const.} - \sum_{k=1}^{K} \log|\det G^{(k)}| - \sum_{i=1}^{N} E[\log q(\hat{\mathbf{s}}_i)], && \text{(2.2.1d)} \end{aligned}$$

where the block diagonal diagonal structure of the global unmixing matrix ($G^{(1,\ldots,K)}$) introduces a summation in the second term.

The partial derivative of the cost function is employed to find the gradient;

$$\Delta g_{ij}^{(k)} = -\frac{\partial J_{IVA}}{\partial g_{ij}^{(k)}} = g_{ij}^{(k)-H} - E[\varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)}) x_j^{(k)*}], \quad (2.2.2)$$

where $\left[(G^{(k)-1})^H\right]_{ij} \equiv g_{ij}^{(k)-H}$. The natural gradient [38] is then applied by multiplying through by $G^H G$:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{N} (\delta_{il} - E[\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \hat{s}_i^{(k)*}]) g_{lj}^{(k)}, \quad (2.2.3)$$

where δ_{il} is the Kronecker delta, i.e. when $i = l$, $\delta_{il} = 1$, and zero otherwise. Therefore Equation (2.2.3) is the update rule for the batch version of natural gradient IVA (NG-IVA). The term $\varphi^{(k)}(\cdot)$ is a multivariate score function which can be based on a multivariate super-Gaussian source prior, and is written in the general case as:

$$\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) = -\frac{\partial \log q(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)})}{\partial \hat{s}_i^{(k)}}. \tag{2.2.4}$$

Based on a source prior, representing the content of frequency domain information of the original signals, the original source prior is written as:

$$q(\mathbf{s}_i) \propto exp\Big(- ((\mathbf{s}_i - \mu_i)^H \mathbf{\Sigma}_i^{-1} (\mathbf{s}_i - \mu_i))^{\frac{1}{2}} \Big), \tag{2.2.5}$$

as proposed in the original formulation [35] (Figure 2.6), where $\mathbf{s}_i = (s_i^{(1)} \ldots s_i^{(K)})$. Note that the 'hat' symbol $(\hat{\cdot})$ is omitted as it is an assumption being about the original sources. A more thorough discussion on the selection of the multivariate score function $\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)})$ can be found in Chapter 4. By setting the mean values to zero and setting the covariance matrix (Σ) to the identity matrix, the non-linear score function is derived as:

$$\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}, \tag{2.2.6}$$

and is the main component on maintaining dependencies across frequency bins.

2.3 Small room acoustics and speech

Reverberant, noisy and multi-source environments, such as a room, pose a significant challenge in signal processing systems particularly in online applications. Generally speaking, multi-sensor array systems are required to enhance or cancel a target signal source by means of spatial filtering so that the target, or other measured signals, can be processed more efficiently. Source separation methods and algorithms are no exception. In the following chapter an outline of the datasets and techniques employed for source separation for convolutive mixtures systems are detailed.

2.3.1 Room Impulse Responses

Realistic audio mixtures are convolutive, meaning that an observation at discrete time, t, would have contributions of previous time samples of an original signal, s. This effect is due to reverberant environments where sound pressure waves can take several paths of different physical lengths and attenuations, hence the delays in time and scaling quantities. Reverberant environments can be modelled by impulse responses (IRs), and the term real room impulse response (RIR) is used to describe non-artificial or simulated impulse responses. RIRs can be considered as FIR filters which describe the acoustic path of a sound pressure wave within an enclosed environment [49], therefore:

$$x(t) = \sum_{\tau=0}^{L} h(\tau)s(t-\tau), \tag{2.3.1}$$

where $x(t)$ is the observation at discrete time t, $h(t)$ is the causal filter impulse response modelling the acoustics of a room, and $s(t)$ is the original source.

2.3.1.1 Reverberation time

Reverberation time (RT) is the time period that it takes for the energy of an impulse response to decay below a certain threshold, usually set

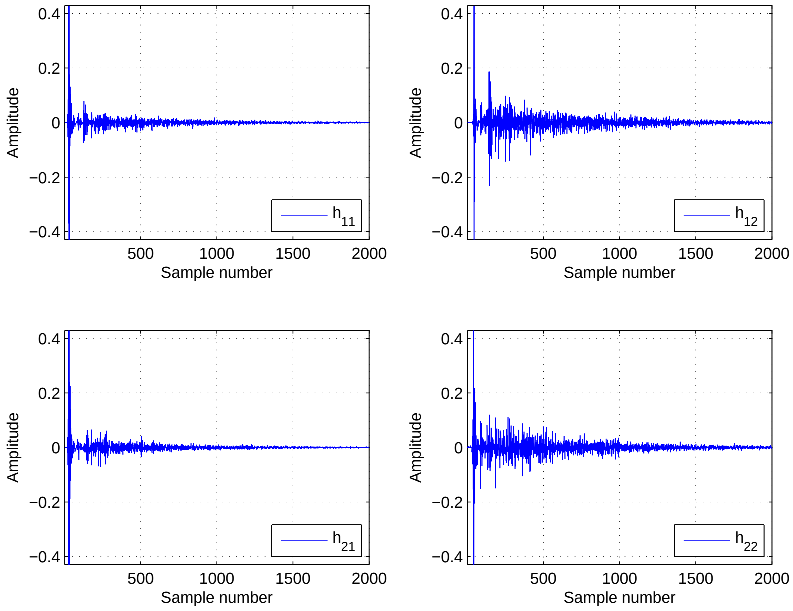

Figure 2.7. Example BRIRs for a two-source two-microphone scenario, where s_1 is 0° at 1m and s_2 is 45° at 1m from the centre of the microphone array.

in decibels, a common threshold is 60dB and is written as RT_{60}.

Throughout the thesis the reverberation time is calculated using the Schroeder integral method [50], where a decay curve (E) is defined in its continuous form as:

$$EC(t_c) = \int_{t_c}^{\infty} h^2(t_c) dt_c, \tag{2.3.2}$$

where t_c denotes the continuous time index. A normalised discrete decay curve can be found by:

$$EC(t) = \frac{\sum_{\tau=t}^{\infty} h(\tau)^2}{\sum_{\tau=0}^{\infty} h(\tau)^2}. \tag{2.3.3}$$

Linear regression would then be used to find an estimate of a line which would cross the horizontal axis at -60dB, a MATLAB implementation of this can be found in [51]. Further details on measuring reverberation time and decay curves can be found in [52].

2.3.2 Critical distance

The critical distance is the point in an enclosed environment where the energy of the direct path (component) is equal to the energy of the reverberant paths [49]. Critical distance can be approximated using the following equation:

$$r_c \approx 0.1\sqrt{\frac{\text{vol}}{\pi RT_{60}}}, \tag{2.3.4}$$

where vol is the volume of the room in m^3.

2.3.3 Image method

The image method (IM) is a well-known method in acoustics for estimating an impulse response within a simulated small room [53]. When compared to the random uncertain nature of real RIRs, the image method is of distinctly artificial nature. Whilst providing suitable IRs for 'proof of concept' methods, for robust testing this method is not considered to be suitable. In later chapters RIRs, particularly from [54] are employed. A study into the uncertainties of IRs can be found in [55] and a method which attempts to exploit the image method for source separation can be found in Appendix A.

2.3.4 Binaural real impulse responses

Real binaural room impulse responses (BRIRs) are used throughout the thesis [54]. BRIRs are recording using a KEMAR (Knowles Electronics Manikin for Acoustic Research) dummy head to simulate the effect of a human head within a real acoustic environment. The dataset consists of source azimuth angle locations of (0°, 15°, 30°, 45°, 60°, 75° and 90°) at distances of 0.15m, 0.40m and 1.00m from the equidistant point between the two ears on the KEMAR dummy head which is placed at various positions in a room. The only position that is considered in this thesis is the 'centre' position at [2.5, 4.5, 1.5]m in a room with dimensions of [5, 9, 3.5]m, where the KEMAR dummy head was placed in the centre of the room approximately 1.5m from the ground facing lengthwise in the room. The room impulse response RT_{60} time is 565ms based on the method described in Section 2.3.1.1 and the BRIRs are sampled at 44.1kHz but are downsampled to 8kHz for the purpose of the experiments in this thesis. An example of BRIRs used in the 2×2 case is shown in Figure 2.7.

2.3.5 TIMIT Acoustic-Phonetic Continuous Speech Corpus

The TIMIT dataset, which is a speech database of phonetically rich speech signals, is widely used throughout this document [56]. The corpus consists of eight different native American English accents with a number of male and female speakers for each accent. Recordings of utterances are provided at 16kHz, but were downsampled to 8kHz for experiments detailed in this thesis. This provides a standard speech library so that results are comparable to other studies.

2.4 Performance parameters

The signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) ratio are defined in [57], and are used throughout the thesis as a measure of separation performance of the estimated sources.

The decomposition of an estimated source signal is based on the following model:

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif}, \tag{2.4.1}$$

furthermore SDR is defined as:

$$SDR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}, \tag{2.4.2}$$

and SIR is defined as:

$$SIR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2}, \tag{2.4.3}$$

where $||\cdot||^2$ denotes the energy of a signal, s_{target} is a measure of the part of the estimated source which can be attributed to a filtered version of the original source, e_{interf} is the interference contribution from other present sources and e_{artif} is anything else that cannot be attributed to the contributions from other sources such as distortion introduced by a BSS algorithm. Effectively, SIR only takes into account the interfering sources affecting an estimated source, however, the SDR also considers interfering sources and in addition takes into account any additive noise within an estimated source and any artifacts (e.g. filtering effects). Throughout the thesis it is assumed of the scaling ambiguity that the sources have the same variance at the microphones so that SDR and SIR is 0dB at the microphone observations.

In addition, it is assumed that there is an allowed FIR filter length of 1024 samples. Separation methods vary throughout the thesis (e.g. in the time and frequency domains) and it was felt that such an allowed filter length gave all methods a fair chance of reaching maximum potential performance whilst allowing a reasonable time delay in the context of the methods presented. By increasing the length of the allowed filter, the length of the subspace projected onto it is increased and thus allows for longer potential time delays. See Section III B of [57], for further details.

The parameters of the decomposition in Equation (2.4.1) are found by:

$$s_{target} = P_{s_j}\hat{s}_j \tag{2.4.4}$$

$$e_{interf} = P_{\mathbf{s}}\hat{s}_j - P_{s_j}\hat{s}_j \tag{2.4.5}$$

$$e_{noise} = P_{\mathbf{s},\zeta}\hat{s}_j - P_{\mathbf{s}}\hat{s}_j \tag{2.4.6}$$

$$e_{artif} = \hat{s}_j - P_{\mathbf{s},\zeta}\hat{s}_j, \tag{2.4.7}$$

where P is a matrix projection on to a subspace, for example P_{s_j} is the projection onto the subspace s_j. The projections allowing for time invariant filters are defined as:

$$P_{s_j} = \mathrm{proj}((s_j^{\tau})_{0\leq\tau\leq L-1}) \tag{2.4.8}$$

$$P_{\mathbf{s}} = \mathrm{proj}((s_{j'}^{\tau})_{1\leq j'\leq N, 0\leq\tau\leq L-1}) \tag{2.4.9}$$

$$P_{\mathbf{s},\mathbf{\imath}} = \mathrm{proj}((s_{j'})_{1\leq j'\leq N}, (\zeta_i^{\tau})_{1\leq i\leq M})_{0\leq\tau\leq L-1}), \tag{2.4.10}$$

where τ is a delay and L, in this case, is the maximum allowed delay, set at 1024 in this thesis unless stated which effectively allows for the length of a projection to be longer, thus allowing for longer delays in the unmixing filters.

2.5 Independent component & vector analyses for convolutive mixtures

2.5.1 Addressing the permutation problem

To address the permutation problem within FD-BSS a straightforward method which measures the distance between unmixing matrices for two adjacent frequency bins is proposed in [58]. The two possible distances between two unmixing matrices for the 2×2 case are measured as:

$$\mathcal{D}_1 = \sum_{i,j} |g_{i,j}^{(k)} - g_{i,j}^{(k-1)}| \tag{2.5.1}$$

$$\mathcal{D}_2 = \sum_{i,j} |g_{i,j}^{(k)(fliped)} - g_{i,j}^{(k-1)}|, \tag{2.5.2}$$

where $G^{(k)(fliped)}$ is found by multiplying by a permutation matrix:

$$G^{(k)(fliped)} = G^{(k)} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{2.5.3}$$

If $\mathcal{D}_2 < \mathcal{D}_1$, then $G^{(k)} \leftarrow G^{(k)(fliped)}$, where ← indicates an assignment, in this case this implies that a permutation between adjacent frequency bins is likely to have occurred. This method of correcting for the permutation problem is written here for the case of two sources $N = 2$. Whilst possible to implement a similar method for $N > 2$, this would add to the complexity as there would be $N!$ possible permutations at each frequency bin thereby increasing the code complexity at each frequency bin; as the target is to potentially implement an online version of this method only $N = 2$ is considered. Previous studies such as [31] discuss restricting the length of the time domain filters to correct for the permutation problem. This method is not considered for FD-ICA in this thesis as a simplistic approach is required.

With IVA it is not necessary to correct for the permutation problem as it directly addresses the permutation problem in its formulation by modelling sources with multivariate probability distributions, thus capturing higher-order inter-frequency dependencies. The batch versions of NG-ICA and NG-IVA are both used as described in Sections 2.1.4 and 2.2.

2.5.2 Choice of window

Consider a system which splits a time domain signal into overlapping blocks, applies a window to the blocks, transforms to the time-frequency domain, performs some processing and transforms it back to the time domain by means of an inverse discrete Fourier transform (DFT) and overlap-add technique. Such a system needs to ensure that the reconstructed time domain signal has constant energy, i.e. perfect reconstruction. The condition for perfect reconstruction for a window with 50% overlap is $\text{window}(t)^2 + \text{window}(t + (K/2))^2 = 1$, where $K/2$ is half the window length, where K is normally the number of frequency bins, however for the purposes of applying a window it also happens to be the length of the window. The window proposed in [35] is a Hann window, described as:

$$\text{window}_{orig}(t) = \left(\frac{1}{2}\Big(1 - \cos\Big(\frac{2\pi t}{K-1}\Big)\Big)\right). \tag{2.5.4}$$

However, [59] proposes the following window:

$$\text{window}_{new}(t) = \sin\left(\frac{\pi}{2}\sin^2\Big(\frac{\pi}{K}\Big(t + \frac{1}{2}\Big)\Big)\right), \tag{2.5.5}$$

by setting the time index and FFT size to example values, e.g. $t = 256$ and $K = 2048$, it is evident that $\text{window}_{new}(t)$ satisfies the condition for prefect reconstruction unlike the Hann window, i.e. $\text{window}(t)^2 + \text{window}(t + (K/2))^2 = 0.0524 + 0.9476 = 1$.

2.5.3 Whitening

It is common in ICA style algorithms to decorrelate mixtures at the microphones (or sensors). The mixtures are whitened so that the observations are uncorrelated with one another and that the variance is set to unit variance, known as whitening or sphering the data. Consequently, the covariance of some data $\mathbf{x}$ becomes the identity matrix $C_{\mathbf{x}} = I$, due to whitening. A whitening matrix can be found using PCA [60], which can be implemented using various techniques including singular value decomposition and eigen analysis. The whitening matrix, Q_w, is defined as:

$$Q_w = \Lambda_{eig}^{-\frac{1}{2}} E_{eig}^{H}, \tag{2.5.6}$$

where Λ_{eig} is a diagonal matrix of eigenvalues and E_{eig} is a matrix whose columns are eigenvectors of a cross-correlation matrix $C_{\mathbf{x}} = E[\mathbf{x}\mathbf{x}^H]$, which is the covariance matrix of the observations, assuming that the observation signals have zero mean.

Typically, for the experiments within this thesis the observation data remains unwhitened, as it is not a strict requirement of IVA and methods in this thesis are derived with online and real-time operation in mind. In such a scenario whitening matrices may not be available at every instance in time. However, a “whitening light” is implemented in some cases where input signal observations are divided by their standard deviation, causing the input signals to have unit variance.

2.5.4 Experimental results

To demonstrate a blind source separation system the techniques discussed so far are employed. Averaged results are given by taking the mean results of 22 mixtures[2] created from a mixture of male-female speakers taken from the TIMIT database. Mixing impulse responses are taken from the BRIR dataset described in [54], where s_1 is placed at 0° and s_2 is placed at 45° at 1.0m away from the centre of a two-microphone array. A distance of 1.0m was chosen as it exceeds the critical distance given by Equation (2.3.4). Both the batch versions of NG-ICA and NG-IVA with two different windows are compared with a preprocessing stage of whitening the observation data. Results are given for a two-microphone two-source scenario in SDR and SIR with an allowed filter tap length of 1024 in Table 2.1. Due to its poorer results for convolutive mixtures ICA style methods are not considered for the remainder of the thesis.

	window$_{orig}$		window$_{new}$	
	SDR (dB)	SIR (dB)	SDR (dB)	SIR (dB)
NG-ICA	5.25	6.30	6.44	7.65
NG-IVA	12.77	14.81	13.16	15.02

Table 2.1. Results for the batch versions of NG-ICA and NG-IVA for convolutive mixtures at 1.0m, where; K=1024, $\eta = 0.001$.

Results in Table 2.1 show improved performance for NG-IVA when compared to NG-ICA, as expected. IVA is much better suited to ensuring that there are fewer permutations within the frequency bins. To show that perfect reconstruction is desirable in a source separation scenario the two different windows proposed in Section 2.5.2 are compared, and significantly improve results. It is interesting to note that when the window proposed in [59] (window$_{new}$) is used, the values for NG-ICA increase by approximately 1.2dB for SDR and 1.3dB for SIR. There is

[2] 22 mixtures were chosen so that they were consistent with the same mixtures in Chapter 4 of this thesis due to practical constraints in obtaining results in Chapter 4.

also an improvement in performance for NG-IVA however it is less pronounced. For the remainder of the thesis when working with frequency domain data which is transformed into the time domain, the proposed window: window$_{new}$, is employed. A 2D schematic of the simulated room layout is shown in Figure 2.8.

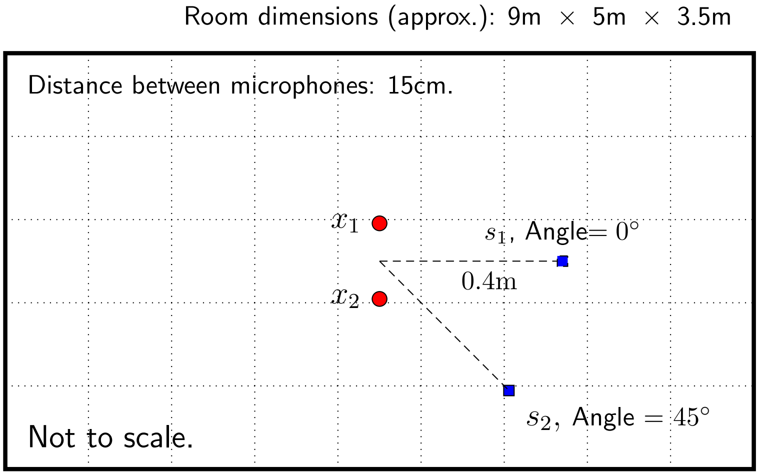
Room dimensions (approx.): 9m × 5m × 3.5m
Distance between microphones: 15cm.
x_1
s_1, Angle= 0°
0.4m
x_2
s_2, Angle = 45°
Not to scale.

Figure 2.8. 2D plan of room setup and locations of sources (blue) and microphones (red). s_1 is placed at 0.4m from x_1 and also x_2.

2.6 Audio-visual blind source separation

Colin Cherry outlined in his original paper how visual information could be used to aid the source separation process [1]. Audio-visual source separation has become an important aspect of the CPP with increasing interest, a recent review of the main techniques [18], highlights several areas of the field. As more and more emerging technologies are likely to be equipped with cameras (e.g. smartphones, wearable technology and robotic human machine interfaces), interest in audio-visual BSS will increase.

Generally, previous audio-visual BSS methods aid the source separation process by exploiting video cues to localise a source. By knowing the location of a source this can improve the convergence speed of algorithms and potentially address the permutation problem when used in conjunction with ICA-style methods.

Previous work into audio-visual source separation for moving and non-stationary sources can be found in [32], [61], where a 3D position tracker is used to identify the location of a source. Then based on this information, an appropriate BSS algorithm is selected depending on the movement of the source; a similar framework to this is proposed in Chapter 5 of this thesis. In [62] a time-frequency masking approach is described that exploits direction of signal arrival. A more complex environment is described in [63] where the number of speakers is not fixed and move in and out of the environment.

Knowledge of the location of the speakers can be estimated using audio techniques rather than video cues [64], [65]. However, audio localisation for simultaneously active speakers in a reverberant room environment is difficult [62], [63].

Other works use the video information differently such as [66], which exploit pauses in speech to identify silent periods so that one source is silenced. Video localisation is also not always effective, especially if a human face is not visible to at least two cameras [32]. Therefore audio-visual modalities with multiple camera integration is the most suitable choice for source localisation, however it is beyond the scope of this thesis.

A video-informed noise source suppression technique using a ‘voxel’ model (a 3D grid with a simulated room) that attempts to exploit the IM can be found in Appendix A.

Potential drawbacks, such as time required to provide an accurate estimate, typically associated with some BSS algorithms, prompts the need for more efficient audio-visual methods, that possibly avoid higher-order statistics, such as will be presented in Chapter 3. In this context, [67] uses assumed video information to enhance a source by adaptive filtering. It is important to reiterate at this point that although this thesis does not directly deal with the identification and tracking of sources it is assumed that this information is available (eg. by implementing one of the video tracking methods found in [32], [61], [62]). An explanation how a video tracking system may work whilst exploiting known speaker locations is given in the following sections and chapter.

2.6.1 Channel Estimation as a Pure Delay

A straightforward way of exploiting speaker location is to use the known location of a speaker and model the acoustic path between the speaker and a microphone as a pure delay (as the full impulse response representing this acoustic path is unknown and is complicated to estimate). Furthermore, it is possible to exploit this pure delay by assuming a mixing matrix, or more likely mixing matrices if operating in the frequency domain, then finding the inverse of these matrices to find a first estimate of the unmixing matrices.

$$\hat{h}_{ml}^{(k)} = e^{-jd\cos(\theta_l)k/c}, \tag{2.6.1}$$

where c is the speed of sound in air, $\hat{h}$ is the estimated pure delay impulse response, θ_l is the angle of arrival and d is the physical distance between a speaker and a microphone. The unmixing matrix at each frequency bin can be initialised by:

$$G_{init}^{(k)} = \begin{bmatrix} \hat{h}_{11}^{(k)} & \cdots & \hat{h}_{1l}^{(k)} \\ \vdots & \ddots & \vdots \\ \hat{h}_{m1}^{(k)} & \cdots & \hat{h}_{ml}^{(k)} \end{bmatrix}^{-1}, \tag{2.6.2}$$

for ICA style methods the initial unmixing matrix is then whitened by a whitening matrix $Q_w^{(k)}$, so that $G_{init}^{(k)} = Q_w^{(k)} G^{(k)}$, this method guides an ICA style method to a solution by exploiting direction of arrival information of the speakers. In FastICA there is noticeable improvement in convergence [32], in addition this also addresses the permutation problem. In [32] the inverse is not explicitly calculated, however this initial 'guess' is whitened, decorrelating the rows of $G^{(k)}$. In the online version of ICA, and indeed in IVA, the full observation matrix X_i for the i-*th* source, needed to calculate the whitening matrix, is not available, therefore other methods of increasing the speed of convergence must be found. In initial experiments with NG-IVA and "intelligent" initialisation, whilst improving speed initially the uninitialised version soon overtook the intelligently initialised version.

2.6.2 Target cancellation by subtraction

The target cancellation method exploits video cues to identify the location of a target source and uses this as a priori information to orientate a microphone array, so that the target source is at an equidistant position from two microphones which work as a pair, as described in [67]. A subtraction of the observations, to find an estimate of $\hat{s}_1$, can be written as:

$$x'_1(t) = x_1(t) - x_2(t) = \hat{s}_1(t) \tag{2.6.3}$$

$$x'_2(t) = x_1(t) + x_2(t). \tag{2.6.4}$$

By exploiting the equidistance property as well as some additional processing of the detected microphone signals, the noise reference (from a second source which is not at a equidistant position between the pair of microphones) is isolated and can be used as a noise reference in an adaptive filtering scheme. The noise reference could be multiple speakers or background noise and the position of the second source is not critical to the functionality of the method. In practice, however, the room environment is likely to be highly reverberant, hence a simple subtraction will not work. Later work in this thesis will address the issue by describing a pair of FIR filters.

2.7 Summary

This chapter presented various previous relevant methods and techniques for convolutive blind source separation, together with some techniques which exploit the video modality to improve and enhance the source separation process. Datasets such as TIMIT, the BRIR database and SDR and SIR performance parameters used throughout this thesis were inroduced.

A summary of the basic techniques involved with source separation systems have been outlined in this chapter. Also, included was a preliminary study of convolutive blind source separation for speech mixtures, comparing NG-ICA and NG-IVA. A highlighted issue was the permutation problem in some FD-BSS methods and how this was addressed in IVA.

The next chapter details novel work in the time domain to find a pair of cancellation filters which remove a source at the observation for a 2×2 case.

Chapter 3

TIME-DOMAIN FILTER DESIGN FOR TARGET SOURCE CANCELLATION

3.1 Introduction

This chapter introduces a time domain null-steering beamforming technique which is shown to cancel a source of interest (target source) from an array of two microphones. Following this, an adaptive filtering process is employed to then use the remaining sources as a noise reference, effectively separating the sources. The motivation for the methods presented in this chapter is to devise an online method which avoids higher order statistics (as opposed to classical BSS methods such as ICA).

By coupling a proposed audio method with a video system providing location of speakers along with formulating the method in the time domain, it is expected that such a system will reduce method computational complexity and overcome the circularity problem [68].

A beamformer, which spatially filters measurements from an array of microphones (or another type of sensor), is often employed to achieve such selectivity [69]–[71]. With broadband signal sources, such as speech, such beamformers are commonly implemented in the frequency domain. In some applications, however, the size of the array can be limited, so only two microphones can be employed. In this context, in [69], a frequency domain generalised sidelobe canceller (GSC) has been proposed. The processing at each discrete frequency, k, in a GSC is represented in Figure 3.1. On the left-hand side of the diagram is a lattice structure, which at the output of the adder enhances the target signal, whereas at the bottom, due to the subtraction, it blocks the target signal so that the input to the adaptive filter nominally contains only other speech signals.

Within the framework of a GSC [69], the signal $u^{(k)}$, in which the target signal has been blocked, is defined as $u^{(k)}(n) = \mathbf{b}^{(k)H}\mathbf{x}^{(k)}(n)$, where n is the time block index of a STFT, $\mathbf{b} = 1/2[e^{j\Delta_k/2}, -e^{-j\Delta_k/2}]^H$ is the blocking vector, $\mathbf{x}^{(k)}(n)$ is a vector of the short-time Fourier transforms of the time-domain quantities $x_{\{1,2\}}$ and Δ_k is the 'uncertainty in angle arrival' which, in this study, is the time shift to correct for delay in signal arrival. Further details can be found in [69].

In this chapter two methods are presented to estimate a pair of time-domain finite impulse response filters which, when included in the GSC framework, suppress any undesired signal components which may pass through the blocking channel due to steering error. This pair of filters helps ensure that energy of the cancelled signal is as small as possible. The combined output of the blocking vector and the pair of cancellation filters is written:

$$u^{(k)}(n) = (\mathbf{w}^{(k)}_{\{1,2\}} \circ \mathbf{b}^{(k)})^H \mathbf{x}^{(k)}(n) \tag{3.1.1}$$

where $\circ$ in this instance denotes the Hadamard product. This pair of

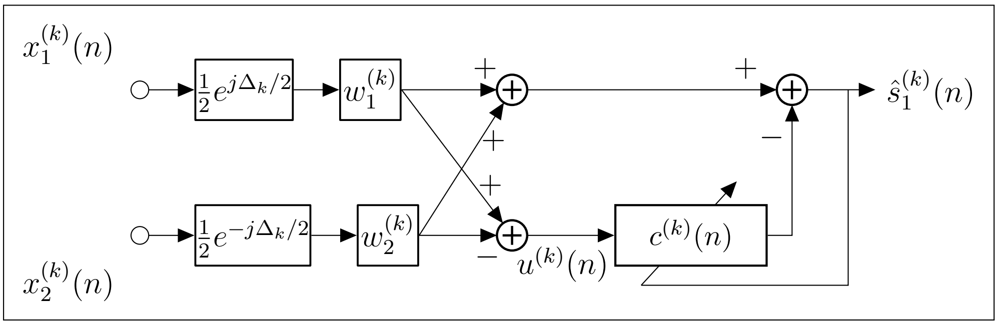

Figure 3.1. Two channel generalised sidelobe canceller in the frequency domain with the addition of the pair of cancellation filters $\hat{\mathbf{w}}_{\{1,2\}}$.

filters ($\mathbf{w}_{\{1,2\}}$) is referred to as 'cancellation filters' in this chapter and the remainder of the thesis, which are also illustrated in Figure 3.1. Figure 3.1 includes also the adaptive filtering stage, where $c^{(k)}(n)$ is a complex frequency domain parameter.

3.2 Why operate in the time domain?

Frequency domain BSS approaches assume that the length of the discrete Fourier transform (DFT) used to convert the time-domain microphone measurements into the frequency domain is significantly longer than the impulse responses of the filters used to model the propagation between the sources and the array microphones.

Due to a DFT a frequency domain mixture, when expressed in the time domain, is only an approximation of linear convolution of the mixing IR and the source signal. In fact, the frequency domain mixture is equivalent to circular convolution in the time domain, this is described mathematically as:

$$\mathbf{x}(t) = H * \mathbf{s}(t) \Longleftrightarrow \mathbf{x}^{(k)}(n) \approx H^{(k)}\mathbf{s}^{(k)}(n) \tag{3.2.1}$$

$$\mathbf{x}(t) = H \circledast \mathbf{s}(t) \Longleftrightarrow \mathbf{x}^{(k)}(n) = H^{(k)}\mathbf{s}^{(k)}(n), \tag{3.2.2}$$

where $*$ denotes convolution, $\circledast$ denotes circular convolution and $\Longleftrightarrow$ denotes the conversion between the time and frequency domains with a DFT and its inverse.

Many frequency domain source separation algorithms which address convolutive mixtures assume Equation (3.2.1). Consequently, these are subject to errors at frame boundaries thereby potentially degrading the separation performance, which is known as the circularity problem. Some studies [72] suggest that the length of the FFT should be at least twice the length of the time domain mixing filters. To avoid any issues of the length of an FFT, the formulation of this method is in the time domain and the circularity problem is avoided.

Suppose that the adaptive filter in Figure 3.1 is operating in the frequency domain. So that it can converge there must be a sufficient number of frequency domain blocks, this requires the impulse responses modelling the propagation environment to be static throughout this period. This assumption is likely to be violated in many applications where the time domain mixing RIR is sufficiently long, and supports operating in the time domain.

In addition, as well as avoiding complex valued signal operations, being formulated in the time domain allows for increased flexibility for being implemented as an online source separation method, thus the proposed overall system is suited to real-time operation.

3.3 Method

3.3.1 Problem Formulation

The observation at each microphone of a two-microphone array can be modeled in the general case in the time-domain as a convolutive mixture from each source of the form:

$$x_j(t) = \sum_{i=1}^{N} h_{ji}(t) * s_i(t), \qquad j = 1, 2, \tag{3.3.1}$$

where s_i is the speech signal generated by the i-th source, h_{ji} is the filter that models the effect of a reverberant environment between the i-th source and the j-th microphone, t is the discrete time index, x_j is the detected signal at the j-th microphone. Throughout the chapter, source number $i = 1$ is the target source that is to be cancelled.

In the training phase the microphones are pre-steered (included in Figure 3.1 as the blocking vector b) so that $\mathbf{h}_{11} \approx \mathbf{h}_{21}$, where $\mathbf{h}_{11} = [h_{11}(1), \ldots, h_{11}(L)]^T$, $\mathbf{h}_{21} = [h_{21}(1), \ldots, h_{21}(L)]^T$, and L is the length of a time domain filter, as this gives the system the best chance of cancelling the target by using the initially observed signals from the microphones. This would be implemented by exploiting the geometry of the acoustic environment by ensuring the position the target signal source is equidistant between the microphones, so that in terms of early reverberation the IRs would essentially be equivalent and the time difference of arrival would small (ideally $\delta_k = 0$). A pair of cancelling filters would then correct for the fact that $\mathbf{h}_{11} \approx \mathbf{h}_{21}$. The operation of the blocking vector, $\mathbf{b}$ is carried out by pre-steering the microphone array towards a target source, and for brevity is dropped in the rest of the thesis is assumed to be automatically acting on the observation vector $\mathbf{x}$.

The core problem formulation is to find a pair of FIR cancelling filters ($\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$), where $\mathbf{w}_1 = [w_1(1), \ldots, w_1(L)]^T$ and $\mathbf{w}_2 = [w_2(1), \ldots, w_2(L)]^T$, so that:

$$u(t) = \sum_{\tau=0}^{L-1} x_1(t-\tau)w_1(\tau) - \sum_{\tau=0}^{L-1} x_2(t-\tau)w_2(\tau) \approx 0. \qquad (3.3.2)$$

A model of the longitudinal wave propagation of a sound pressure wave with respect to the two microphone array is illustrated in Figure 3.2. The sources are assumed to be far field, hence the field front lines have been drawn at a tangent to the direction of sound propagation. The source s_1 is at its training position equidistant to both microphones. Source s_2 is placed at an arbitrary position and the cancellation pair of filters, $\hat{\mathbf{w}}_{\{1,2\}}$, are shown in Figure 3.2.

To find the pair of cancellation filters, an error vector, $\boldsymbol{\epsilon}_1$, is therefore formulated as:

$$\boldsymbol{\epsilon}_1 = \epsilon_1(w_1, w_2) = (X_1\mathbf{w}_1 - X_2\mathbf{w}_2), \qquad (3.3.3)$$

where $\epsilon_1(w_1, w_2)$ is an error function and X_1 and X_2 are the convolution matrices, so that when the matrix is multiplied by a vector the resultant vector is the convolution of $x_j(t)$ and $w_j(t)$, (i.e. $x_j(t) * w_j(t)$) which are formed from $x_1(t)$ and $x_2(t)$ observation signals (which themselves are convolutions of the target source with h_{11} and h_{21} respectively assuming the other sources are silent during training), thus the convolution matrices are formed as a Toeplitz-style matrix structure:

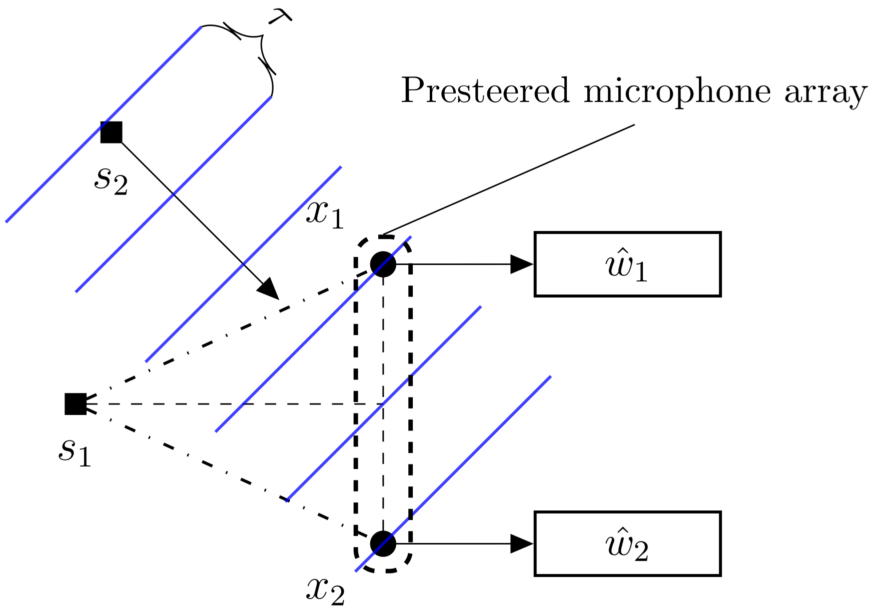

Figure 3.2. 2D plan of microphone and source positions. Blue lines indicate longitudinal wave fronts of speech signals.

$$X_j = \begin{bmatrix} x_j(1) & 0 & \cdots & 0 & 0 \\ \vdots & x_j(1) & \ddots & 0 & \vdots \\ x_j(T) & \vdots & \ddots & x_j(1) & 0 \\ 0 & x_j(T) & \ddots & \vdots & x_j(1) \\ \vdots & 0 & \ddots & x_j(T) & \vdots \\ 0 & 0 & \cdots & 0 & x_j(T) \end{bmatrix} \qquad j \in \{1,2\}, \qquad (3.3.4)$$

where the width of the matrices is the length of the cancellation filter pair to be found, L, and T is the length of the time-domain training data.

3.3.2 Alternating Gradient Descent Method

This method assumes that $\hat{\mathbf{w}}_{\{1,2\}}$ is estimated during a training phase where source s_2 is silent. A cost function for the alternating gradient descent method (GD method) is derived from the error vector from Equation (3.3.3), which yields:

$$J_1 = ||\boldsymbol{\epsilon}_1||_2^2, \{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2\} = \arg \min_{\mathbf{w}_1, \mathbf{w}_2} J_1, \qquad \text{s.t.}||\hat{\mathbf{w}}||_2 = 1. \qquad (3.3.5)$$

The assumption is made that $X_1 \neq X_2$ (i.e. they differ sufficiently so that Equation (3.3.3) cannot be factorised as $X_1(\mathbf{w}_1 - \mathbf{w}_2)$ or $X_2(\mathbf{w}_1 - \mathbf{w}_2)$). Taking the partial derivatives of the cost function, J_1, with respect to the filters to be estimated, $\mathbf{w}_1$ and $\mathbf{w}_2$, yields:

$$\frac{\partial J_1}{\partial \mathbf{w}_1} = 2X_1^T X_1 \mathbf{w}_1 - 2X_1^T X_2 \mathbf{w}_2 \qquad (3.3.6a)$$

$$\frac{\partial J_1}{\partial \mathbf{w}_2} = 2X_2^T X_2 \mathbf{w}_2 - 2X_2^T X_1 \mathbf{w}_1. \qquad (3.3.6b)$$

To minimise the cost function, J_1, the two expressions for the gradient, $\frac{\partial J_1}{\partial \mathbf{w}_1}$ and $\frac{\partial J_1}{\partial \mathbf{w}_2}$ are included in a gradient descent scheme, which updates filter weights according to a change proportional to the gradient of the cost function. Thus, this yields the update equations for the estimated filters:

$$\hat{\mathbf{w}}_1^{\ell+1} = \hat{\mathbf{w}}_1^{\ell} + \eta(X_1^T X_2 \hat{\mathbf{w}}_2^{\ell} - X_1^T X_1 \hat{\mathbf{w}}_1^{\ell}) \tag{3.3.7a}$$

$$\hat{\mathbf{w}}_2^{\ell+1} = \hat{\mathbf{w}}_2^{\ell} + \eta(X_2^T X_1 \hat{\mathbf{w}}_1^{\ell+1} - X_2^T X_2 \hat{\mathbf{w}}_2^{\ell}), \tag{3.3.7b}$$

where $(\cdot)^{\ell}$ denotes the iteration number and η denotes the learning rate. Notice that in Equation (3.3.7a) $\hat{\mathbf{w}}_2$ is fixed and the update is performed with respect to $\hat{\mathbf{w}}_1$, whereas the reverse applies in Equation (3.3.7b), hence this is an alternating descent. Likewise, the order of Equation (3.3.7a) and Equation (3.3.7b) could be reversed (so that $\hat{\mathbf{w}}_2^{\ell+1}$ is found before $\hat{\mathbf{w}}_1^{\ell+1}$ at each iteration). The scale factor of 2 has been factored out and absorbed by η. The condition $||\hat{\mathbf{w}}_2||^2 = 1$ is applied so that the trivial zero solution is avoided, equally $||\hat{\mathbf{w}}_1||^2 = 1$ could also be applied, though only one condition is used so the remaining filter has more freedom to reach its optimised value. This is especially important if the amplitudes of observations x_1 and x_2 are different. The constraint is applied by adding the update equation:

$$\hat{\mathbf{w}}_2^{\ell+1} = \hat{\mathbf{w}}_2^{\ell+1} / ||\hat{\mathbf{w}}_2^{\ell+1}||, \tag{3.3.8}$$

after Equation (3.3.7b). This constrained optimisation corresponds to modifying the cost $J_1 = J_1 + \lambda_{Lag}(||\mathbf{w}_2|| - 1)$, where λ_{Lag} is a Lagrange multiplier. Such an approach to canceller design has been adopted in stereophonic echo cancellation [73], and has been known to exhibit poor convergence due to the correlation between the two signal channels.

An overview of the GD method is in Algorithm 1. The next section introduces a method which exploits the singular value decomposition (SVD) to find the filter pair $\hat{\mathbf{w}}_{\{1,2\}}$ without using an iterative process.

Algorithm 1 Alternating gradient descent method to find the pair of cancellation filters $\hat{\mathbf{w}}_{\{1,2\}}$.

```
Input: Convolution matrices of the microphone observations.
Output: Pair of cancellation filters $\hat{\mathbf{w}}_{\{1,2\}}$.
for $\ell = 1$ to maximum number of iterations do
    $\hat{\mathbf{w}}_1^{\ell+1} \leftarrow \hat{\mathbf{w}}_1^{\ell} + \eta(X_1^T X_2 \hat{\mathbf{w}}_2^{\ell} - X_1^T X_1 \hat{\mathbf{w}}_1^{\ell})$
    $\hat{\mathbf{w}}_2^{\ell+1} \leftarrow \hat{\mathbf{w}}_2^{\ell} + \eta(X_2^T X_1 \hat{\mathbf{w}}_1^{\ell+1} - X_2^T X_2 \hat{\mathbf{w}}_2^{\ell})$
    $\hat{\mathbf{w}}_1^{\ell+1} \leftarrow \frac{\hat{\mathbf{w}}_1^{\ell+1}}{||\hat{\mathbf{w}}_1^{\ell+1}||}$ or $\hat{\mathbf{w}}_2^{\ell+1} \leftarrow \frac{\hat{\mathbf{w}}_2^{\ell+1}}{||\hat{\mathbf{w}}_2^{\ell+1}||}$
end for
return $\hat{\mathbf{w}}_{\{1,2\}}$
```

3.3.3 Principal Angles Method

In a similar fashion to the previous method, the pair of filters, $\hat{\mathbf{w}}_{\{1,2\}}$, is estimated during a training phase. During the training phase only the target signal speech source (s_1) is active whilst the other source (s_2) is assumed to be silent.

The novelty in this approach is that the method of principal angles (PA method) is exploited to find the filter estimates ($\hat{\mathbf{w}}_1$ and $\hat{\mathbf{w}}_2$), as described in [74] and originally proposed in [75], which should overcome the slow convergence in the gradient descent method. An orthonormal basis for the convolution matrices is needed to implement the method of principal angles; taking the QR decomposition of X_1 and X_2, yields $X_1 = Q_1 R_1$ and $X_2 = Q_2 R_2$. The QR decomposition decomposes a matrix into an orthonormal matrix (Q, so that $Q^T Q = I$) and an upper triangular matrix (R). The error vector is rewritten as;

$$\boldsymbol{\epsilon}_2 = \epsilon_1(w_1, w_2) = \epsilon_2(\tilde{w}_1, \tilde{w}_2) = (Q_1\tilde{\mathbf{w}}_1 - Q_2\tilde{\mathbf{w}}_2), \tag{3.3.9}$$

where $\epsilon_1(w_1, w_2)$ and $\epsilon_2(\tilde{w}_1, \tilde{w}_2)$ are error functions. Also, $\tilde{\mathbf{w}}_1 = R_1\mathbf{w}_1$ and $\tilde{\mathbf{w}}_2 = R_2\mathbf{w}_2$. The minimisers of a new cost function are then found as:

$$J_2 = ||\boldsymbol{\epsilon}_2||_2^2,\ \{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2\} = \arg\min_{\tilde{\mathbf{w}}_1.\tilde{\mathbf{w}}_2} J_2, \tag{3.3.10}$$

subject to $||\tilde{\mathbf{w}}_1||_2 = ||\tilde{\mathbf{w}}_2||_2 = 1$. Unlike the gradient descent method discussed previously, the two constraints can be applied simultaneously as there is no longer a problem with the amplitudes due to the orthonormal basis. To find the principal angles and principal vectors of the orthonormal subspaces Q_1 and Q_2, the singular value decomposition is taken of $Q_1^TQ_2$, so that $[U, \Lambda, V^T] = SVD(Q_1^TQ_2)$. The constraints $||\tilde{\mathbf{w}}_1||_2 = 1$ and $||\tilde{\mathbf{w}}_2||_2 = 1$ are inherently introduced to the method by exploiting the properties of the SVD avoiding the trivial solution $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2 = \mathbf{0}$. The cost function J_2 is rewritten as:

$$\begin{aligned} J_2 &= ||Q_1\tilde{\mathbf{w}}_1 - Q_2\tilde{\mathbf{w}}_2||_2^2 && (3.3.11)\\ &= \tilde{\mathbf{w}}_1^T\tilde{\mathbf{w}}_1 + \tilde{\mathbf{w}}_2^T\tilde{\mathbf{w}}_2 - 2\tilde{\mathbf{w}}_1^TQ_1^TQ_2\tilde{\mathbf{w}}_2, && (3.3.12)\end{aligned}$$

therefore reducing J_2 is equivalent to maximising $\tilde{\mathbf{w}}_1^TQ_1^TQ_2\tilde{\mathbf{w}}_2^T$ as $\tilde{\mathbf{w}}_1^T\tilde{\mathbf{w}}_1 = 1$ and $\tilde{\mathbf{w}}_2^T\tilde{\mathbf{w}}_2 = 1$, thus:

$$\arg\min_{\tilde{\mathbf{w}}_1.\tilde{\mathbf{w}}_2} J_2 \equiv \arg\max_{\tilde{\mathbf{w}}_1.\tilde{\mathbf{w}}_2} \tilde{\mathbf{w}}_1^TQ_1^TQ_2\tilde{\mathbf{w}}_2. \tag{3.3.13}$$

By exploiting the SVD:

$$(\tilde{\mathbf{w}}_1^T Q_1^T Q_2 \tilde{\mathbf{w}}_2) = \tilde{\mathbf{w}}_1^T (U\Lambda V^T)\tilde{\mathbf{w}}_2 = \tilde{\mathbf{w}}_1^T (\sum_m \lambda_m \mathbf{u}_m \mathbf{v}_m)\tilde{\mathbf{w}}_2, \tag{3.3.14}$$

by selecting $\tilde{\mathbf{w}}_1 = \mathbf{u}_1$ and $\tilde{\mathbf{w}}_2 = \mathbf{v}_1$ in Equation (3.3.14), where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the vectors from the rows of U and V which correspond to the largest largest singular value, denoted λ_1, where the subscript $(\cdot)_1$ denotes the largest singular value. In turn, λ_1 corresponds to the smallest angle between the orthonormal bases Q_1 and Q_2 [74].

The equalising filters are the columns of U and V which correspond to λ_1 (as they maximise Equation (3.3.14)), multiplied by the inverse of R_1 and R_2 to allow for the basis change introduced by the QR decomposition, thus:

$$\hat{\mathbf{w}}_1 = R_1^{-1}\mathbf{v}_1 \tag{3.3.15a}$$

$$\hat{\mathbf{w}}_2 = R_2^{-1}\mathbf{u}_1, \tag{3.3.15b}$$

therefore the filter pair $\hat{\mathbf{w}}_{\{1,2\}}$ has been estimated. The full PA method is described in Algorithm 2.

3.3.4 Normalised Least Mean Square

An NLMS algorithm is employed to recover the target source s_1 after it has been cancelled from one of the microphone observations. First proposed in [76], the least mean square (LMS) adaptive filter is considered to be a 'classic' adaptive filter and is well-known in the field of signal processing [77].

A brief derivation of the normalised least mean square (NLMS)

Algorithm 2 Principle angles method to find the pair of cancellation filters $\hat{\mathbf{w}}_{\{1,2\}}$. Note that V_1 and U_1 denote the columns of V and U which correspond to the largest singular value, λ_1.

```
Input: Convolution matrices of the microphone observations.
Output: Pair of cancellation filters $\hat{\mathbf{w}}_{\{1,2\}}$.
1: $[Q_1, R_1] \leftarrow QR(X_1)$
2: $[Q_2, R_2] \leftarrow QR(X_2)$
3: $[V, \Lambda, U] \leftarrow SVD(Q_1^T Q_2)$
4: $\mathbf{v} \leftarrow V_1$
5: $\mathbf{u} \leftarrow U_1$
6: $\hat{\mathbf{w}}_1 \leftarrow R_1^{-1}\mathbf{v}$
7: $\hat{\mathbf{w}}_2 \leftarrow R_2^{-1}\mathbf{u}$
8: return $\hat{\mathbf{w}}_{\{1,2\}}$
```

algorithm is provided for completeness, for details see [77], [78]. A new error, ϵ_3, is formed as:

$$\epsilon_3(t) = d(t) - \mathbf{x}'^T(t)\mathbf{c}(t), \tag{3.3.16}$$

where $(\cdot)'$ notates an altered version of the original microphone observations (outputs of the lattice structure in Figure 3.1) and $\epsilon_3(t)$ is different from the error vectors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, as it is not a vector and represents the error at the adaptive filtering stage rather than the target cancellation stage. $\mathbf{c}(t)$ is written here as a time domain quantity and is defined by $\mathbf{c}(t) = [c(1), \ldots, c(L)]^T(t)$, i.e. the vector of filter co-efficients at discrete time t, which for clarity is reintroduced for this section.

The weights of the filter are updated as follows:

$$\mathbf{c}(t+1) = \mathbf{c}(t) + \eta_{\mathsf{LMS}}\hat{\nabla}(t), \tag{3.3.17}$$

where $\hat{\nabla}$ is the instantaneous gradient estimate of the error function at time t. To find $\hat{\nabla}$, the mean square error $(||\epsilon_3||_2^2)$ is minimised by taking the partial derivatives;

$$\hat{\nabla}(t) = \left[\frac{\partial \epsilon_3(t)^2}{\partial c_1}, \ldots, \frac{\partial \epsilon_3(t)^2}{\partial c_L}\right]^T = -2\epsilon_3(t)\mathbf{x}'(t), \tag{3.3.18}$$

combining Equation (3.3.17) and Equation (3.3.18), the update rule for the LMS algorithm becomes:

$$\mathbf{c}(t+1) = \mathbf{c}(t) + \eta_{\mathsf{LMS}}\epsilon_3(t)\mathbf{x}'(t), \tag{3.3.19}$$

where the factor of 2 is absorbed by η. To form the NLMS algorithm, the update term is divided by the energy of the input signal, therefore Equation (3.3.19) is altered:

$$\mathbf{c}(t+1) = (\beta)\mathbf{c}(t) + \frac{\eta_{\mathsf{NLMS}}(1-\beta)}{||\mathbf{x}'(t)||^2}\epsilon_3(t)\mathbf{x}'(t), \tag{3.3.20}$$

where β is a positive constant (usually $0.9 < \beta < 1.0$). Equation (3.3.20) describes the normalised least mean square update rule, with a constant β which has the effect of updating $\mathbf{c}$ with a weighted version of instantaneous gradient, causing the algorithm to have a 'memory' of previous values between time indices.

The performance of the two cancellation filter design approaches is compared in the next section.

3.4 Experimental Setup

Firstly, the pair of cancellation filters employed as a null-steering beamformer are evaluated in the general case, and in the second half of the section results are presented to show how these filters along with assumed speaker locations can be used in speech source separation.

Sampling rate (f_s)	8kHz
Reverberation time (RT_{60})	565ms
Learning rate (η_{NLMS})	0.275
Memory factor (β)	0.9
Length of $w_{\{1,2\}}$	810 samples
Length of $\mathbf{c}$	1024 samples
Angles tested	{15°, 45°, 75°}
TIMIT speakers used	{faks0, mbjk0, fjre0, mdab0}

Table 3.1. Experimental conditions for PA method for source separation results.

Twelve mixtures created from four different speakers (two male, two female) taken from the TIMIT dataset were used, with all available utterances for each speaker concatenated to form longer speech signals of 246 seconds. Binaural room impulse responses (BRIRs) from a classroom were measured with a dummy head between two microphones [54] and then resampled to 8kHz. See Table 3.1 for full experimental conditions. A two-dimensional room plan is shown in Figure 3.3, where s_1 is at $0.4m$ from the centre of the microphone array.

3.5 Experimental Results

3.5.1 Cancellation Filter Performance

The cancellation filter methods were compared by calculating the value of the respective cost functions with the estimated filter vectors for the PA method and the GD method, i.e. $||X_1\hat{\mathbf{w}}_1 - X_2\hat{\mathbf{w}}_2||^2$ for both methods. BRIRs and speech signal inputs are used to train the cancellation filters, where the target signal source was positioned at 0°, and at a distance 40cm from the centre of the microphone array as marked in Table 3.2.

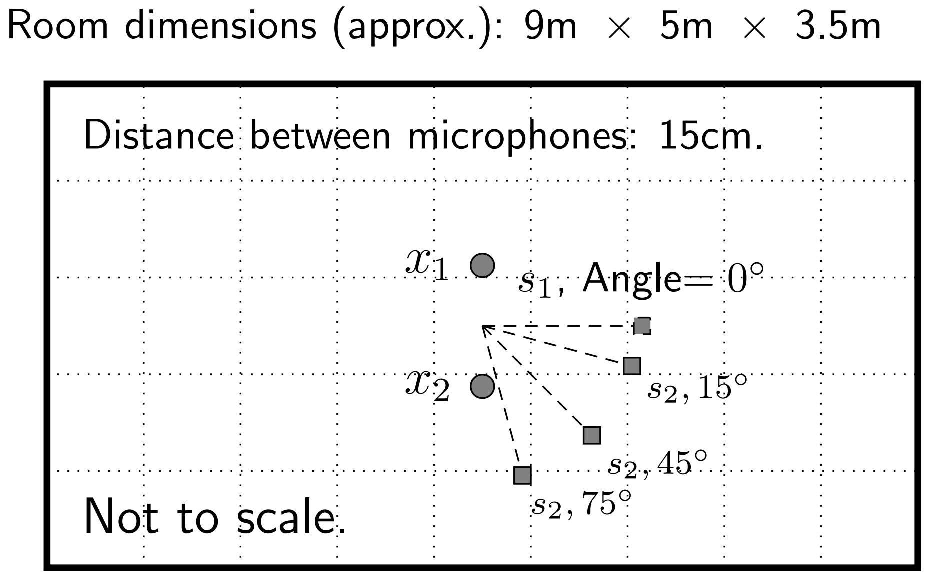
Room dimensions (approx.): 9m × 5m × 3.5m
Distance between microphones: 15cm.
x_1
s_1, Angle= $0°$
x_2
$s_2, 15°$
$s_2, 45°$
$s_2, 75°$
Not to scale.

Figure 3.3. 2D room plan of microphone and source positions.

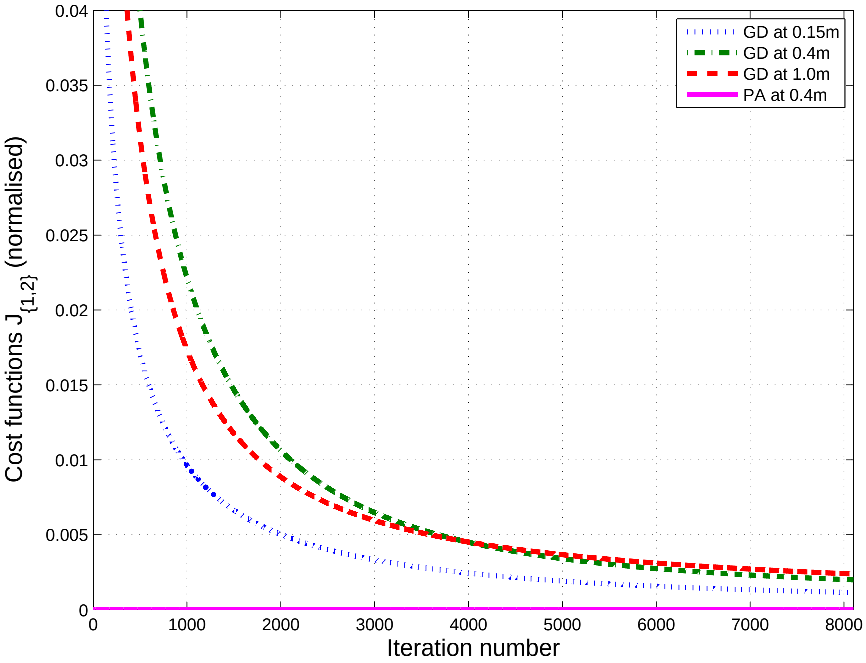

Figure 3.4. Convergence performance of the GD method, where $\eta = 1 \times 10^{-6}$. The strong correlation between both microphone signals x_1 and x_2, as they share a common source convolved with similar IRs, cause slow convergence. Also included is the cost function value achieved by the PA method which is much nearer zero.

Strong correlation between the microphone signals x_1 and x_2 causes slow convergence for the GD method as shown in Figure 3.4. Normalised values of the cost function, J_1, are given after 8100 iterations of the update equations where the cancellation filters' lengths are 810 taps. To train the cancellation filters for Figure 3.4, speech signals of 10000 samples were used which was chosen to limit the size of X_1 and X_2 to save on computational load. The plots corresponding to s_1 at 0.4m and 1.0m in Figure 3.4 overlap at approximately iteration 4000, it is suggested this is due to the random nature, inherent in audio source separation, of the mixing impulse responses used for this particular experiment.

Table 3.2 shows values for J_1 (GD method) and J_2 (PA method). From Table 3.2 it can be seen that the principal angles method offers better performance than the GD method. Values are shown for normalised cost, that is to say the initial value from the cost function divided by the length of the estimated filter. The slow convergence of the alternating GD method also reduces performance. Lower normalised cost function values could be achieved if the GD method was run for more iterations, but this would introduce a potential delay in real-time systems. This is an advantage of the PA method as it finds the optimal filters without the need of update iterations. For a training length of 8000 samples and filter length of 810, the PA method has an execution time of 5.44 seconds, which increases to 467.67 with training length of 24000 samples and filter length of 810 on a desktop PC running MATLAB.

Cancellation filters of different lengths were calculated for the PA method, then λ_1 was found from the SVD of $Q_1^T Q_2$. This corresponds to

Distance (m)	PA (normalised cost)	GD (normalised cost)
0.15	4.04×10^{-8}	1.42×10^{-2}
0.40	4.04×10^{-8}	1.38×10^{-2}
1.00	4.04×10^{-8}	1.38×10^{-2}

Table 3.2. Values of the cost function with estimated filters, for various distances from the centre of the microphone array. Filters of length 810 were estimated for both methods.

the smallest angle between the two orthonormal subspaces (the inverse cosine is taken as this represents the dot product) and thereby leads to the best cancellation performance. As expected, the performance improves as the length of the cancellation filters increases, see Figure 3.5. Note that speech signals are used to estimate the cancellation filters.

3.5.2 Principal Angles as a Beamformer

Figures 3.6 and 3.7 show the beampattern of the outputs of the lattice structure (adders in Figure 3.1) with the filter pair $\hat{\mathbf{w}}_{\{1,2\}}$. Figure 3.7 gives the output for the fixed beamformer channel (top output of lattice structure in Figure 3.1). Figure 3.6 gives the response of the blocking channel (top output of lattice structure (subtract) in Figure 3.1), as described in [71]. As the filter lengths of $\hat{\mathbf{w}}_{\{1,2\}}$ are relatively long and because there are only two microphones in the array, this causes several sidelobes which can be seen in Figures 3.7 and 3.6. Figure 3.6 shows a null at 0°, therefore the filter structure acts as a null-steering beamformer. The beampatterns are defined in the frequency domain

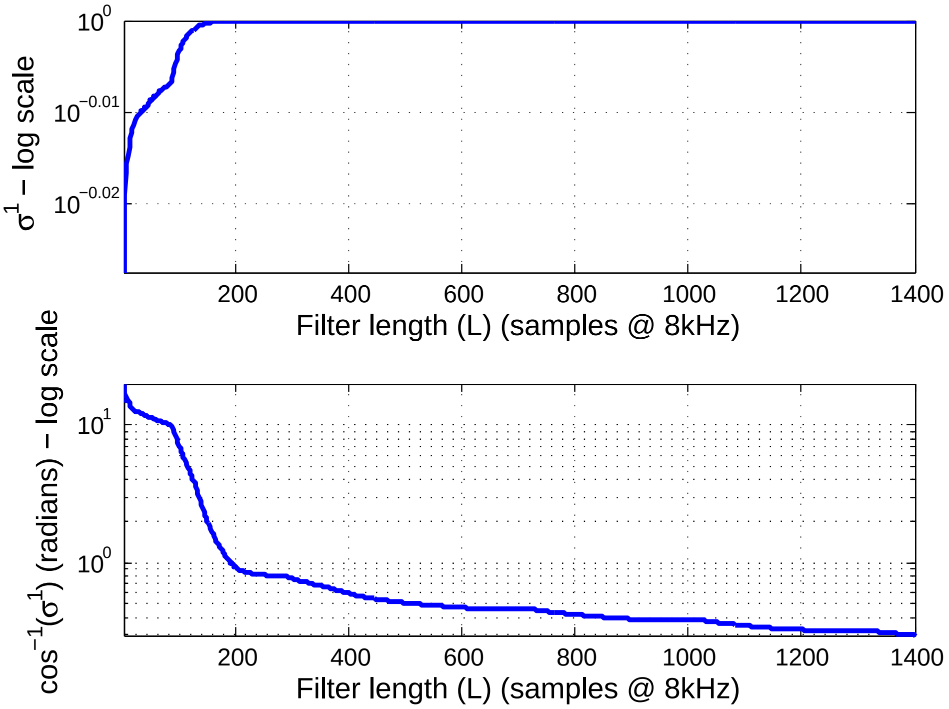

Figure 3.5. Angle between the pair of filters $\hat{\mathbf{w}}_{\{1,2\}}$. The top plot is the largest singular value from the SVD and the bottom plot is the largest singular value expressed in radians. The cancellation filter pair $\hat{\mathbf{w}}_{\{1,2\}}$ is trained with 16000 samples of speech.

as:

$$r_1^{(k)}(\theta) = (\hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2)^H \boldsymbol{\psi}^{(k)}(\theta) \tag{3.5.1a}$$

$$r_2^{(k)}(\theta) = (\hat{\mathbf{w}}_1 + \hat{\mathbf{w}}_2)^H \boldsymbol{\psi}^{(k)}(\theta), \tag{3.5.1b}$$

where $\boldsymbol{\psi}$ is defined as:

$$\boldsymbol{\psi} = [1, e^{j2\pi k\tau_2(\theta)}, e^{j2\pi k\tau_3(\theta)}, \ldots, e^{j2\pi k\tau_N(\theta)}]^T, \tag{3.5.2}$$

where $\tau_i(\theta)$ and $i = \{2 \ldots K\}$, are time delays due to propagation and any tap delays from the zero phase reference to the point at which the i-th weight is applied.

3.5.3 Video-Informed Source Separation Application

In this section the results for the PA method with an adaptive filtering scheme are presented as an alternative to classical higher-order statistics source separation methods. An array of two microphones is pre-steered towards the target so that IRs between a speaker and the two microphones, which are positioned close together (0.15m), are approximately equal, $\mathbf{h}_{11} \approx \mathbf{h}_{21}$, as the microphones are the same distance from the target source.

The microphones are assumed to be pre-steered by video information which provides the location of the target speech source. In practice, a microphone array would be orientated towards the target source using a mechanical device. The use of video information is much more robust to background noise than an audio based method for source localisation. The extraction of localisation information from video information for pre-steering the array is outside the scope of this method,

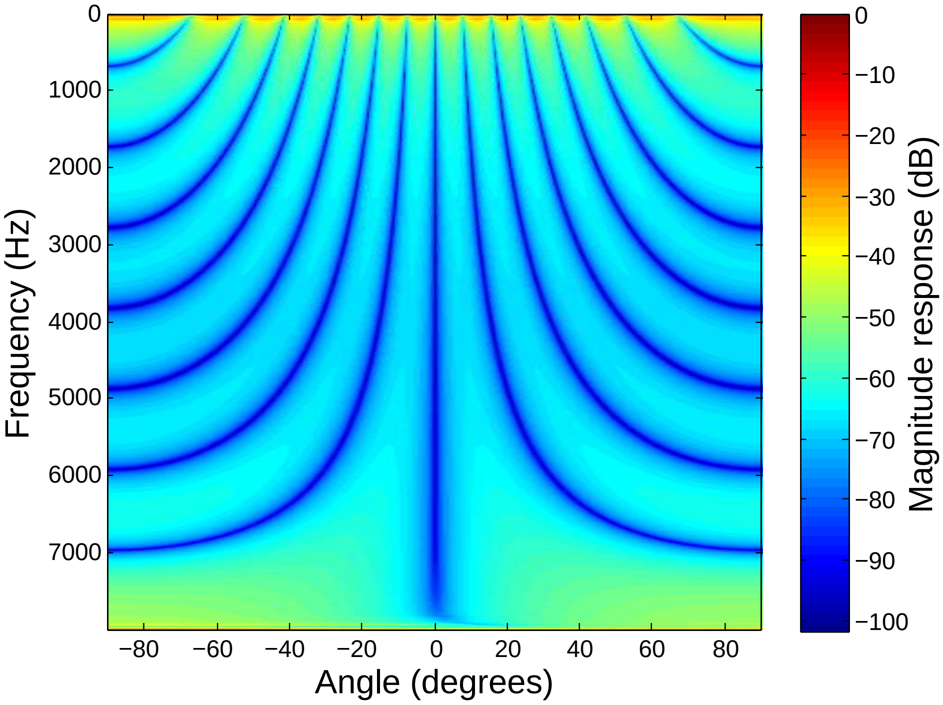

Figure 3.6. Output of null-steering beamformer for $r_1^{(k)}(\theta)$ (blocking channel), when the lengths of cancellation filters are $L = 1100$, trained with BRIRs. The null created by the blocking vector is clear at 0°.

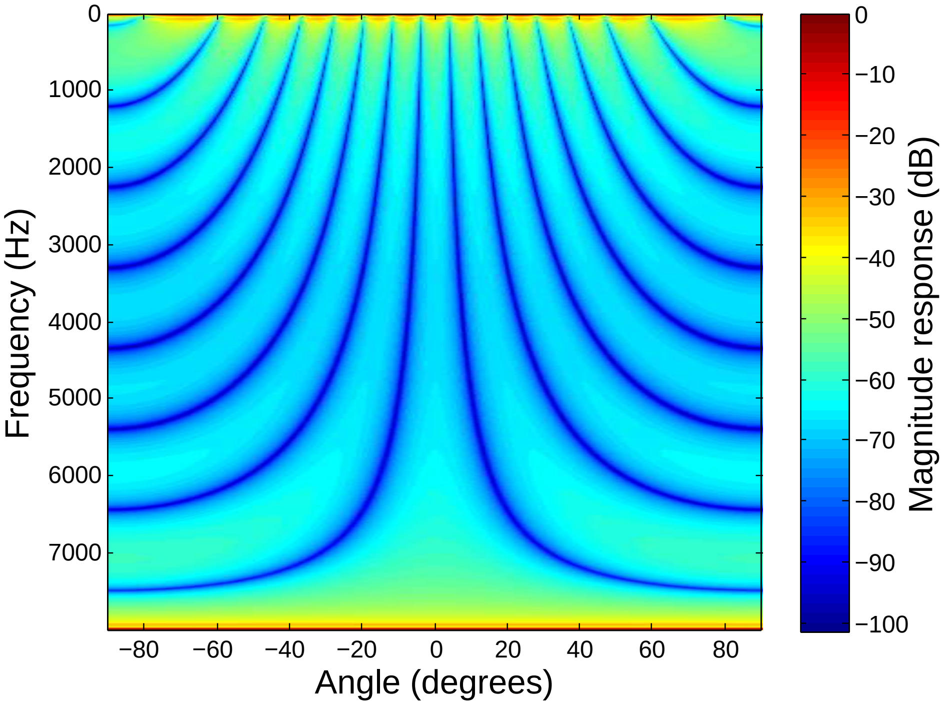

Figure 3.7. Output of null-steering beamformer for $r_2^{(k)}(\theta)$, when the lengths of cancellation filters are $L = 1100$, trained with BRIRs.

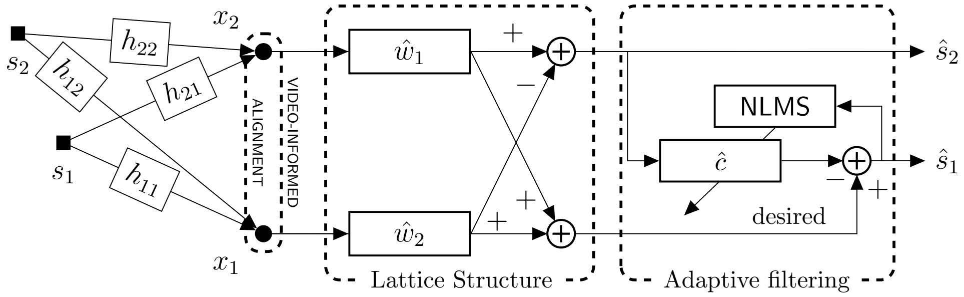

Figure 3.8. System overview for the two-microphone configuration, the microphone array is pre-steered towards s_1, the cancellation filters and an NLMS adaptive filtering stage. The overall outputs are $\hat{s}_1$ and $\hat{s}_2$. The dashed boxes indicate the various subsystems of the overall audio-video system including the video formed alignment.

but further details can be found in [32], [61], [62].

In the training phase, the same BRIRs and speech signal inputs as before are used to create mixtures at each microphone, where only the target source is present. The estimated cancellation filters, $\hat{\mathbf{w}}_{\{1,2\}}$, are found for an angle of 0° and a distance of 0.40m from the centre of the microphone array. After the training phase, the second source (s_2) is then added at 15°, 45° and 75° and 0.40m from the array.

At a particular distance, the target source ($\hat{s}_1$) is cancelled from the mixture leaving the other source ($\hat{s}_2$) which is employed as a noise reference. The cancelled target source s_1 is then recovered by using $\hat{s}_2$ as a noise reference in a NLMS adaptive filtering scheme. A diagram of the full system, including GSC lattice structure and NLMS adaptive filtering scheme, is given in Figure 3.8, including the mixing process.

The method is evaluated in the two-microphone two-source scenario, see Tables 3.3 to 3.5. Average performance values for and SDR and SIR are given for mean values of $\hat{s}_1$ and $\hat{s}_2$.

Table 3.3 shows the mean SDR and SIR results for sources at 0°

and 75°. Mean values are given for the geometric mean of the original performance ratios (i.e. not in a log scale). The SIR values are particularly good, mostly due to the successful separation of the source $\hat{s}_2$ (18.52dB - 25.09dB).

Unusually, the average SIR values are rather high. These values have been calculated by averaging across two sources where one estimated source (in this case $\hat{s}_2$) is especially (>20dB) good and estimated source $\hat{s}_1$ is lower some cases (<5dB). The adaptive filtering stage does not recover the source $\hat{s}_1$ adequately due to the statistically non-stationary nature of speech, an issue which is addressed in later chapters. Also, despite moving the source s_2, there seems to be little change in performance values in Table 3.4 and Table 3.5, indicating that the position of s_2 has little effect on the proposed method.

The filters $\hat{\mathbf{w}}_{\{1,2\}}$ and the room impulse responses cause the outputs of the algorithm $\hat{s}_1$ and $\hat{s}_2$ to be filtered versions of the original sources. However, significant SIR is achieved with a peak value of 27.18dB. The additional filtering on both estimated sources $\hat{s}_1$ and $\hat{s}_2$ causes lower average SDR values, however the effect can be reduced by additional post-processing, as in [66]. The improved SIR ratios also suggest that signal leakage is not a major problem in the operation of the adaptive filter.

3.6 Summary

Two methods have been proposed for designing time-domain cancellation filters. The more conventional alternating gradient descent based method was shown to converge slowly and to perform badly in terms of the cost function value, even after a significant number of update

	SDR (dB)	SIR (dB)
Mixture 1	8.532	18.52
Mixture 2	9.598	19.73
Mixture 3	9.777	20.01
Mixture 4	11.19	21.56
Mixture 5	8.869	19.56
Mixture 6	7.636	20.49
Mixture 7	13.81	25.09
Mixture 8	11.53	22.06
Mixture 9	12.33	22.04
Mixture 10	11.85	22.38
Mixture 11	11.09	22.07
Mixture 12	11.90	22.69

Table 3.3. Averaged batch results when s_1 is at 0° and s_2 is at 75° and 0.4m away from the centre of the microphone array.

	SDR (dB)	SIR (dB)
Mixture 1	8.046	18.97
Mixture 2	8.819	18.87
Mixture 3	9.298	19.26
Mixture 4	10.79	20.91
Mixture 5	8.788	19.19
Mixture 6	7.124	20.05
Mixture 7	11.73	23.57
Mixture 8	10.98	20.86
Mixture 9	10.67	20.54
Mixture 10	11.39	21.68
Mixture 11	8.739	21.43
Mixture 12	10.94	21.91

Table 3.4. Averaged batch results when s_1 is at 0° and s_2 is at 45° and 0.4m away from the centre of the microphone array.

	SDR (dB)	SIR (dB)
Mixture 1	10.58	22.35
Mixture 2	10.65	20.64
Mixture 3	10.84	20.7
Mixture 4	12.65	23.62
Mixture 5	10.94	21.84
Mixture 6	11.35	23.00
Mixture 7	14.93	27.18
Mixture 8	14.10	24.57
Mixture 9	13.09	23.99
Mixture 10	12.24	24.75
Mixture 11	13.40	24.20
Mixture 12	13.44	24.47

Table 3.5. Averaged batch results when s_1 is at 0° and s_2 is at 15° and 0.4m away from the centre of the microphone array.

iterations. An alternative novel method of principal angles was introduced, which minimises the cost function without the need of iterative updates and gives a much lower cost function value.

Both methods are formulated in the time-domain to ensure that any IR of a particular environment can be adequately covered by the cancellation filters. Once the pair of cancellation filters have been estimated, they become a part of a GSC style structure.

The behavior of the resulting system is applied to a source separation context. The method may be used as a stand-alone source separation method, for a two-source two-microphone scenario, or can be used as a pre-processing stage for a more conventional blind source separation algorithm in the under-determined case ($M < N$).

It can be argued that the time domain method proposed is advantageous as the permutation problem, typically associated with FD-BSS methods, is inherently avoided by not operating in the frequency domain.

The method is provided as a 'proof of concept' method, as results shown are given for averaged (mean) SDR and SIR, while results for the target signal, s_1, are not always consistent due to the adaptive filtering stage of the method. A drawback of the proposed method is that the PA method exploits the SVD, which has a computational complexity of L^3 (where L is the length of the cancellation filters). Whilst being acceptable for a training phase within a method, such computational complexity may not be acceptable in online and real-time systems (particularly if there is a need to implement such a system on a low performance embedded system). In the following chapter an online method which does not require a training phase is presented.

Chapter 4

INDEPENDENT VECTOR ANALYSIS IN REAL-TIME WITH STUDENT'S T SOURCE PRIOR

4.1 Introduction

A major problem for FD-ICA is the permutation ambiguity across all frequencies inherent to FD-BSS. In [35], therefore, FD-IVA was introduced which directly addresses the permutation problem by maintaining the dependencies between the frequency bins in the algorithmic formulation. By using a multivariate super-Gaussian distribution as the source prior the resulting score function maintains dependencies between frequency bins, unlike one based upon a univariate distribution as used in ICA style methods.

Previously, an online (thus real-time) version of NG-IVA was formulated in [79] which exploits the multivariate super-Gaussian distribution. This algorithm is discussed further in [80] where an expectation-maximisation approach is used to estimate the source prior. The auxiliary version of IVA implemented in real-time in [81]. In addition, various implementations of online/real-time ICA based techniques are presented in [58], [82]–[86] (but these all have to address the permutation problem by means of various post-processing techniques, potentially adding a significant computational complexity when implemented on an embedded system, such as a digital signal processor).

In this chapter a Student's t source prior is introduced for the first time in online IVA and therefore incorporated into the online NG-IVA algorithm. Distributions with heavier tails are more suited to speech [87], [88], particularly voiced utterances, as they better model the dependency between higher amplitude data points in a frequency domain speech signal [89]. This differs from the multivariate super-Gaussian distribution as originally proposed in the original formulation [35][1]. If a special condition of the multivariate super-Gaussian, the bivariate version, is considered, this has a Laplacian shaped marginal distribution when one given value is set to zero, and is Gaussian-like otherwise. This dependency in shape makes the multivariate super-Gaussian distribution suitable for modelling the interrelationships in frequency domain speech signals.

However, the heavier tails of the Student's t source prior means that it is better suited to frequency domain speech signals, particularly voiced utterances. The proposed source prior was implemented within online NG-IVA as an embedded application on a Texas Instruments digital signal processing platform and will be shown to perform well in terms of separation performance when compared to the original online

[1] The author is aware that the term 'super-Gaussian' could be considered vague in this context as it represents a family of probability density functions, however the original literature [35] uses this for a specific probability density function, which is defined later in this chapter. For consistency this terminology is used in this thesis.

NG-IVA algorithm. The importance of choosing a suitable value for the degrees of freedom for the Student's t distribution is also discussed.

4.2 Method

4.2.1 Online Natural Gradient Independent Vector Analysis

The derivation of online NG-IVA is similar to that of the batch version of NG-IVA introduced in Chapter 2. The block index (n) is introduced to the unmixing model to emphasise the iterative nature over time of the online version.

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} g_{ij}^{(k)}[n] x_j^{(k)}[n] \tag{4.2.1}$$

where $g_{ij}^{(k)}[n]$ is the unmixing coefficient at time block n, frequency bin k between source i and microphone j, $\hat{s}_i$ is the i-th estimated source of N estimated sources, and $x_j^{(k)}[n]$ is the observation value at time block n and frequency bin k.

The cost function (J_{IVA}) of the IVA algorithm uses the Kullback-Lieber divergence (denoted by $\mathcal{KL}(\cdot)$) between the joint probability distribution of the estimated sources and the product of their marginal probabilities as a measure of independence:

$$\begin{aligned} J_{IVA} &= \mathcal{KL}(p(\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N) || \prod_{i=1}^{N} q(\hat{\mathbf{s}}_i)) && \text{(4.2.2a)} \\ &= \text{const.} - \sum_{k=1}^{K} \log|\det G^{(k)}| - \sum_{i=1}^{N} E[\log q(\hat{\mathbf{s}}_i)] && \text{(4.2.2b)} \end{aligned}$$

where $q(\cdot)$ is an approximated pdf of the original sources, see chapter 2 for a full derivation of NG-IVA.

To minimise the cost function (J_{IVA}) a natural gradient approach is employed by taking the partial derivatives with respect to the individual separating filter coefficients ($g_{ij}^{(k)}$), the increments for the filter co-efficients are given by:

$$\Delta g_{ij}^{(k)} = -\frac{\partial J_{IVA}}{\partial g_{ij}^{(k)}} = g_{ij}^{(k)-H} - E[\varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})x_j^{(k)*}], \tag{4.2.3}$$

where $[(G^{(k)-1})^H]_{ij} = g_{ij}^{(k)-H}$, $(\cdot)^*$ denotes the complex conjugate and $(\cdot)^H$ denotes a Hermitian transpose. Then by multiplying by the scaling matrices to find the natural gradient, it follows that:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{N} (\delta_{il} - E[\varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})\hat{s}_i^{(k)*}])g_{lj}^{(k)}, \tag{4.2.4}$$

where δ_{il} is the Kronecker delta, i.e. when $i = l$, $\delta_{il} = 1$, and zero otherwise. The expectation in equation (4.2.4) is dropped to form the online block wise algorithm and thus yields:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{N} (\delta_{il} - \varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})\hat{s}_l^{(k)*})g_{lj}^{(k)}, \tag{4.2.5}$$

which gives the instantaneous estimate of the gradient and is the major difference between the original (batch) NG-IVA and the online version in this chapter.

The non-linear score function ($\varphi(\cdot)$), which maintains the dependencies between frequency bins, is given in the general case by

$$\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) = -\frac{\partial \log q(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)})}{\partial \hat{s}_i(k)}. \tag{4.2.6}$$

A nonholomonic constraint is also implemented as in [79], meaning that the direction of the update equation is restricted, therefore (4.2.5), becomes:

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{N} (\Lambda_{il}^{(k)} - \varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})\hat{s}_l^{(k)*})g_{lj}^{(k)}, \tag{4.2.7}$$

where $\Lambda_{ii}^{(k)} = \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)})\hat{s}_i^{(k)*}$ and zero otherwise (i.e. $\Lambda_{il}^{(k)} = 0$). $\Lambda^{(k)}$ is a diagonal matrix based on the non-linear score function. Faster convergence performance of online NG-IVA is thereby generally observed as the diagonal elements of $\Lambda_{il}^{(k)} - \varphi^{(k)}(\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})\hat{s}_l^{(k)*}$ (i.e. when $i = l$) are always zero and are therefore more robust to fast changes in input energy level. The introduction of the nonholomonic constraint has a practical advantage as it reduces the complex multiplications at each frequency bin by N. The block-wise update equation, which includes a gradient normalisation, for the separating filter coefficients is given by:

$$g_{ij}^{(k)}[n+1] = g_{ij}^{(k)}[n] + \eta\sqrt{(\xi^{(k)}[n])^{-1}}\Delta g_{ij}^{(k)}[n], \tag{4.2.8}$$

where η is the learning rate. The normalisation factor ($\xi^{(k)}[n]$) is defined as:

$$\xi^{(k)}[n] = \beta\xi^{(k)}[n-1] + (1-\beta)\sum_{i=0}^{M}|x_i^{(k)}[n]|^2/M, \tag{4.2.9}$$

where β is the smoothing factor ($\beta < 1$). The normalisation factor also improves the robustness of the algorithm as it is more tolerant to sudden changes in input signal energy by dividing by the sample root mean square (RMS) of the input signal. Practically, a small constant γ is added to the term $(\xi^{(k)}[n])^{-1}$ in Equation 4.2.8, where $\gamma << 1$, to avoid the case $(\xi^{(k)}[n])^{-1} = \infty$. The following section introduces the alternative Student's t source prior.

4.2.2 Alternative Student's t Source Prior

A Student's t multivariate pdf is proposed as an alternative to the original super-Gaussian source prior. This alternative source prior improves the modelling of the dependency between the high amplitude data points in a frequency domain speech signal, that is characteristic of the signals around the formant frequencies of vowel sounds in human speech. Thus the heavier tails of the new source prior match such frequency domain speech signals more accurately than the original super-Gaussian source prior.

The heavier tails can be seen in the univariate version of the Student's t distribution; various values of the degrees of freedom parameter (υ) were plotted with the original super-Gaussian distribution as a comparison (Figure 4.1).

The original score function is derived on the basis of a multivariate super-Gaussian distribution, given by:

$$q(\mathbf{s}_i) \propto exp\Bigg(- \Big((\mathbf{s}_i - \mu_i)^H \mathbf{\Sigma}_i^{-1} (\mathbf{s}_i - \mu_i) \Big)^{\frac{1}{2}} \Bigg), \tag{4.2.10}$$

and by setting the mean to zero and the covariance matrix to the identity matrix (as the frequency bins are uncorrelated due to the orthogonality of Fourier bases). The original non-linear score function is given as:

$$\varphi^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}. \tag{4.2.11}$$

As in [90], a multivariate Student's t distribution takes the form:

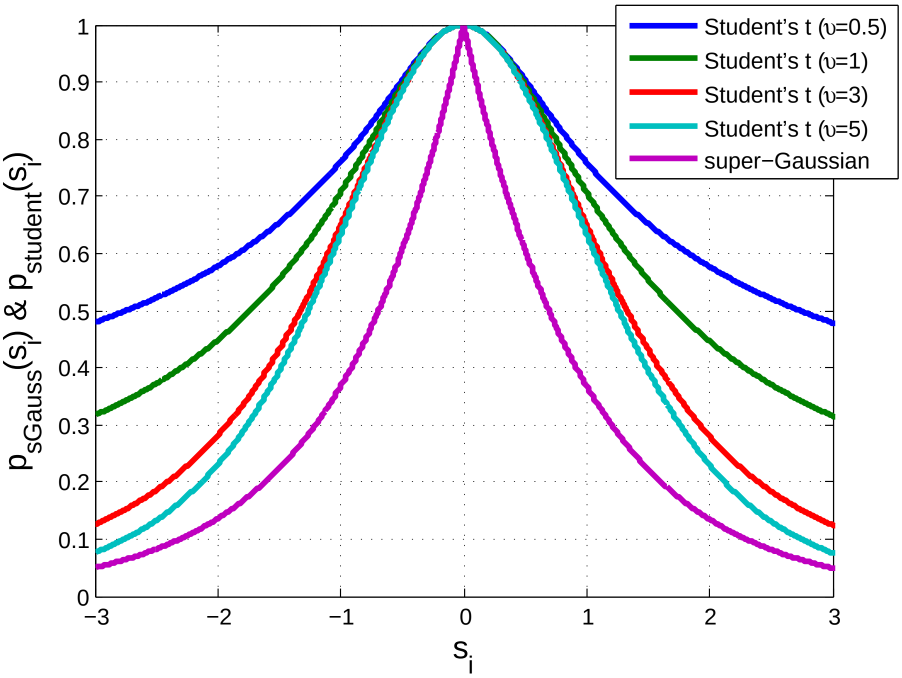

Figure 4.1. Univariate version of the Student's t distribution with different degrees of freedom parameter (υ), with a univariate super-Gaussian distribution.

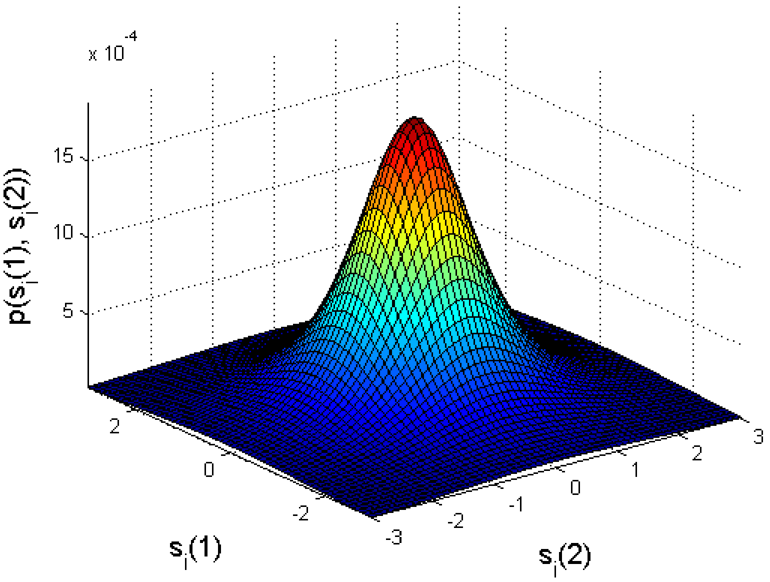

Figure 4.2. Bivariate example of the Student's t distribution with the degrees of freedom parameter υ set to two. This could represent the real or imaginary part of complex frequency domain data.

$$q(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^H \mathbf{\Sigma}_i^{-1} (\mathbf{s}_i - \mu_i)}{\upsilon}\right)^{((\upsilon+K)/2)}. \tag{4.2.12}$$

An example of the multivariate version of the Student's t distribution is plotted in Figure 4.2. The degrees of freedom parameter (υ) controls the leptokurtic nature of the pdf. As υ decreases the tails become heavier whereas as it increases the pdf becomes more Gaussian-like.

Similar to the original super-Gaussian multivariate source prior, the multivariate Student's t can be shown to model the higher-order dependencies between frequency bins in IVA as $p(\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_N) \neq \prod_{i=1}^{N} q(\hat{\mathbf{s}}_i)$, i.e. the product of the marginal distributions is not equal to the joint distribution when the covariance matrix is diagonal, therefore the joint distribution is dependent.

By assuming zero mean ($\mu_i = \mathbf{0}$) and setting the covariance matrix ($\mathbf{\Sigma}_i$) to the identity matrix (due to the orthogonality of the Fourier bases), a new non-linear score function is derived which replaces equation (4.2.11):

$$\varphi_{new}^{(k)}(\hat{s}_i^{(1)} \dots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{1 + (1/v)\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}. \tag{4.2.13}$$

The choice of the degrees of freedom (v) becomes important in the online version of NG-IVA as will be shown in the results section.

4.3 Experimental Setup

4.3.1 Floating point TI TMS320C6713 platform

The online version of IVA was implemented on a Texas Instruments TMS320C6713 floating point digital signal processing platform (TI DSP) (Figure 4.3). Features of the board include a TI C6713 floating point digital signal processor (Harvard architecture), an AIC23 codec (analogue to digital converter (ADC)), 16 MB of external memory, line-in/out socket and headphone in/out socket. This TI DSP processor differs from other microcontroller systems as it has been optimised for digital signal processing, through hardware multiply accumulate (MAC) provision, together with hardware circular and bit-reversed addressing capabilities [91].

The NG-IVA algorithm was implemented in C [92], [93] using the fast Fourier transform (FFT) code provided by TI [94]. Not including the FFT code, the approximate time to execute the update equations (4.2.1), (4.2.7) - (4.2.9) and (4.2.13) for one time block for 2048 frequency bins was 43ms (approx 9.8 million instruction cycles).

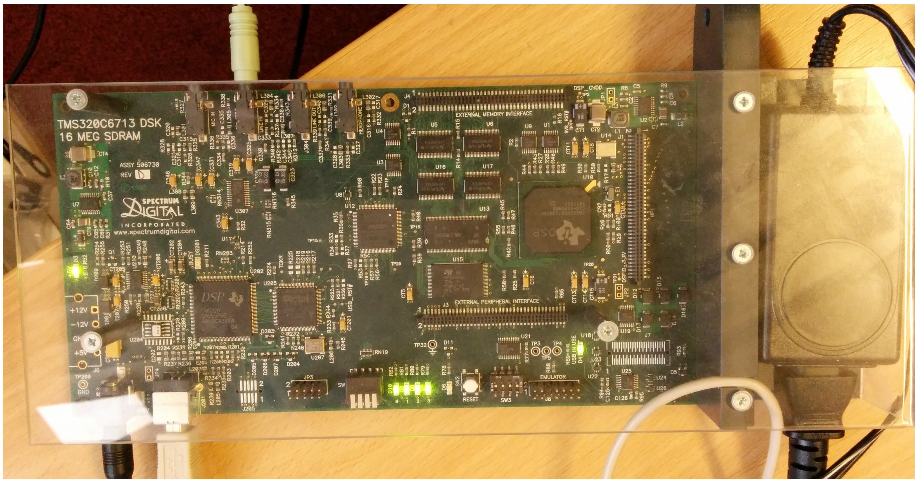

Figure 4.3. Texas Instruments TMS320C6713 floating point digital signal processing development board.

4.3.1.1 Two-Channel FFT Implementation

As the TI DSP performs the update operations of online IVA in the frequency domain, forward FFTs and reverse FFTs need to be executed in real-time (i.e. transforming the input time domain data into the frequency domain, operating the update equations and transforming the output frequency domain data back into the time domain, all need to be finished before the arrival of the following time frame). To achieve this a procedure to process two FFTs with one pass of an FFT with a small computational overhead is implemented. The description for this "buy one get one free trick" is as follows; consider two real time domain signals (such as two channel inputs to a digital signal processor), which are defined as $a(t)$ and $b(t)$ (note that t is the discrete time index rather than the time block index), and are denoted, $A^{(k)}$ and $B^{(k)}$ in the frequency domain. A complex valued signal $z(t)$ is defined as:

$$z(t)_{re} = a(t) \tag{4.3.1a}$$

$$z(t)_{im} = b(t), \tag{4.3.1b}$$

where the real valued signals $a(t)$ and $b(t)$ have been interleaved to form a complex valued signal $z(t)$, which at first sight makes no sense, however it will be shown how this can been used to exploit periodicity in the frequency domain. The signals $a(t)$ and $b(t)$ can be expressed in terms of $z(t)$:

$$a(t) = \frac{z(t) + z^*(t)}{2} \tag{4.3.2a}$$

$$b(t) = \frac{-j(z(t) - z^*(t))}{2}. \tag{4.3.2b}$$

Taking the DFT (implemented by an FFT algorithm), yields:

$$A^{(k)} = \frac{FFT[z(t)] + FFT[z^*(t)]}{2} \tag{4.3.3a}$$

$$B^{(k)} = \frac{-j(FFT[z(t)] - FFT[z^*(t)])}{2}. \tag{4.3.3b}$$

The DFT of $z^*(t)$ is considered:

$$\begin{aligned} FFT[z^*(t)] &= \sum_{t=0}^{T-1} z^*(t)e^{-j2\pi kt/T} \\ &= \left\{ \sum_{t=0}^{T-1} z(t)e^{+j2\pi kt/T} \right\}^* \\ &= \left\{ \sum_{t=0}^{T-1} z(t)e^{-j2\pi(T-k)t/T} \right\}^*, \end{aligned} \tag{4.3.4}$$

where T is the length of the discrete time signal, therefore $FFT[z^*(t)] = \{Z^{(T-k)}\}^*$. $A^{(k)}$ and $B^{(k)}$ can now be found with only one pass of an FFT, with a small computational overhead to reverse the sequence of the frequency domain signal and T complex multiplications.

$$A^{(k)} = \frac{Z^{(k)} + \{Z^{(T-k)}\}^*}{2} \tag{4.3.5a}$$

$$B^{(k)} = \frac{-j(Z^{(k)} - \{Z^{(T-k)}\}^*)}{2}. \tag{4.3.5b}$$

The Fourier transform of the composite signal is defined by :

$$Z^{(k)} = A^{(k)} + jB^{(k)}. \tag{4.3.6}$$

Taking the inverse FFT, yields:

$$z(t) = a(t) + jb(t), \tag{4.3.7}$$

thus, in a similar fashion to the forward FFT, the inverse only requires one pass of the inverse FFT with a small computational overhead. The main motivation for using this two-channel procedure is to increase code speed to ensure all necessary processing is finished before the arrival of the next time frame. See Algorithm 3 for a description of the full update equations and the method for one time frame including FFTs.

Algorithm 3 Online IVA update function and forward/reverse FFTs, *update_iva*$(\cdot)$

```
Input: Time domain data from input buffer of length $T$ (i.e. two concatenated buffers of length $T/2$).
Output: Time domain data to be moved into the output buffer of length $T$.
 1: Calculate the two-channel forward FFT; $\mathbf{x}_{\{1,2\}}^{\{1,\ldots,K\}}[n] \leftarrow FFT(\mathbf{x}_{\{1,2\}}(t))$
 2: for  each frequency bin (k), do
 3:     $\hat{s}_i^{(k)}[n] \leftarrow \sum_{j=1}^{N} g_{ij}[n]^{(k)} x_j^{(k)}[n]$
 4:     Implement the source prior: $\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \leftarrow \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}$, or
        $\varphi_{new}^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \leftarrow \frac{\hat{s}_i^{(k)}}{1+(1/v)\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}$
 5:     $\Delta g_{ij}^{(k)} \leftarrow \sum_{l=1}^{N} (\Lambda_{il} - \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \hat{s}_l^{(k)}) g_{lj}^{(k)}$
 6:     $\xi^{(k)}[n] \leftarrow \beta \xi^{(k)}[n-1] + (1-\beta) \sum_{i=0}^{M} |x_i^{(k)}[n]|^2 / M$
 7:     $g_{ij}^{(k)}[n+1] \leftarrow g_{ij}^{(k)}[n] + \eta \sqrt{(\xi^{(k)}[n] + \gamma)^{-1}} \Delta g_{ij}^{(k)}$, where $\epsilon << 1$.
 8: end for
 9: Increment time block index $(n)$.
10: Calculate two-channel reverse FFT; $\hat{\mathbf{s}}_{\{1,2\}}(t) \leftarrow FFT^{-1}(\hat{\mathbf{s}}_{\{1,2\}}^{\{1,\ldots,K\}}[n])$
```

When implementing step 4 in Algorithm 3 there is a choice between using the original source prior (a multivariate super-Gaussian) and the proposed (multivariate Student's t), the proposed form will be shown to yield better performance, however the original is worth considering as it is possible to use the fast inverse square root [95] which has the potential to increase code speed and is optimised on several embedded platforms, including TI C67x series assembly language where it is available as a single assembly instruction [93]. Ease of implementation and power consumption are worth considering when implementing on power sensitive systems which require batteries such as hearing aids [96].

4.3.2 Methodology and Room Layout

Online NG-IVA was tested on a two-speaker, two-sensor scenario. To recreate a realistic room environment, BRIRs of 565ms as used in [54] were employed, and speakers were placed in two configurations, configuration 'A' where s_1 is at 0° and s_2 is at 45° and configuration 'B' where s_1 is at 0° and s_2 is at 30°, see Figure 4.4 for a 2D room plan and the locations of sources and microphones employed; the BRIRs were convolved with speech files from randomly selected individual speakers across three accents from the training part of the TIMIT dataset. Table 5.1 details the full experimental conditions for the TI DSP, including

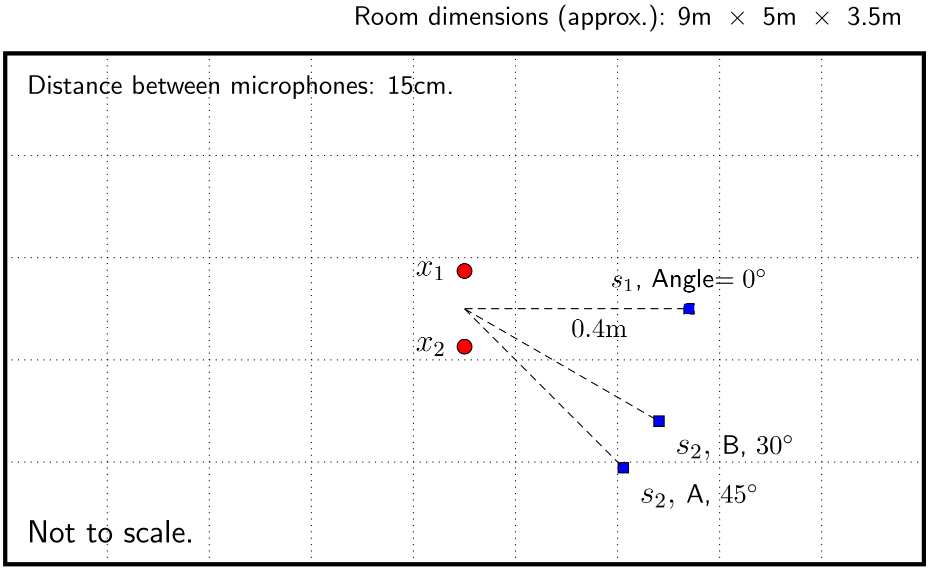
Room dimensions (approx.): 9m × 5m × 3.5m
Distance between microphones: 15cm.
x_1
x_2
s_1, Angle= 0°
0.4m
s_2, B, 30°
s_2, A, 45°
Not to scale.

Figure 4.4. 2D plan of room setup and locations of sources (blue) and microphones (red).

Sampling rate (f_s)	8kHz
FFT length (K)	2048
Reverberation time (RT_{60})	565ms
Window function	$win = \sin\left(\frac{\pi}{2}\sin^2\left[\frac{\pi}{2K}\left(k+\frac{1}{2}\right)\right]\right)$
Overlap ratio	50%
Student's t learning rate (η)	1.0 (with a scaling factor of 200)
Super-Gaussian learning rate (η)	0.4
Degrees of freedom considered (υ)	[0.5, 1.0, 2.0, 2.652]
s_1 position	0° at 0.4m
s_2 position 'A'	45° at 0.4m
s_2 position 'B'	30° at 0.4m

Table 4.1. Experimental conditions for TI DSP results.

values for the learning rate (η) and degrees of freedom parameter (υ) which are discussed in the experimental results section.

Male and female speakers were swapped between two positions which were both 0.4m away from the microphone array at 0° and 45° (for s_2 position 'A') relative to the centre of the array. Utterances from each speaker across different accents were selected to form the anechoic recordings of the speech sources, the utterances were then concatenated to form longer speech signals, up to 300s (i.e. each speaker was repeating what they were saying with the full range of utterances available for that speaker). These speech mixtures were then played via a PC sound card into the line in of the TI DSP for processing, the separated sources were audible via headphones attached to the headphone out jack of the TI DSP.

Performance of the separated mixtures are based on two measurements, namely the SIR and SDR, as the original speech sources are available. SIR takes into account the interfering sources affecting an estimated source, whereas the SDR also considers interfering sources and in addition takes into account any artefacts (e.g. filtering effects) and any noise (e.g. due to quantisation error from the ADC) within an estimated source.

Unmixing matrices (G) were saved at every five seconds in the external memory of the TI DSP. The results given are based on the unmixing matrices obtained and simulated with (unrepeated) speech mixtures, this was to ensure that there were no problems with aligning the estimated sources with the original speech signals in time (which is necessary for accurate results with the BSS Evaluation Toolbox [57]) induced by any latency delay between the PC sound card and the TI DSP. In the following chapter SDR and SIR values are calculated with a moving window over the estimated signal in the time domain, that yields a slight difference in steady state performance.

4.4 Experimental Results

A comparison of averaged SIR and SDR convergence plots for different values of υ with a 50% window overlap are shown in Figures 4.5 and 4.6 respectively, where η was chosen for fast convergence for a range of mixtures and values of υ without the algorithm becoming unstable. The value for η was kept constant for all Student's t plots ($\eta = 1.0$, with a scaling factor of 200). For comparison a typical performance curve for the super-Gaussian source prior, and a learning rate was chosen so that initial convergence is similar to that of the case of $\upsilon = 1.0$ for the Student's t distribution. In this case a learning rate of $\eta = 0.4$ was chosen for the super-Gaussian source prior, so that it had similar initial convergence to the Student's t source prior. Larger values of υ make online NG-IVA converge faster but seem more erratic in the steady state, and in some cases there is a notable deterioration in SDR/SIR performance. This is a typical example of a trade-off between fluctuation in the steady state with larger learning rate values and longer convergence time with smaller learning rate values. It is assumed in a realistic scenario that speakers would remain essentially physically stationary for at most 300 seconds, hence is the reason why results are shown up to 300 seconds.

The convergence plots in Figure 4.5 and 4.6 show that the Student's t distribution performs better after convergence. Plots have been calculated by taking the arithmetic mean of the SDR and SIR ratios of the results over 22 mixtures. Results were limited to 22 mixtures in this case as processing results from the TI DSP was time consuming as the TI DSP needed to be supervised as results were being obtained in real-time. There was a degradation in performance when compared to the results calculated with MATLAB, this could be for a variety of factors including SDR and SIR results being based on unmixing matrices saved every five seconds meaning that time domain signals are not exactly reconstructed as they are from the headphone out jack of the TI DSP (rather than basing the results on a sliding time window over the 'raw' reconstructed time signal). It is also noted in other studies that there is a slight degradation in performance when such algorithms are implemented [97]. In addition there was scaling within the TI DSP possibly introduced by the AIC23 codec (analogue to digital converter), consequently a scaling factor of 200 was introduced into the update equation (Equation 4.2.8) to ensure an acceptable convergence rate.

Convergence of the algorithm with the online NG-IVA algorithm with Student's t source prior is confirmed (in MATLAB) in Figure 4.7, and for the original source prior in Figure 4.8 by calculating the

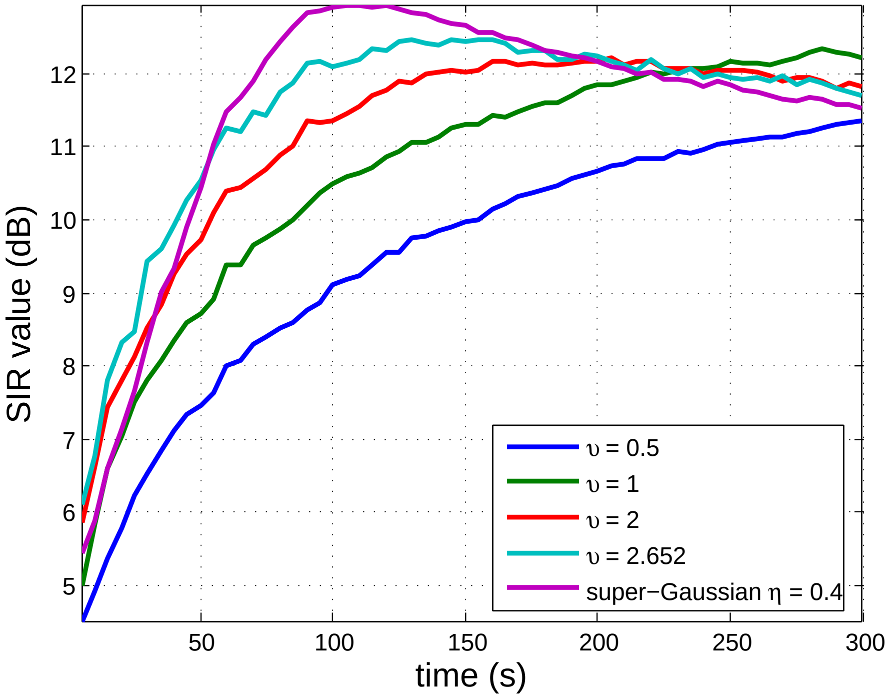

Figure 4.5. SIR convergence for different values of υ. η=1.0, except for the super-Gaussian plot where it is 0.4. Plots have been averaged over 22 mixtures which include male and female speakers.

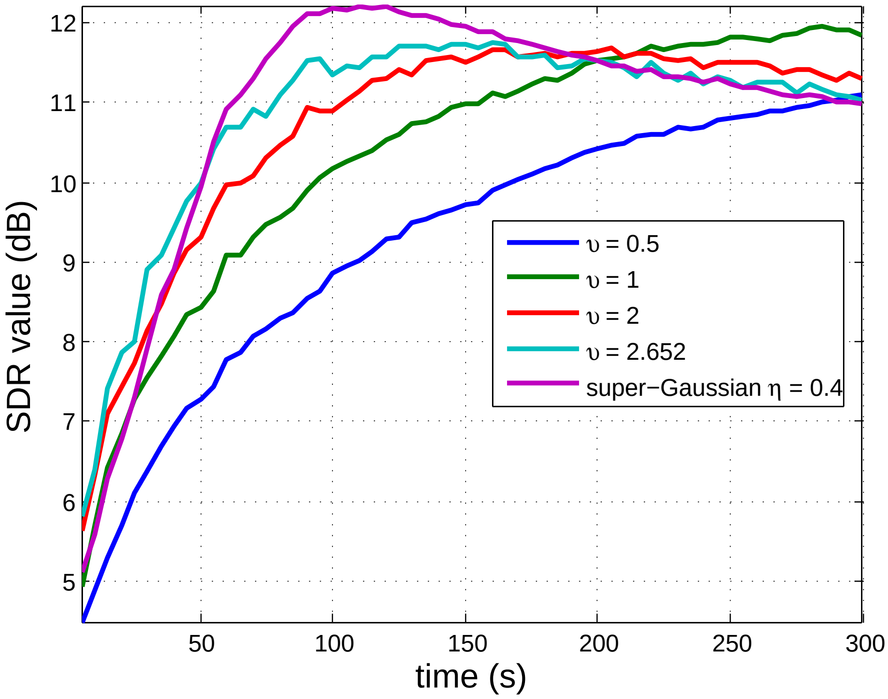

Figure 4.6. SDR convergence for different values of υ. η=1.0, except for the super-Gaussian plot where it is 0.4. Plots have been averaged over 22 mixtures which include male and female speakers.

mean-squared sum (MSS) of the instantaneous gradient of the unmixing matrices, given by:

$$MSS = \frac{1}{KMN} \sum_{i,j,k} |\Delta g_{i,j}^{(k)}[n]|^2. \tag{4.4.1}$$

Figures 4.7 and 4.8 confirm the convergence of the algorithm with both source priors when considering the mean-squared sum, however the instantaneous gradient with the original source prior seems more erratic especially in the early stages of convergence, whereas the instantaneous gradient of the proposed source prior settles quickly.

For details of the two speakers positions, see Figure 4.4, and Table 5.1. Results for online NG-IVA with 50% overlap with s_2 in position 'A' are given in SDR (Figure 4.9) and SIR (Figure 4.10) and show improved performance of approximately 1dB in SDR and 0.75dB in SIR, for the proposed source prior over a period of approximately 300 seconds. The error bars in Figures 4.9 and Figure 4.10 correspond to the standard deviation. The standard deviation improvement is approximately 0.2dB in SDR and 0.3dB in SIR.

However convergence time is not as good as that in [35] (where a steady performance state is reached in within 20 seconds), there are two reasons for this; realistic reverberant binaural room impulse responses are used in this experimental setup, rather than room impulse responses generated by the image method [53] which are highly artificial in nature. Secondly, in an attempt to ensure that the FFT has a sufficient length to cover the length of the time-domain room impulse response, more frequency bins are used for the unmixing filters (2048, compared to 256), thus it takes longer for all the unmixing filters to converge for all

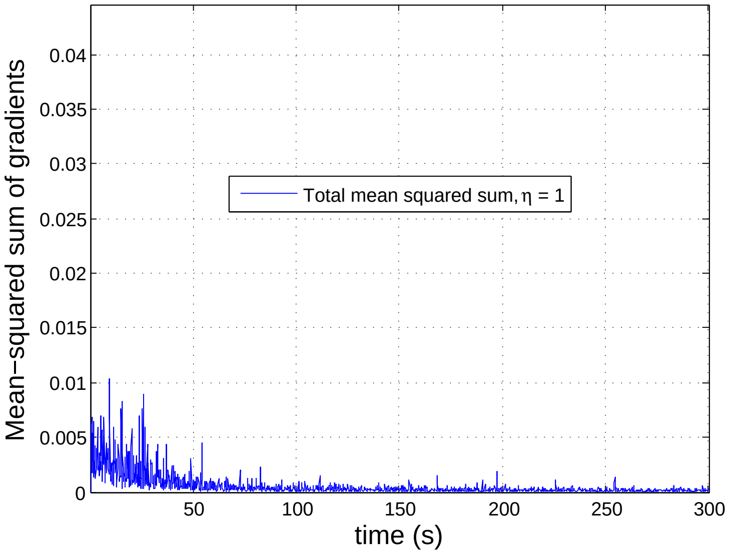

Figure 4.7. Mean-squared sum of the instantaneous gradient of online IVA with score function based on the Student's t source prior, averaged over 22 mixtures. (N.B. no scaling factor was necessary as the variances of x_1 and x_2 were set to one.)

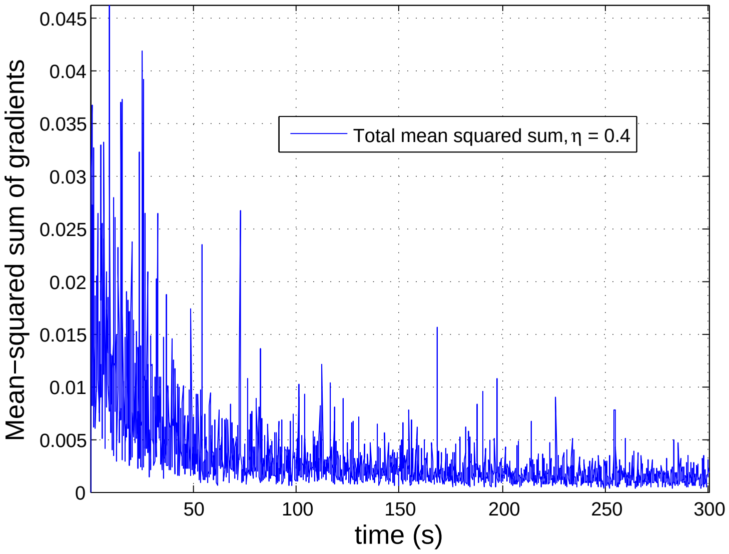

Figure 4.8. Mean-squared sum of the instantaneous gradient of online IVA with score function based on the original source prior, averaged over 22 mixtures.

frequency bins. Convergence time, particularly for moving sources, is addressed in the following chapter.

SDR and SIR results are also given for s_2 at 30° with respect to the centre of the microphone array (position ‘B’) in Figures 4.11 and 4.12. As s_1 and s_2 are spatially closer together there is an observable drop in performance when compared to position A, for example steady state SDR performance in position A is approximately 12dB, compared to a steady state performance of 9.5dB in position B. Likewise, the error bars in Figures 4.11 and Figure 4.12 correspond to the standard deviation. In terms of SDR and SIR (Figures 4.11 and 4.12) there was a noticeable reduction in standard deviation using a Student’s t source prior, approximately 0.3dB and 1.0dB, respectively.

A point of interest in Figures 4.9, 4.10, 4.11 and 4.12 is that the super-Gaussian source prior initially overshoots, the explanation offered for this is as frequency domain unmixing matrices are subsampled from the TI DSP every five seconds. Therefore, when the unrepeated time domain signal estimates are reconstructed, there is a bias towards one of the sources due to the varying length of the unrepeated time domain sources depending on the way sources are either cropped or padded with zeros to ensure that the full range of speech utterances for one speaker is the same length as full range of speech utterances for the other speaker.

In addition, the potential combined effect of latency delay (between the PC sound card) and delay between the input and output to the TI DSP, means that it is challenging to align the time domain output of the TI DSP with the mixtures played via a PC sound card, and was not considered feasible in this study. Simulations within MATLAB do

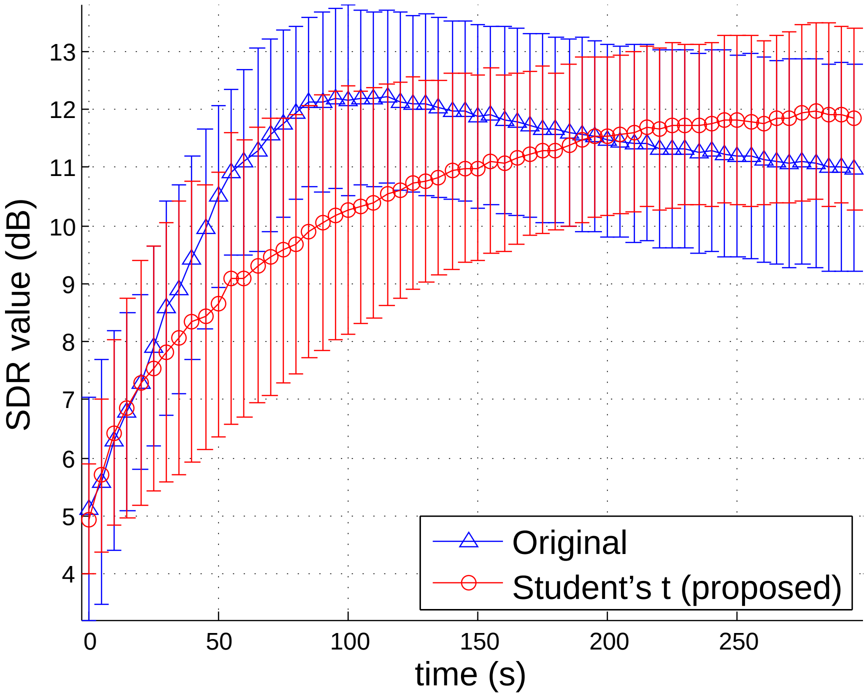

Figure 4.9. Convergence of NG-IVA as averaged SDR over 22 male-female speech mixtures with a 50% overlap between time frames,, where s_2 is at position A. The bars indicate the maximum and minimum standard deviation of the SDR.

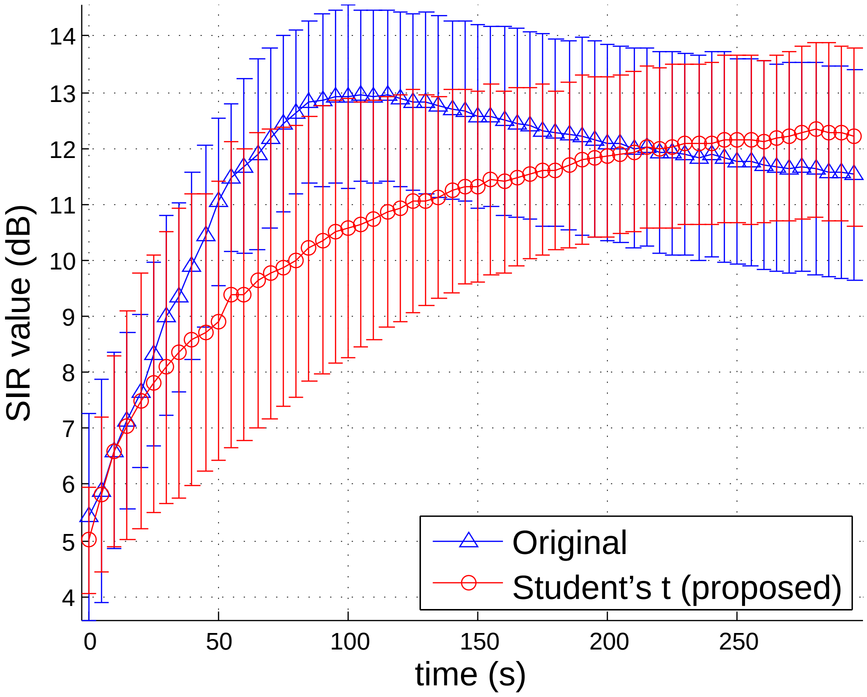

Figure 4.10. Convergence of NG-IVA as averaged SIR over 22 male-female speech mixtures with a 50% overlap between time frames, where s_2 is at position A. The bars indicate the maximum and minimum standard deviation of the SIR.

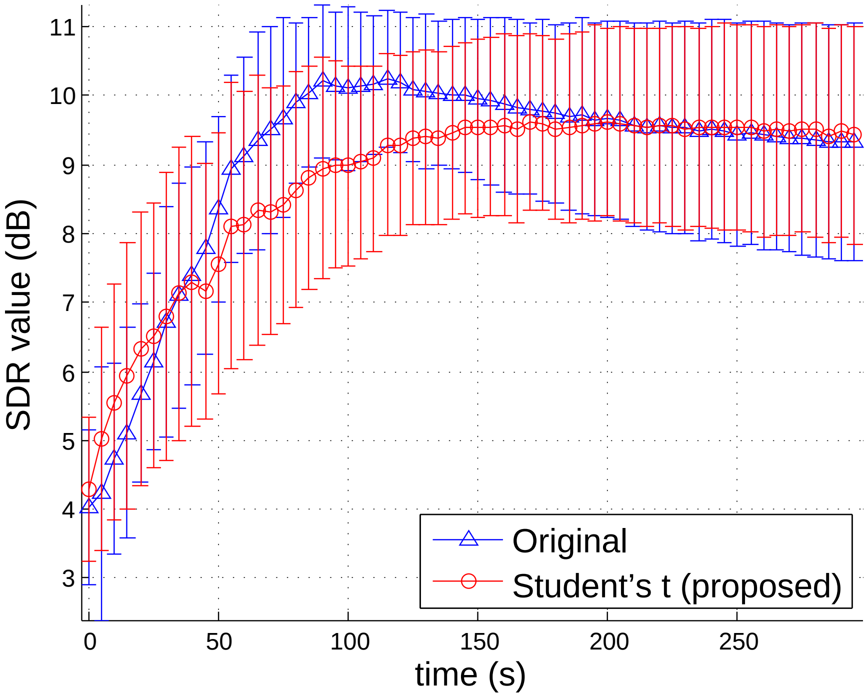

Figure 4.11. Convergence of NG-IVA as averaged SDR over 22 male-female speech mixtures with a 50% overlap between time frames, where s_2 is at position B.

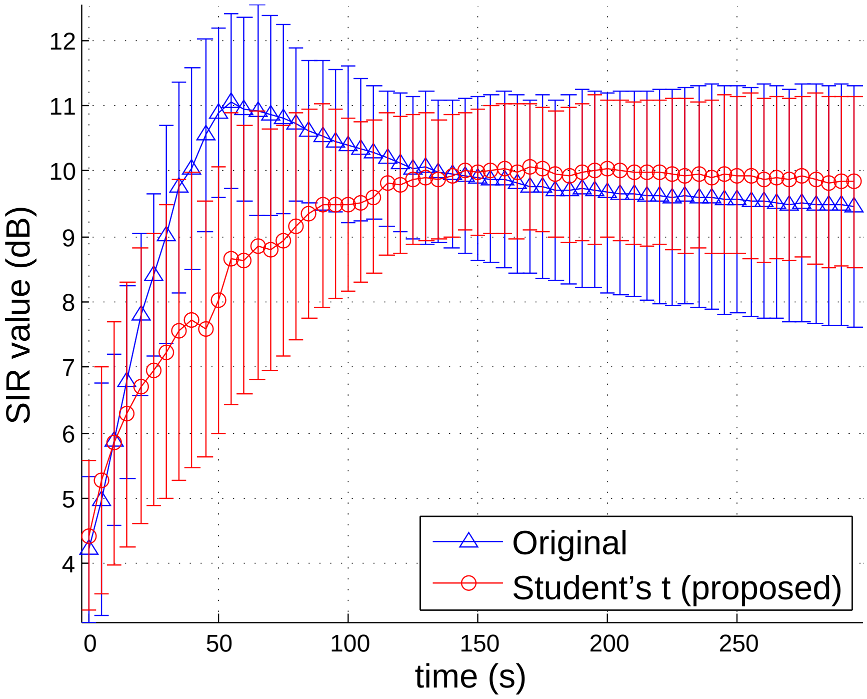

Figure 4.12. Convergence of NG-IVA as averaged SIR over 22 male-female speech mixtures with a 50% overlap between time frames, where s_2 is at position B.

not have the same overshoot characteristic as all necessary time domain signals are available. Figures 4.13 and 4.14 show results obtained from MATLAB, and do not exhibit this overshoot.

To verify that the proposed method worked in a more controlled environment, i.e. one where all necessary signals were available and there was no need to reconstruct the estimated signals with the unmixing matrices, NG-IVA with the original and Student's t were implemented in MATLAB, Figures 4.13 and 4.14. There is an improvement in performance as expected as there is no need to subsample unmixing matrices every 300 seconds. Based on the second half of the experimentation time (150s-300s) the average improvement of the converged algorithm was 0.96dB for SDR and 0.61dB for SIR. The intermediate matrices that otherwise would have been discarded cause improvement in performance as these provide more recent unmixing matrices as they are based on more recent input values. As noted previously, other studies [97] have also noted a degradation in performance in real-time. In addition, the performance graphs in [35] were based on MATLAB simulations rather than outputs from the DSP used in the same paper. In Figures 4.13 and 4.14 the learning rate was $\eta = 1.2$ for the super-Gaussian and $\eta = 1.4$ for the Student's t source priors. The degrees of freedom parameter was set to $\upsilon = 1.0$.

4.5 Summary

An online (real-time) algorithm for NG-IVA has been presented with an alternative source prior, based on a multivariate Student's t distribution, this gives an improved model for dependency amongst high amplitude data points in a frequency domain speech signal due to the

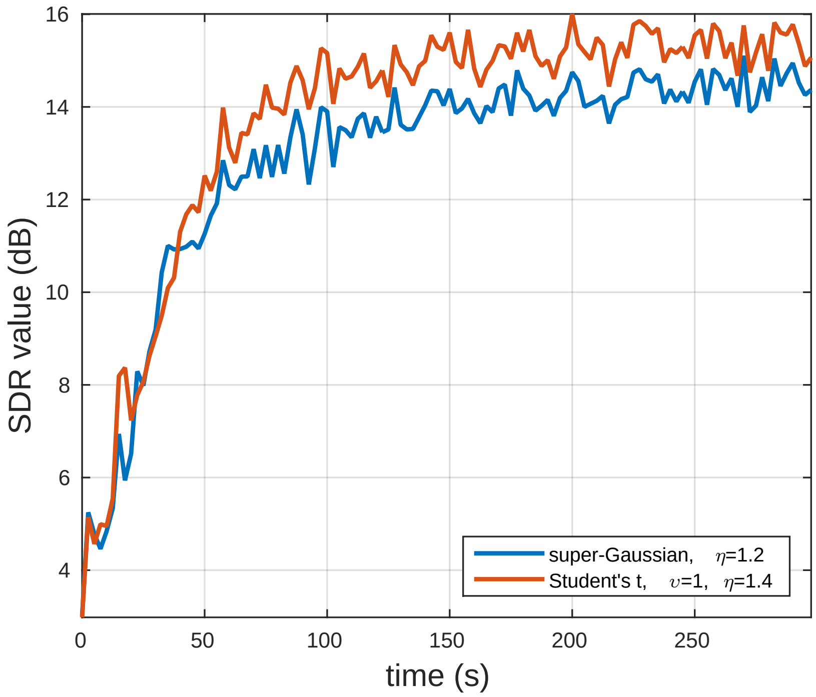

Figure 4.13. Convergence of NG-IVA as averaged SDR over 22 male-female speech mixtures with a 50% overlap between time frames, where s_2 is at position B.

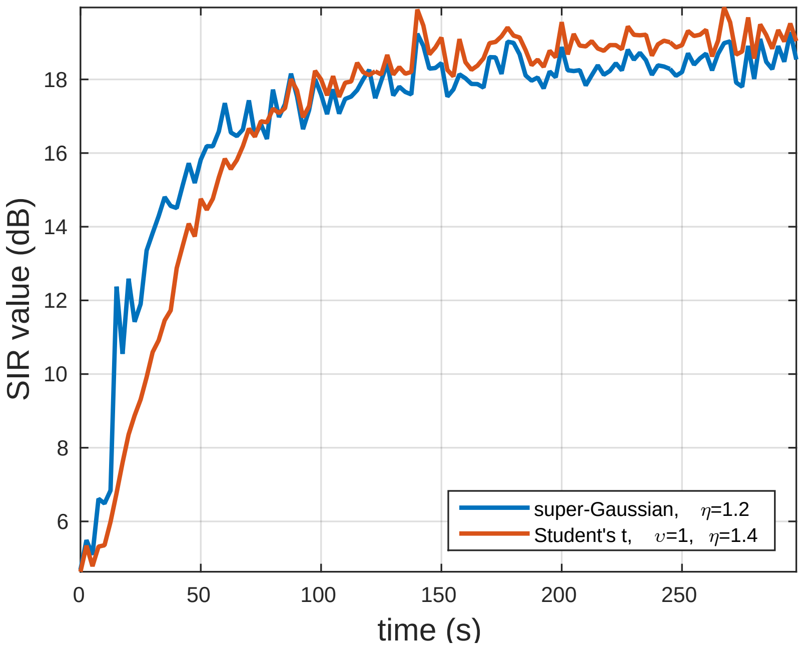

Figure 4.14. Convergence of NG-IVA as averaged SIR over 22 male-female speech mixtures with a 50% overlap between time frames, where s_2 is at position B.

heavier tails of the Student's t distribution. The reverberant mixtures used are more realistic, therefore more of a challenge to separate, than those used in some previous studies. In addition, the importance of the degrees of freedom within the Student's t distribution was highlighted.

Results have shown improved performance in terms of SDR and SIR when compared to the original NG-IVA, which has a source prior based on a multivariate super-Gaussian distribution. Real-time NG-IVA was able to be implemented as an embedded application on a TI TMS320C6713 DSP platform, a common floating-point DSP platform, due to its lower complexity when compared to the batch version. Binaural real room impulse responses were used to validate the derived method, thus in the future such a method could be applied to hearing aid technology [98], [99].

In the next chapter directional information and video cues are incorporated into online NG-IVA to address the problem of moving signal sources.

Chapter 5

A COMBINED AUDIO-VISUAL BEAMFORMING-IVA METHOD FOR SOURCE SEPARATION OF MOVING SOURCES

5.1 Introduction

Having addressed the separation performance of online NG-IVA in the previous chapter, this chapter introduces a novel way of exploiting video cues to improve the convergence speed of online NG-IVA and improve separation performance in the context of moving speech sources.

The time-varying nature of the mixing system within source separation for moving sources is the main problem to be addressed in this chapter. The proposed method is an audio-visual approach [61] which attempts to separate speech sources by employing a microphone array which is pre-steered by utilising video cues towards a target source as the sources move. An online source separation algorithm, online NG-IVA, is used as it is able to adapt to changes to the mixing environment, which is assumed to be enclosed such as a small room. As discussed previously in Chapter 4, online NG-IVA is suited to convolutive mixtures as it accounts for the permutation problem in FD-BSS.

As well as improving initial convergence of online NG-IVA, the method also addresses temporary lapses, 'dips', in separation performance due to the physical movement of sources, a known problem in real-time source separation for moving sources [97]. A related geometric source separation approach is taken in [100]. The method proposed in this chapter could be considered as an extension of geometric source separation, as it uses a combination of online NG-IVA and a pair of FIR filters which acts as a null-steering beamformer so that one source is cancelled from one of the inputs to online NG-IVA, thus giving online NG-IVA a 'head-start' when separating the mixtures. This could be thought of as a preprocessing step which in part removes the crosstalk from one of the mixtures.

Colin Cherry, who originally proposed the machine cocktail party problem, had outlined that video information could be exploited, in a similar manner to which a human speaker might use eyesight, to aid the source separation process [1], [2]. Along with [61], another method which takes an audio-visual approach to source separation is [63]. An overview of audio-visual source separation can be found in [18]. Also, previously researchers have used a null-steering beamformer to address the permutation problem in FD-BSS [101], however the proposed method takes an entirely different approach as IVA already addresses the permutation problem in FD-BSS.

The proposed method is compared to online NG-IVA, which is shown in previous studies [79] to recover from step-wise position changes of the sources. However, it will be shown that convergence performance is improved after such step-wise changes in source position and performance in terms of SDR and SIR are improved with the proposed method.

5.2 Method Overview

5.2.1 System Model

The model can be considered in two stages, firstly, a video-tracking and speaker identification stage, which is not the focus of this chapter, more details of video tracking method can be found in [32], [61], [62]. Secondly, an audio source separation stage which exploits video cues which is the main focus of this chapter. Originally proposed in [32], the schematic diagram of the overall system's framework is shown in Figure 5.1, which clearly highlights the video tracking, source separation and decision making stages.

The video localisation of speakers could be achieved with an array of more than one camera so that the location of the sources can be found by a combination of the codebook method for background subtraction [102], which would provide a 2-D outline of the speakers and the Tsai calibration (non-coplanar) technique [103], which would use the intersections of the extracted 2-D outlines to provide a 3-D location of the sources. The locations of the sources are then used in the visual-tracking step of the video tracking stage. The video tracking is achieved by a state of the art tracking algorithm such as [104]. From this, the velocity information can be used to decide which method to use, either online NG-IVA or online NG-IVA with the FIR lattice-structure (see Algorithm 4). Although the video tracking and identification is beyond the scope of this thesis the brief description above is included for completeness.

Algorithm 4 Decision making for combined beamforming/NG-IVA method for moving speech sources. See Figure 5.1.

```
Input: Location data of target source and microphone observations.
Output: Binary decision on which method to be implemented.
1: if All sources are physically stationary for a given number of time
   blocks ($D$), i.e. $\sum_{d=n-D+1}^{n}|v_{d,i}|= 0$ then
2:     Operate online NG-IVA and bypass the FIR lattice structure
       (normal NG-IVA algorithm).
3: else
4:     Combined beamforming NG-IVA method with lattice structure
       (online NG-IVA with a null-steered beamformer).
5: end if
```

The decision stage in Algorithm 4 decides between two cases, the physically stationary sources case and the moving sources case. A source is considered physically stationary if a source has an instantaneous velocity of 0 across D previous time blocks, $\sum_{d=n-D+1}^{n}|v_{d,i}|= 0$ where $v_{d,i}$ is the instantaneous velocity information at time block d for the i-th source. Practically, there is a tolerance of small movement in the sources, to ensure that the method is more robust to subtle changes in, for example, human movement.

The key contribution of the proposed method, the FIR filter-lattice structure, is outlined in the following section.

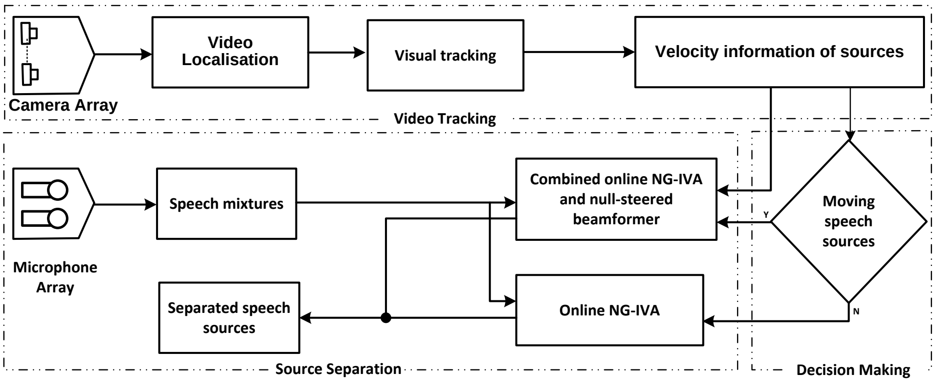

Figure 5.1. Overall system diagram including video tracking stage (which is beyond the scope of this thesis), the decision making stage and the audio source separation stage (the focus of this chapter).

5.2.2 FIR filter-lattice structure

The same FIR filter-lattice structure as used in Chapter 3 is used to cancel a second moving source, which acts as a null-steered beamformer. The main assumption with this method is that the source being cancelled (assumed to be s_1) is always at essentially a perpendicular location between two microphones, this is achieved by pre-steering a two-microphone array (possibly by employing a mechanical device) towards s_1 with information obtained from video cues. As mentioned previously, the tracking and identification is considered beyond the scope of this thesis.

The other main assumption is that $M = N = 2$ (i.e. both the number of sources and observations are equal to two), however extra sources (possibly noise sources) can be grouped together as a second source.

Maintaining the perpendicular position condition between the target speaker and the two-microphone array is achieved by orientating this two-microphone array towards a target speaker so that the two mixing filters corresponding to the target source (in this case s_1) and the two microphones are approximately equivalent. However, possible error due to misalignment or late reverberation needs to be considered, this motivates the need for two filters to correct this so that the subtraction of the two observations (thus cancelling the target source) can be performed.

A major advantage of this method is that as the IVA step is being provided with a reasonable estimate of one of the sources (in this case s_2), regardless of source position, therefore the average SDR and SIR values across both sources remain ‘stable’ for the case of physi-

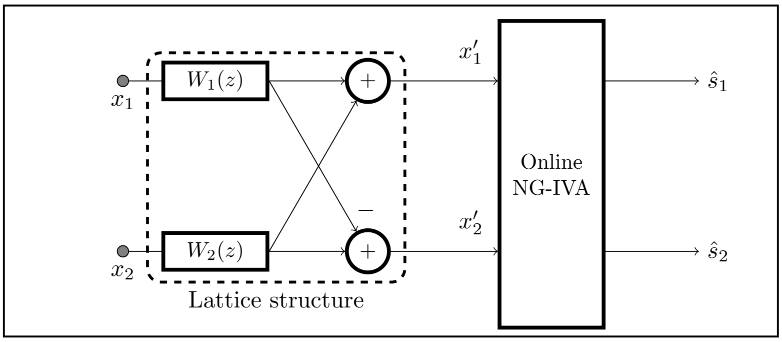

Figure 5.2. System diagram for combined online NG-IVA null-steering beamforming method, the pair of FIR filters are shown here in z domain notation; discrete time t is dropped on all signals for convenience. The lattice structure can be bypassed to give the online NG-IVA algorithm, when the sources are stationary.

cally moving sources, i.e. the effects of performance lapses, due to the movement of sources, are reduced. Another advantage is that convergence using the proposed method is expected to be faster than normal NG-IVA after a source has moved in a step-wise fashion.

Assuming that the acoustic environment does not change during experimentation, $\mathbf{w}_{\{1,2\}}$ will also not change. To allow for movement of s_1 the microphone array will orientate itself accordingly using information from the video system, as mentioned previously. In addition, the lattice structure is not subject to the position of source (s_2), which has freedom to move within the acoustical environment as online NG-IVA will allow for any changes in the related impulse responses.

The two inputs to the online NG-IVA algorithm are the filtered version of one observation, minus the other filtered observation and the addition of the two filtered observations, thus:

$$x'_1(t) = \sum_{\tau=0}^{L-1} x_1(t-\tau)w_1(\tau) + \sum_{\tau=0}^{L-1} x_2(t-\tau)w_2(\tau) \tag{5.2.1a}$$

$$x'_2(t) = \sum_{\tau=0}^{L-1} x_2(t-\tau)w_2(\tau) - \sum_{\tau=0}^{L-1} x_1(t-\tau)w_1(\tau), \tag{5.2.1b}$$

where $(\cdot)'$ denotes the altered version of $x_{\{1,2\}}$, t is the discrete time index, τ is a discrete time delay and L is the length of the time domain FIR filter. These are written in the time domain here so that they are consistent with the notation in Chapter 3, however the pair of FIR filters could be implemented in the frequency domain if desired, Figure 5.2 shows these equations in diagrammatic form.

5.2.3 Method of Principal Angles

The training process to find the pair of time domain FIR filters $\hat{\mathbf{w}}_{\{1,2\}}$, as described in full in Chapter 3 and included here for completeness, is as follows; a pair of convolution matrices are formed, denoted X_1 and X_2, performing a QR decomposition yields;

$$X_1 = Q_1 R_1 \tag{5.2.2a}$$

$$X_2 = Q_2 R_2. \tag{5.2.2b}$$

An error vector is written, as in Chapter 3, as;

$$\epsilon_1(w_1, w_2) = \epsilon_2(\tilde{w}_1, \tilde{w}_2) = \boldsymbol{\epsilon}_2 = (Q_1\tilde{\mathbf{w}}_1 - Q_2\tilde{\mathbf{w}}_2), \tag{5.2.3}$$

where $\epsilon_1(w_1, w_2)$ and $\epsilon_2(\tilde{w}_1, \tilde{w}_2)$ are error functions. Also, $\tilde{\mathbf{w}}_1 = R_1\mathbf{w}_1$ and $\tilde{\mathbf{w}}_2 = R_2\mathbf{w}_2$. The minimisers of a new cost function are then found as:

$$J_2 = ||\boldsymbol{\epsilon}_2||_2^2, \ \{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2\} = \arg\min_{\tilde{\mathbf{w}}_1 . \tilde{\mathbf{w}}_2} J_2, \tag{5.2.4}$$

subject to $||\tilde{\mathbf{w}}_1||_2 = ||\tilde{\mathbf{w}}_2||_2 = 1$. The two constraints are applied simultaneously due to the orthonormal basis. To find the principal angles and principal vectors of the orthonormal subspaces Q_1 and Q_2, the singular value decomposition is taken of $Q_1^T Q_2$, so that $[U, \Lambda, V^T] = SVD(Q_1^T Q_2)$. The constraints $||\tilde{\mathbf{w}}_1||_2 = 1$ and $||\tilde{\mathbf{w}}_2||_2 = 1$ are inherently introduced to the method by exploiting the properties of the SVD avoiding the trivial solution $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2 = \mathbf{0}$. The cost function J_2 is rewritten as:

$$J_2 = ||Q_1\tilde{\mathbf{w}}_1 - Q_2\tilde{\mathbf{w}}_2||_2^2 \tag{5.2.5}$$

$$= \tilde{\mathbf{w}}_1^T\tilde{\mathbf{w}}_1 + \tilde{\mathbf{w}}_2^T\tilde{\mathbf{w}}_2 - 2\tilde{\mathbf{w}}_1^T Q_1^T Q_2 \tilde{\mathbf{w}}_2, \tag{5.2.6}$$

therefore reducing J_2 is equivalent to maximising $\tilde{\mathbf{w}}_1^T Q_1^T Q_2 \tilde{\mathbf{w}}_2$ as $\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1 = 1$ and $\tilde{\mathbf{w}}_2^T \tilde{\mathbf{w}}_2 = 1$, thus:

$$\arg\min_{\tilde{\mathbf{w}}_1 . \tilde{\mathbf{w}}_2} J_2 \equiv \arg\max_{\tilde{\mathbf{w}}_1 . \tilde{\mathbf{w}}_2} \tilde{\mathbf{w}}_1^T Q_1^T Q_2 \tilde{\mathbf{w}}_2. \tag{5.2.7}$$

By exploiting the SVD:

$$(\tilde{\mathbf{w}}_1^T Q_1^T Q_2 \tilde{\mathbf{w}}_2) = \tilde{\mathbf{w}}_1^T (U\Lambda V^T)\tilde{\mathbf{w}}_2 = \tilde{\mathbf{w}}_1^T (\sum_m \lambda_m \mathbf{u}_m \mathbf{v}_m)\tilde{\mathbf{w}}_2, \tag{5.2.8}$$

by selecting $\tilde{\mathbf{w}}_1 = \mathbf{u}_1$ and $\tilde{\mathbf{w}}_2 = \mathbf{v}_1$ in Equation (5.2.8), where $\mathbf{u}_1$ and $\mathbf{v}_1$ are the vectors from the rows of U and V which correspond to the largest largest singular value, denoted λ_1, where the subscript $(\cdot)_1$ denotes the largest singular value. In turn, λ_1 corresponds to the smallest angle between the orthonormal bases Q_1 and Q_2 [74].

The equalising filters are the columns of U and V which correspond to λ_1 (as they maximise Equation (5.2.8)), multiplied by the inverse of R_1 and R_2 to allow for the basis change by the QR decomposition, hence:

$$\hat{\mathbf{w}}_1 = R_1^{-1}\mathbf{v}_1 \tag{5.2.9a}$$

$$\hat{\mathbf{w}}_2 = R_2^{-1}\mathbf{u}_1, \tag{5.2.9b}$$

thus the pair of equalisation filters ($\hat{\mathbf{w}}_{\{1,2\}}$) is found; Equations (5.2.1a) and (5.2.1b) use this pair of equalising FIR filters to implement the lattice structure. It is assumed the training stage is done ‘offline’ and any time taken to train $w_{\{1,2\}}$ is not included in the experimental results below.

5.2.4 Online Natural Gradient Independent Vector Analysis

Online NG-IVA is used as in the previous chapter with a score function derived from a super-Gaussian source prior (defined as: $q(\mathbf{s}_i) \propto exp\Big(-((\mathbf{s}_i - \mu_i)^H \mathbf{\Sigma}_i^{-1}(\mathbf{s}_i - \mu_i))^{\frac{1}{2}}\Big)$, in Chapter 4), included for completeness, thus the update equations are:

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{N} g_{ij}[n]^{(k)} x_j^{(k)}[n] \tag{5.2.10a}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \tag{5.2.10b}$$

$$\Delta g_{ij}^{(k)} = \sum_{l=1}^{N} (\Lambda_{il} - \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \hat{s}_l^{(k)*}) g_{lj}^{(k)} \tag{5.2.10c}$$

$$\xi^{(k)}[n] = \beta \xi^{(k)}[n-1] + (1-\beta) \sum_{i=0}^{M} |x_i^{(k)}[n]|^2 / M \tag{5.2.10d}$$

$$g_{ij}^{(k)}[n+1] = g_{ij}^{(k)}[n] + \eta \sqrt{(\xi^{(k)}[n] + \gamma)^{-1}} \Delta g_{ij}^{(k)}, \tag{5.2.10e}$$

as derived in Chapter 4.

5.3 Methodology

Speech signals of approximately 250 seconds long were used to test the proposed method, where the second interference source is moved in a step-wise fashion at a third of the overall experimental time (approximately 84 seconds) to simulate a moving source. Table 5.1 details the experimental conditions in full for the experiments.

The pair of FIR cancelation filters was found in a training stage where s_2 was silent. Three seconds of training data from the TIMIT database were used to form the convolution matrices $X_{\{1,2\}}$ and thus train the pair of FIR filters $\hat{\mathbf{w}}_{\{1,2\}}$.

A 2D room plan is illustrated in Figure 5.3. Source s_1 was kept at the same position perpendicular to both microphones throughout experimentation (as a straightforward method to simulate the pre-steering of the microphone array due to video cues), source s_2 begins at 75° for

Sampling rate	8kHz
FFT length (K)	2048
Window function	$win = \sin\left(\frac{\pi}{2}\sin^2\left[\frac{\pi}{2K}\left(k+\frac{1}{2}\right)\right]\right)$
Overlap ratio	50%
Length of FIR equalisation filters (L)	900 taps
RT_{60}	565ms
Overall mixture length	251s
Source movement time	84s

Table 5.1. Experimental conditions for combined NG-IVA-Beamforming results, where $k \in \{1, \ldots, K\}$.

all experiments then moves in a step-wise fashion to either 45°, 30°, or 15°, except for the first experiment where its position remains constant at 75°.

Results were averaged for 56 different male-female mixtures from a range of accents (where the speakers were swapped between positions), the clean speech signals were taken from the TIMIT dataset and the mixing filters were taken from a binaural IR database [54]. All available speech sources for each speaker were concatenated to form longer speech sources. Full experimental parameters can be found in Table 5.1. SDR and SIR values were calculated using the BSS Evaluation Toolbox [57] over a window period of three seconds.

Steady state values in the Experimental results section are calculated by either taking the mean SDR and SIR values across the last 120 seconds of experimental time for the stationary experiment or 60 seconds of experimental time for the moving experiments. This allows sufficient time for experiments to converge.

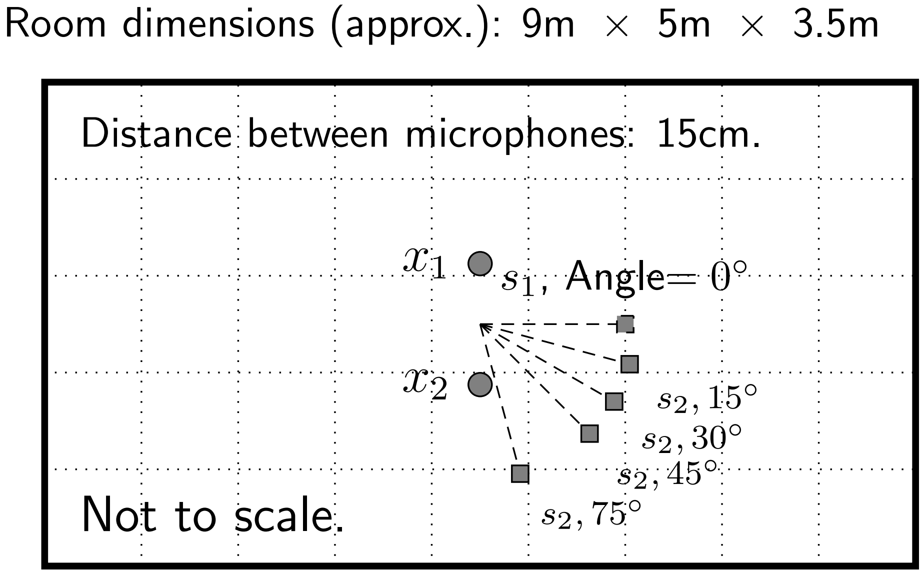

Figure 5.3. 2D room plan of microphone and source positions.

5.4 Experimental Results

In Figures 5.4 and 5.5 the second source (s_2) remained physically stationary throughout experimentation. The improvement in initial convergence time for proposed beamforming-IVA method compared to the original method is outlined in Table 5.2. The convergence time to reach 50% of the steady state is more than halved with the proposed beamforming-IVA method. To attain 75% of the steady state, the convergence time is reduced by 12.07 seconds for SDR and 7.45 seconds for SIR. In Table 5.2 the steady state time average was based on the mean value of the SDR and SIR for the final 120 seconds, the convergence times were based on the amount of time both methods took to reach 50% and 75% of their respective final steady states. This improvement in initial convergence time is observed across all experiments including those where source s_2 moves to a new location, furthermore initial convergence times are considered acceptable for the room impulse responses ($RT_{60} = 565$ms) used.

Steady state performance (based on the final 120 seconds of experiment time) is 16.31dB and 20.33dB for the original online NG-IVA method (SDR and SIR respectively). For the proposed combined online NG-IVA and null-steering beamforming method the results were 13.92dB and 21.31dB for the SDR and SIR respectively. In this case there has been a degradation in SDR performance. This emphasises a need to have intelligent selection of the source separation method (i.e. with or without the proposed beamforming method), so that the performance parameters of the original method can be attained.

Although initial convergence of the proposed method is improved, Figures 5.4 and 5.5 demonstrate little support for the proposed method

	SDR				SIR			
	Time (s)		Steady state (dB)		Time (s)		Steady state (dB)	
(%)	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.
50	8.51	22.90	13.92	16.31	10.57	23.75	21.31	20.33
75	22.00	34.07			29.28	36.73		

Table 5.2. Convergence time of stationary case (75° → 75°), where the final steady state value used has been calculated by averaging over the final 120 seconds of experiment time. (Prop. = proposed combined null-steered beamformer with NG-IVA, Orig. = online NG-IVA).

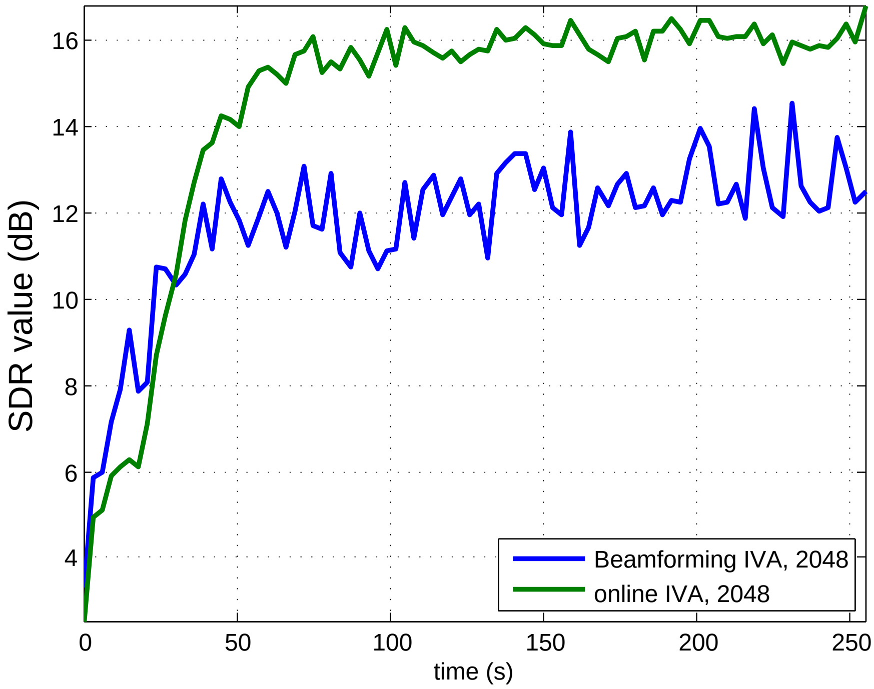

Figure 5.4. An SDR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where the second source is physically stationary.

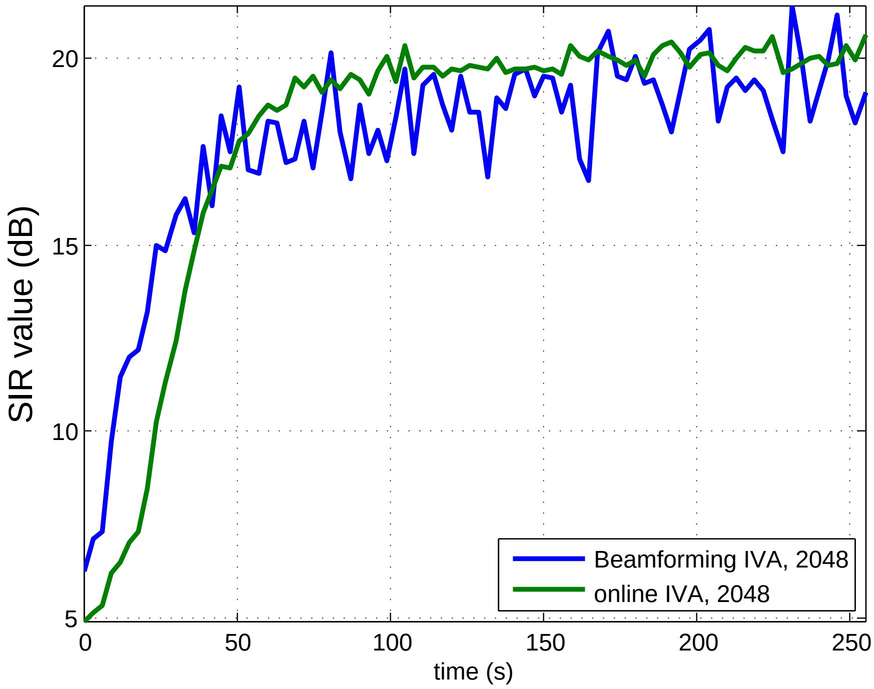

Figure 5.5. An SIR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where the second source is physically stationary.

as the original NG-IVA gives better performance in the steady state in terms of SDR, only the SIR is slightly improved (Table 5.2). The advantage of the proposed method is more pronounced when the position of s_2 is changed in a step-wise fashion, Figures 5.6 - 5.11. In a realistic scenario where the proposed combined method is implemented it is suggested that the lattice structure in Figure 5.2 is bypassed when the stationary scenario is detected by the video system, thus the method is able to achieve better steady state performance shown in Figure 5.4 and Figure 5.5.

Figures 5.6 and 5.7 show results of the step-wise movement of s_2 from an angle of 75° to an angle of 45°. The performance is slightly worse in terms of SDR when compared to the original method, see Figure 5.6. Though there is an obvious improvement in convergence time and final value after the step-wise movement of s_2, in terms of SIR in Figure 5.7. The details of gain in steady state performance in the final 60 seconds of experiment time for all the moving source and stationary cases is summarised in Table 5.3. Steady state time is based on 60 seconds for the moving cases (rather than 120 seconds for the stationary case) because it is harder to ensure that the methods would have recovered and be in steady state over the last 120 seconds of experiment time. This shows that in terms of SDR there is an improvement for source s_2 moving to 30° and 15° and in terms of SIR there is a movement for all moving cases. The stationary case is also provided as a comparison and shows no improvement in the performance for the proposed method. The best performance is observed when s_2 moves from 75° to 15°, which shows the potential of the algorithm to handle fast moving sources.

s_2 (start position → end position)	Gain SDR (dB)	Gain SIR (dB)
75° → 75°	-4.33	-2.74
75° → 45°	-0.71	2.12
75° → 30°	0.94	3.20
75° → 15°	7.16	10.75

Table 5.3. Average improvement in performance for the proposed (compared to original NG-IVA) over the last 60 seconds of overall experiment time.

Convergence time after the lapses in performance due to the stepwise movement of s_2 are summarised in Table 5.4. Convergence times in this table are based on the time it takes to reach 80% of the respective final steady state performance (based on a mean calculated over the final 60 seconds of overall experiment time.) Table 5.4 shows that convergence time is improved for all the configurations tested and in some cases convergence time is halved. Final steady state values are also improved or kept constant between the proposed and original algorithms. SIR performance values are consistent for all final positions of s_2 in the proposed method, however the steady state SIR value declines as the angle between s_2 and s_1 is reduced in the original method (online NG-IVA), which supports the proposed method.

Figures 5.8 and 5.9 are the first experiment (75° to 30°) to suggest strong performance advantage of the proposed research for both performance parameters. Convergence is quicker in both SDR and SIR after the step-wise movement and in the steady state performance of the proposed method performs better for both parameters (0.94dB for SDR and 3.20dB for SIR).

Figures 5.10 and 5.11 show the method in the most difficult case tested (75° to 15°), where s_2 is closest to s_1 in its second position, this case demonstrates the advantages of the method and there is a clear

	SDR				SIR			
	Time		Steady state (dB)		Time		Steady state (dB)	
s_2 pos.	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.	Prop.	Orig.
75° → 45°	16.24	17.99	14.19	14.75	12.53	16.24	21.05	18.47
75° → 30°	11.82	29.57	13.80	13.06	10.39	24.90	20.35	16.51
75° → 15°	22.25	30.72	15.32	7.86	15.68	48.02	21.85	10.71

Table 5.4. Convergence time to reach 80% of the respective steady state values after the step-wise movement of s_2, where the final steady state value used has been calculated by averaging over the final 60 seconds of experiment time. (Prop. = proposed combined null-steered beamformer with NG-IVA, Orig. = online NG-IVA).

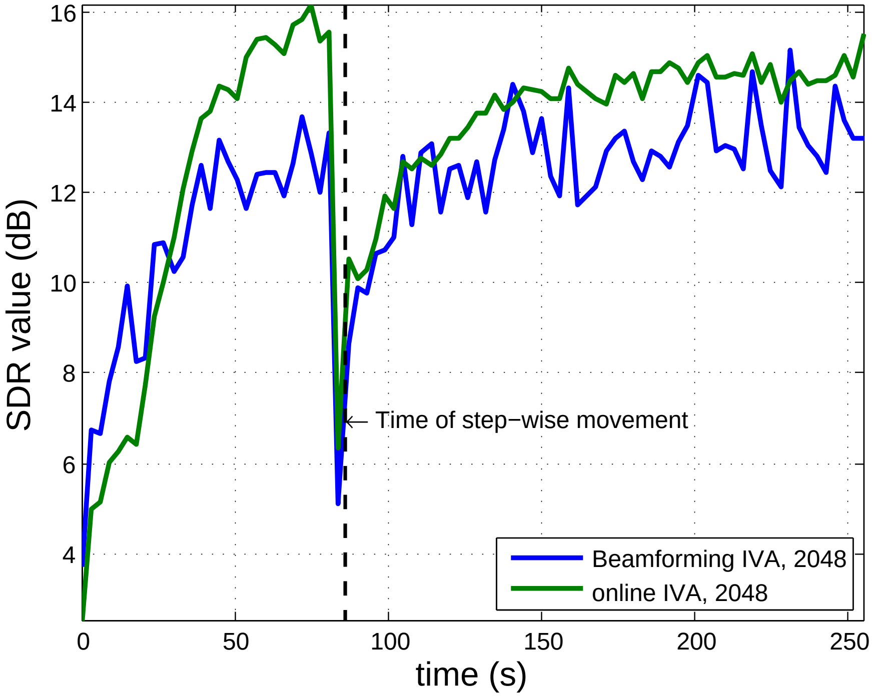

Figure 5.6. An SDR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where s_2 is moving from an angle of 75° to 45°.

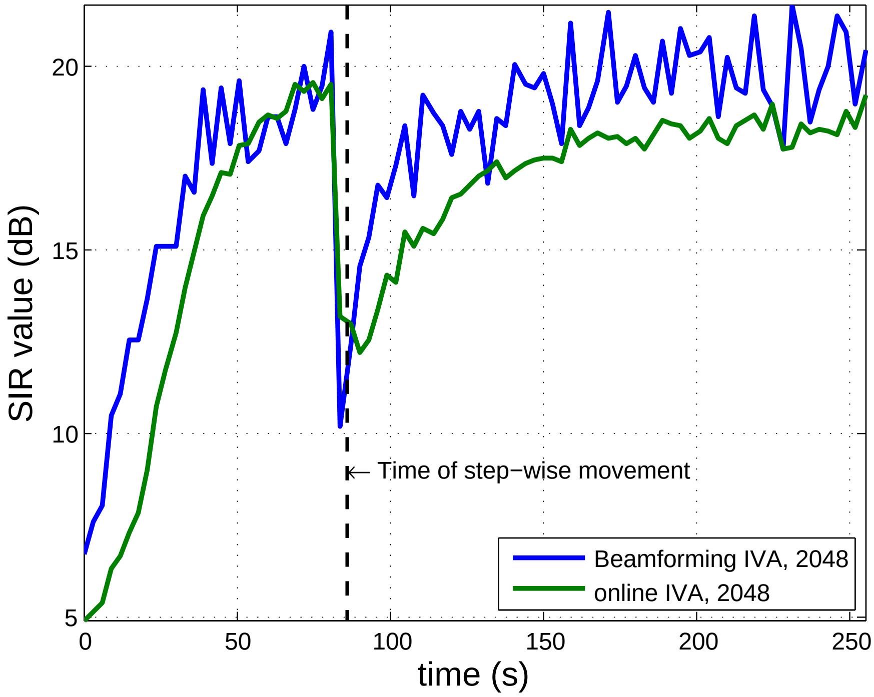

Figure 5.7. An SIR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where s_2 is moving from an angle of 75° to 45°.

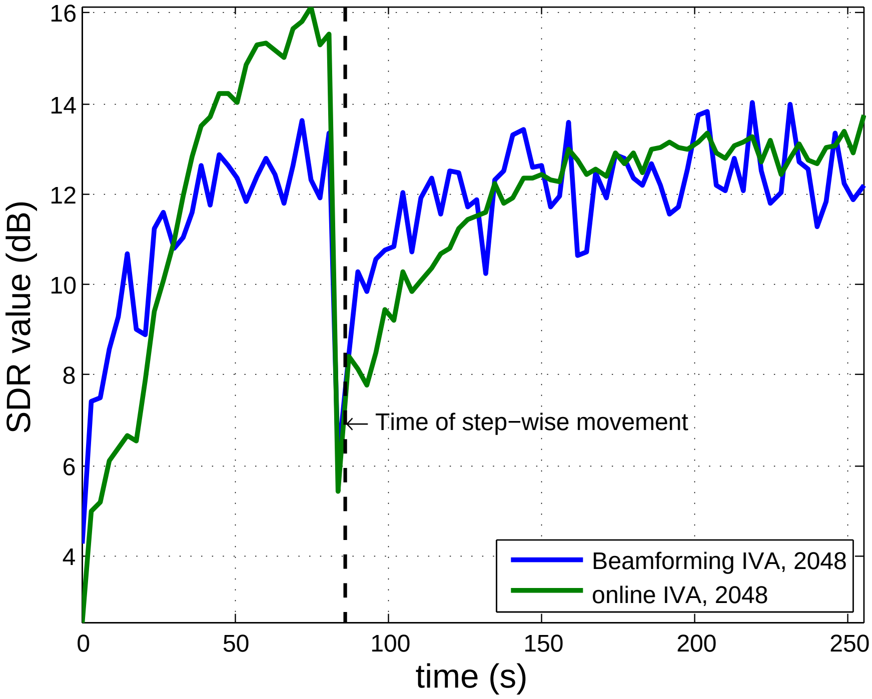

Figure 5.8. An SDR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where s_2 is moving from an angle of 75° to 30°.

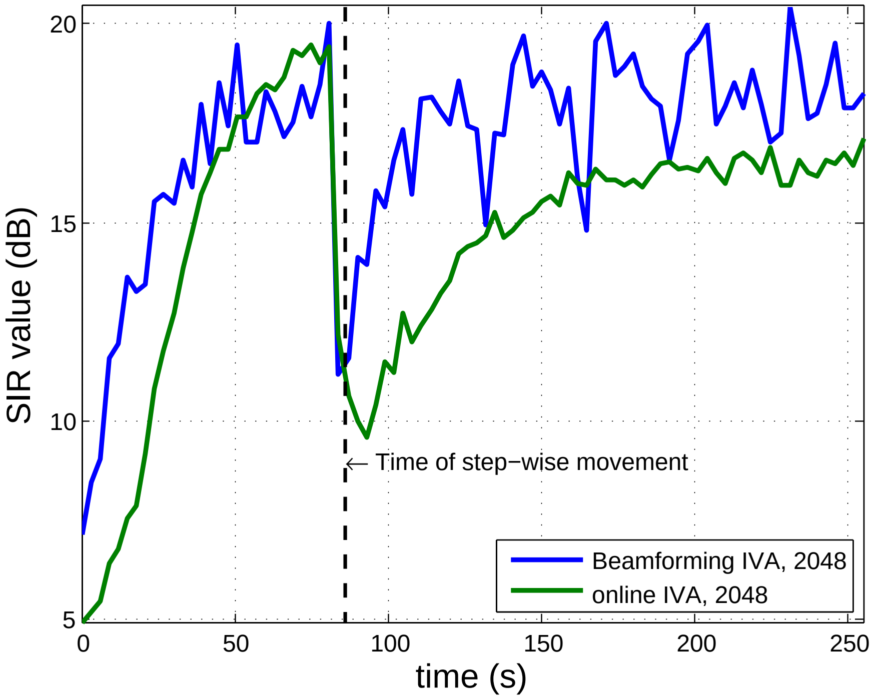

Figure 5.9. An SIR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where s_2 is moving from an angle of 75° to 30°.

improvement of SDR (7.16dB) and SIR (10.75dB) in the steady state (Table 5.3).

A potential disadvantage of the proposed method is that SDR and SIR performance in the steady state is slightly more erratic than the original method, however any disadvantage there is, would be mitigated by improved convergence and improved mean steady state SDR and SIR performance in the last 120 and 60 seconds (Table 5.2 and Table 5.3), and using the proposed framework in Algorithm 4.

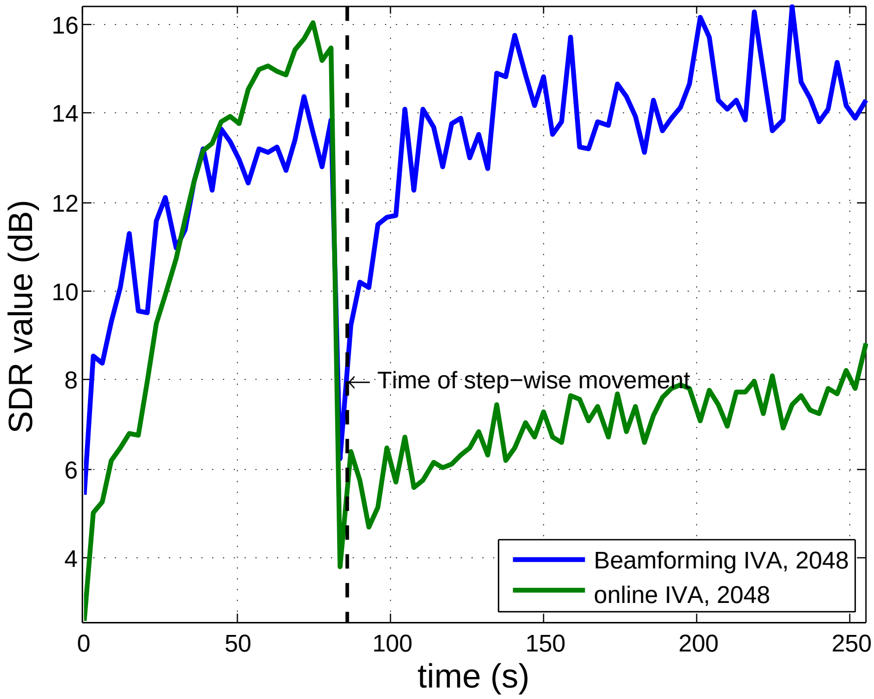

Figure 5.10. An SDR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where the s_2 is moving from an angle of 75° to 15°.

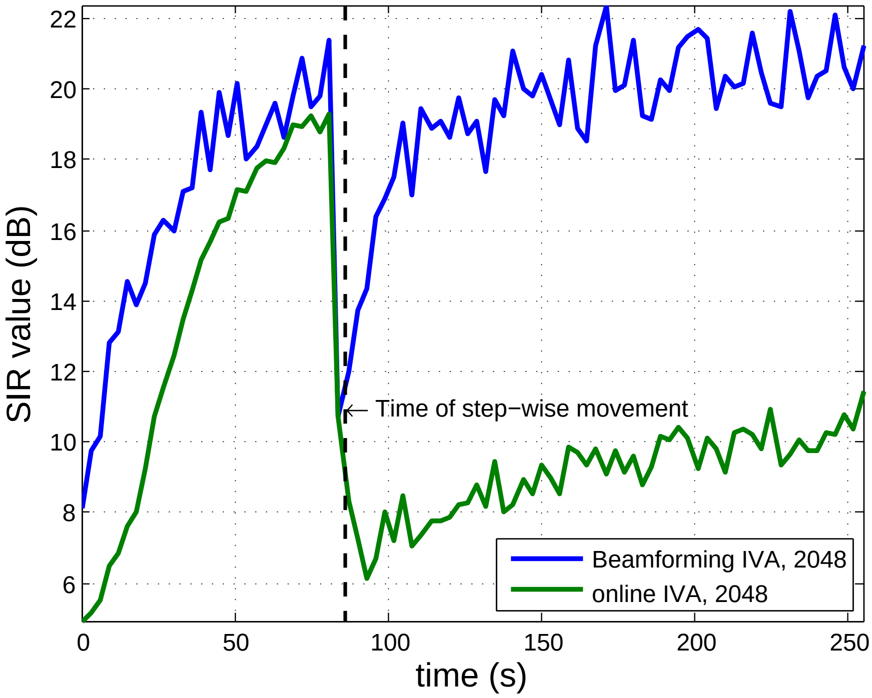

Figure 5.11. An SIR comparison of the combined IVA-beamforming and original online IVA ($\eta = 0.55$), where the s_2 is moving from an angle of 75° to 15°.

5.5 Summary

A method which behaves as a preprocessing step to remove a target source by acting as a null-steering beamformer, along with an online NG-IVA algorithm, has been presented within an established audio-visual source separation framework then applied to the case of step-wise moving sources. Results have been shown for a physically stationary case then several cases of step-wise movement of the second source (s_2) with increasing difficultly (i.e. when the source is in its second position it is nearer s_1 than in the previous test). Results confirm that there is faster convergence time after a second source has moved in a step-wise fashion and initial convergence time has also been improved. The major contribution of this chapter is improving SDR and SIR performance of the proposed method which exceeds the original online NG-IVA algorithm in cases involving fast-moving (simulated by a step-wise movement) non-stationary sources.

Furthermore, the algorithm has the potential to also deal with the over determined case, as one source would be removed from at least two of the observations in advance and has the potential in future to eliminate spatially separated noise sources.

In the next chapter, the contributions of the thesis are summarised, avenues for future research are discussed and closing remarks are provided.

Chapter 6

CONCLUSION

6.1 Overall conclusions

The work presented in this thesis has shown that potential solutions for the cocktail party problem can be implemented in an online manner for application within reverberant environments. In the spirit of the suggestion in Colin Cherry's original paper, video cues, in this case known speaker location, can also be used as a priori knowledge and are able to be exploited to improve the source separation process.

The main issue facing blind source separation systems is reverberation, which is confounded by moving speech sources, such as human speakers moving in a room. Consequently, mixing filters are time varying and therefore adaptive algorithms with an online structure are an obvious choice, as such algorithms are likely to have an in-built ability to adapt to changes in the acoustic environment.

Realistic reverberant environments require that processing is conducted in the frequency domain, which leads to the permutation problem that is inherent in frequency domain ICA. Chapter 2 highlights previous work within the subject of convolutive blind source separation and focuses on the permutation problem. In addition, Chapter 2 provides a comparison between the batch versions of ICA and IVA, and as expected IVA performs betters than ICA as it directly addresses the permutation problem. Also included is information on the measurement parameters and datasets used within the thesis.

The contributions of this thesis satisfy the five research objectives outlined in the introduction. The objectives were addressed by introducing methods which exploit known speaker location or are more suited to speech signals. In this regard, Chapter 3 introduced a novel method to calculate a pair of FIR filters that cancelled a target speaker within a room exploiting known speaker location and the principle angles method (which in turn exploits the SVD). This was formulated in the time domain thereby mitigating any potential approximation effects in circular convolution. An adaptive filtering stage was then applied to recover the remaining speakers. This proof of concept method demonstrates that it is possible to recover sources using only second order statistics, yielding good results above 20dB in experiments. However, in the method described this comes at the cost of a training stage where the estimates of the pair of cancellation filters are found by means of an SVD.

Chapter 4 introduced online natural gradient IVA and a new source prior, namely a multivariate Student's t which was the main contribution of this chapter. The proposed multivariate source prior pdf is more suited to certain speech signals due to its heavier tails and therefore it better represents the content of a frequency domain speech signal. The online NG-IVA algorithm was implemented in real-time on a Texas Instruments digital signal processor platform. Due to problems with subsampling the unmixing matrices and reconstructing the various signals remotely, behaviour of the real-time implementation is unusual.

Despite this, the Student's t pdf was shown to perform better than the originally proposed source prior and performance was confirmed in a MATLAB simulation.

Following the introduction of online NG-IVA, a method which improves the convergence of online NG-IVA was described in Chapter 5, in the case of moving sources. The pair of cancellation filters and online NG-IVA were combined to produce a solution which relied on known speaker locations. Significant improvement to the convergence of online NG-IVA was achieved. In some cases, such as the case where s_2 moves from 75° to 30°, convergence time improves by almost 18 seconds for SDR, demonstrating a clear contribution in the case of audio-visual blind source separation for moving sources. However, improved convergence time was traded for reduced steady state performance in some situations. In such cases it is proposed that the lattice structure that the filters are arranged in is turned off, depending whether speakers are stationary or moving, reducing the method to online NG-IVA.

All methods considered in the main body of the thesis deal with reverberant environments. In this respect the impact of this thesis is a stepping stone for future researchers to expand on the ideas and find an all encompassing solution to the cocktail party problem suggested by Colin Cherry.

6.2 Future work

For the PA method for target speaker cancellation proposed in Chapter 3, possible changes include removing the training phase and replacing it with a voice activity detector (VAD) [66], which detects silent periods in speech sources. Such a solution would then retrain the pair of cancellation filters during silent periods. This would lead to implementing the pair of cancellation filters estimated from the principal angles method in real-time.

Within the context of Chapter 4, investigating other appropriate multivariate source prior distributions that would improve the separation of speech for moving source further would be interesting. The formulation for IVA assumes that there are evenly weighted dependencies between the frequency bins, even for frequency bins which are far apart. In the future, frequency bins could be split into frequency bands. By grouping the frequency bins this would give less weighting to frequency bins further apart and more weighting to frequency bins closer together. The band arrangement scheme could be determined to suit speech signals, as well as using different source priors within the bands, such as those discussed in [105]. Another area of future work is to implement online NG-IVA on a field programmable gate array (FPGA) in real-time, potentially using the architecture to execute parts of the online algorithm in parallel, thus increasing the speed of computation time and saving on computational load. Any future work would extract time domain signals from the TI DSP in real-time, so that practical limitations of the current set up can be avoided.

Future work for the combined method in Chapter 5 would include using the IVA-beamforming technique proposed in the Chapter with the Student's source prior. Also, a more robust study into the effect of the learning rate within NG-IVA with the combined method could be carried out. A possible avenue of future study is combining the method with the time-varying learning rate as introduced in [106]. Additionally, the method could be expanded to work on more than one acoustic plane.

In this thesis, the number of frequency bins is assumed to be similar to that of the length of the room impulse response (2048), unless otherwise stated. This figure is chosen so that the IVA algorithm can maintain good SDR and SIR performance values at the outputs, whilst being a realistic number of frequency bins to cover the length of time domain impulse responses. It is possible in future work to investigate reducing the number of frequency bins that IVA operates over (to reduce computational complexity) and still provide acceptable SDR and SIR separation performance values. In addition, experiments in Chapter 5 improved convergence time, however, there was a trade-off for reduced steady state performance, but this could also be the subject of future work.

One goal within the community of researchers investigating the cocktail party problem is to find a more ‘elegant’ solution, potentially mimicking deep rooted biological and sensory mechanisms that a human may use. For example, finding a solution to the underdetermined case is one of these areas, separating mixture in the underdetermined case using videos cues, in a similar way to which a human being only has two ears, yet has the ability to separate more than two speakers. This research takes a step towards providing potential solutions for the moving source case. Humans also use other pieces of a priori knowledge about the target speaker, such as familiarity with the speaker’s voice, such as expected timbre or accent, to assist in understanding the speaker. Within this context, smart initialisation in form of initialising unmixing matrices could also be considered in future research.

A possible weakness in the study is that video information is not explicitly utilised and information such as location of the speakers and velocity of moving sources are assumed. Therefore, unforeseen problems which may arise by using video information have not been investigated. Although there have been studies where the video tracking has been investigated, the author is unaware of a full system (audio and video) being implemented in real-time. This can also be considered for future study.

6.3 Final remarks

With advances in computing power and embedded technology becoming increasingly ubiquitous, the demand for voice automated technology and a solution to the cocktail party problem will undoubtedly grow. The prevalence of a variety of sensors on such embedded technology, not just video cameras, has the potential for exciting new developments in signal processing for natural language processing. A large part of this is blind source separation techniques which will play an important role in the years ahead.

Appendix A

APPENDIX

In some applications, particularly with non-stationary sources in real time, it is desirable to use a more efficient method as higher order methods can consume large amounts of system resources and require too much time to produce accurate estimates.

In this section a different approach is described to audio-visual source separation. It relies on a given set of previously calculated frequency responses (FR) and an unwanted noise source's 3D location from video information provided by an array of video cameras. The proposed method can be viewed as a pre-processing stage before a more conventional BSS algorithm that suppresses a noise source with a known location; in addition it can be used by itself in the 2-microphone 2-source scenario to extract a filtered version of one of the sources.

Throughout this section the method is considered by itself in the general case and there is an equal number of microphones and sources ($M = N$), in line with the basic ICA model. When $N = 2$, the method can be used by itself, although further processing could be used if it was desired to extract the cancelled noise from the mixture. However, if the number of microphones and sources was increased, for example when $M = N = 3$, after noise source suppression two sources would be left in the mixture and a further BSS algorithm could be used such as

ICA or IVA. It is assumed that any movement of the sources is slow, so that a quasi-static assumption can be made, and that the method uses block-wise processing, as a result changes in the room impulse response due to movements of the sources, or other persons or objects, that affect the acoustic environment are considered to be negligible between time blocks. This is a proof of concept method and would not work well in a real environment, due to the artificial nature of image method IRs.

The time-frequency convolutive mixture model in Equation (2.1.18) is used in this model.

The method is divided into two principal stages. The first stage consists of estimation of FRs between the noise source and multiple microphones (Section A.0.1). In a second stage these FRs are used to find a suppression filter to remove the effect of the noise source on the mixture at each microphone (Section A.0.2).

The transforms of known IRs which have been measured over a spatial grid are calculated. From these FRs a weighted linear combination is calculated to estimate an FR at the point where the noise source is measured. It is only necessary to know the FRs around the noise source to remove it from the mixture. Note that the number of microphones always needs to be equal to the number of sources including the noise source, as the related transfer functions are used to create a suppression filter.

A.0.1 Frequency response estimation

The room is divided into cubes known as voxels which are arranged into a non-overlapping 3D spatial grid pattern. Based on work in [107] and [108], part of the motivation for calculating FRs in such a manner is to correct inaccuracies in measured IRs. An FR is estimated by calculating a weighted average of previously known FRs at each corner of the voxel that contains the noise source, calculated by the Fourier transform of an IR [53]. The weighted average depends on the noise source position (p_n), which is provided by video information, as represented in Figure A.1. The purpose of taking averages of FRs is to avoid storing an IR and a transfer function (TF) for every possible point within the room which would be impractical. Weights are assigned to each FR at each corner ($a = 1, \ldots, 8$) and are calculated by:

$$\omega_a = \Big(1 - \frac{x_a^{(\text{voxel})}}{\Omega_L}\Big)\Big(1 - \frac{y_a^{(\text{voxel})}}{\Omega_L}\Big)\Big(1 - \frac{z_a^{(\text{voxel})}}{\Omega_L}\Big) \tag{A.0.1}$$

where ω_a denotes the weight at corner a, Ω_L is the edge length of the voxel and x_a, y_a and z_a are the distances in each dimension between each corner and p_n. The linear combination is then:

$$\hat{h}_{ji}^{(k)} = \sum_{a=1}^{8} \omega_a h_{ji}^{(k)a} \tag{A.0.2}$$

where $h_{ji}^{(k)a}$ is the previously calculated FR at each corner and $\hat{h}_{ji}^{(k)}$ is the estimated FR between p_n and each microphone (j).

A.0.2 Noise source suppression

This method principally exploits the property of two orthogonal vectors (Figure A.2), so that when the dot product between two vectors is calculated the result is 0.

A new filter $\hat{G}_i^{(k)} \in \mathbb{C}^{(N-1)\times N}$ is calculated from the estimated FR vector that removes the source i from the mixture, thus $G^{(k)}(i)\mathbf{x}^{(k)}(i) = \hat{\mathbf{s}}_{\{1,\ldots,i-1,i+1,\ldots,N\}}^{(k)}$. Dropping the frequency bin index, k, for convenience,

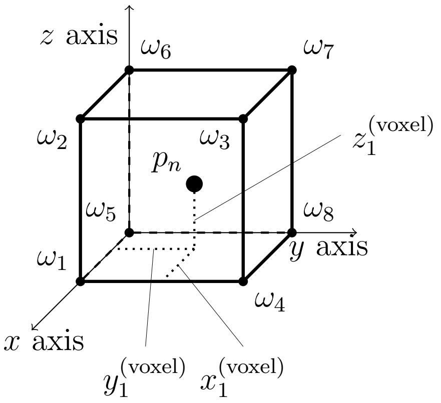

Figure A.1. An example of a voxel within a room, distances are given for corner $a = 1$. p_n is the noise source position within the voxel.

this implies: $\hat{G}_i\hat{\mathbf{h}}_i = 0$ and $\hat{G}_i\mathbf{h}_i \approx 0$, where $\mathbf{h}_i$ represents the vector of actual frequency responses which are unknown.

In a general setup, including the estimated suppression filter $\hat{G}_i$, yields:

$$\hat{G}_i^{(k)}\mathbf{x}^{(k)}(i) = \sum_{j=1}^{N} \hat{G}_i^{(k)}\mathbf{h}_j^{(k)}s_j^{(k)}(i) \tag{A.0.3}$$

The result of Equation (A.0.3) would be a distorted version of the unsuppressed sources, possibly with a small contribution from the suppressed noise source due to a mismatch between the estimated filter ($\hat{G}_i$) and its ideal value (G_i).

To find the filter ($\hat{G}_i$) one constructs an orthogonal projection [74] by:

$$\hat{G}_{i,proj}(k) = (I - \hat{\mathbf{h}}_i^{(k)}(\hat{\mathbf{h}}_i^{(k)H}\hat{\mathbf{h}}_i^{(k)-1})\hat{\mathbf{h}}_i^{(k)H}) \tag{A.0.4}$$

where $\hat{\mathbf{h}}_i$ is the steering vector of the interference source and $\hat{G}_{i,proj}$ is the projection matrix. Singular value decomposition (SVD) is then performed on $\hat{G}_{i,proj}$, so that; $\hat{G}_{i,proj} = U_i\Sigma V_i^H$, where $(\cdot)^H$ denotes the Hermitian transpose. To find the filter, all non-zero values of the diagonal matrix (denoted Σ) are used to identify the corresponding columns in the unitary matrix (denoted U), these columns in the unitary matrix correspond to the non-zero values of $\hat{G}_{i,proj}$. $\hat{G}_i$ is the concatenation of singular vectors associated with the non-null singular values, so $\hat{G}_i = [U_1, \ldots, U_{N-1}]^H$. Given a mixture, the output of the suppression stage is $\mathbf{x}^{(k)\prime}(i) = \hat{G}_i^{(k)}\mathbf{x}^{(k)}(i)$.

A.0.3 Experimental setup & results

Simulated IRs generated by the image method (IM) [53] in an almost anechoic simulation ($T_{60} = 37ms$) (Test 1) and a reverberant simulation ($T_{60} = 100ms$) (Test 2) were used when $N = 2$ to create mixtures using two 15 second speech utterances in a variety of source positions (to simulate the 3D location of a source provided by video camera information). The number of sources is $N = 2$. The FR at each corner of a voxel is calculated by the DFT of IRs generated by the IM for all tests. The voxel size used in all tests was 0.03m. Utterances from the TIMIT database from a male and female speaker were used. One utterance is a 'desired' source and the other acts as a noise source. Sources were positioned in an arc around a two-microphone array at 0.6m, the source is at 0° when it is equidistant to both microphones,

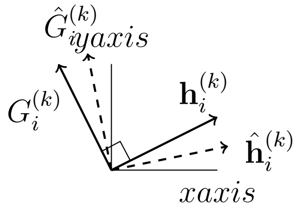

Figure A.2. Orthogonality property, when $N = 2$, where $G_i^{(k)}$ is the suppression filter vector at a particular frequency and $\mathbf{h}_i^{(k)}$ the FR vector at a particular frequency for the noise source i. $\hat{G}_i^{(k)}$ and $\hat{\mathbf{h}}_i^{(k)}$ are the estimated equivalents.

negative angles are to the left and positive ones to the right of the array (Figure A.3).

A.0.4 Simulated Mixtures

Table A.1 (Test 1) confirms that the concept of the method works well in a simulated environment. The estimated version of the second source is clearly audible, with a small amount of noise and in some cases there is a very small contribution from the suppressed source. The method also performed well when the reverberation was increased to $T_{60} = 100ms$ (Table A.2), this shows reduced performance in a slightly reverberant environment.

		Source 2 (SDR - dB)				
NS	Position 1	-	19	21	22	23
	Position 2	25	-	22	25	26
	Position 3	22	17	-	16	20
	Position 4	19	16	11	-	9.6
	Position 5	22	19	17	12	-

Table A.1. Performance of the method with IM mixtures, where $T_{60} = 37ms$ (Test 1). NS is the noise source that is to be suppressed from the mixture leaving Source 2. The optional post-processing stage has not been used in these results. Results are shown as SDR in dB.

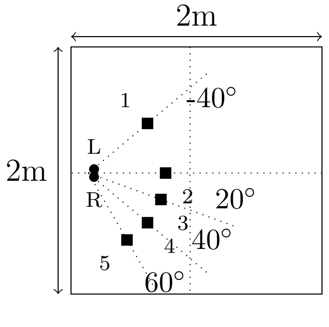

Figure A.3. 2D plan view of the simulated/physical room. Source positions are numbered 1 to 5 (0.6m), the left and right microphones are marked 'L' and 'R' respectively.

		Source 2 (SDR - dB)				
NS	Position 1	-	6.8	9.1	11	11
	Position 2	7.7	-	7.5	8.7	9.6
	Position 3	11	7.3	-	6.2	6.8
	Position 4	10	6.6	4.7	-	5.1
	Position 5	11	8.2	5.8	6.5	-

Table A.2. Performance of the method with IM mixtures, where $T_{60} = 100ms$ (Test 2). NS is the noise source that is to be suppressed from the mixture leaving Source 2. Results are shown as SDR in dB.

A.1 Summary

Further research and study will include improving the estimate of the IRs for a more realistic room environment, development of a method to correct suppression filter mismatch, expansion of the method to suppress more than one source, change of the number of sources and different types of background noise.

References

[1] E. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] E. Cherry and W. Taylor, "Some further experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 26, no. 4, pp. 554–559, 1954.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[4] J. H. McDermott, "The cocktail party problem," *Current Biology*, vol. 19, no. 22, R1024–R1027, 2009.

[5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," 1976.

[6] E. M. Z. Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann, R. R. Goodman, R. Emerson, A. D. Mehta, and J. Z. Simon, "Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party," *Neuron*, vol. 77, no. 5, pp. 980–991, 2013.

[7] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley-Interscience, 2001.

[8] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[9] J. Hérault, C. Jutten, and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé," in *Groupe d'Etudes du Traitement du Signal et des Images, GRETSI*, in French, 10 Colloque sur le traitement du signal et des images, FRA, 1985, 1985, pp. 1017–1020.

[10] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[11] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 02, pp. 163–178, 2000.

[12] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Fetal electrocardiogram extraction by blind source subspace separation," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 5, pp. 567–572, 2000.

[13] Apple Corporation Inc., *About Siri*, `https://support.apple.com/en-gb/HT204389`, Accessed: 20-06-2015, 2015.

[14] Google Inc., *Google Now*, `https://www.google.com/landing/now/`, Accessed: 20-06-2015, 2015.

[15] M. Mandel, R. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans-*

actions on Audio, Speech, and Language Processing, vol. 18, no. 2, pp. 382–394, 2010.

[16] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, 2009.

[17] A. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," *ArXiv preprint arXiv:1504.04658*, 2015.

[18] B. Rivet, W. Wang, S. Naqvi, and J. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125–134, 2014, ISSN: 1053-5888. DOI: 10.1109/MSP.2013.2296173.

[19] A. Cichocki and S.-I. Amari, *Adaptive blind signal and image processing: Learning algorithms and applications.* John Wiley & Sons, 2002, vol. 1.

[20] J. Hérault and C. Jutten, "Space or time adaptive signal processing by neural network models," in *Neural networks for computing*, AIP Publishing, vol. 151, 1986, pp. 206–211.

[21] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[22] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources," *Neural computation*, vol. 11, no. 2, pp. 417–441, 1999.

[23] 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, no. 01, pp. 1–8, 2000.

[24] J. Cardoso and A. Souloumiac, "Blind Beamforming for Non-Gaussian Signals," in *IEE Proceedings on Radar and Signal Processing*, IEE, vol. 140, 1993, pp. 362–370.

[25] A. Hyvärinen, "Independent component analysis: Recent advances," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20 110 534, 2013.

[26] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[27] M. Davies, "Audio source separation," in *Institute of Mathematics and its applications conference series*, Oxford; Clarendon, vol. 71, 2002, pp. 57–68.

[28] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *1997 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, 1997, 4–pp.

[29] F. Duplessis-Beaulieu and B. Champagne, "Fast convolutive blind speech separation via subband adaptation," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, IEEE, vol. 5, 2003, pp. V–513.

[30] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

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

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

[33] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 2158–2168, 1994.

[34] S. Makino, T.-W. Lee, and H. Sawada, *Blind speech separation*. Springer, 2007.

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

[36] T. Adali, M. Anderson, and F. Geng-Shen, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, 2014, ISSN: 1053-5888. DOI: 10.1109/MSP.2014.2300511.

[37] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: Using more sensors than source signals in a noisy mixture," in *Proceedings of ICA*, 2000, pp. 81–86.

[38] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[39] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal processing*, vol. 45, no. 2, pp. 209–229, 1995.

[40] J. Xi and J. Reilly, "Blind separation and restoration of signals mixed in convolutive environment," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, IEEE, 1997, p. 1327.

[41] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *1996 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing*, IEEE, 1996, pp. 423–432.

[42] 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, no. 5, pp. 530–538, 2004.

[43] J. Cardoso, "Multidimensional independent component analysis," in *1998 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, IEEE, vol. 4, 1998, pp. 1941–1944.

[44] A. Hyvärinen and U. Köster, “FastISA: A fast fixed-point algorithm for independent subspace analysis.,” in *14th European Symposium on Artificial Neural Networks*, 2006, pp. 371–376.

[45] I. Lee, T. Kim, and T.-W. Lee, “Independent vector analysis for convolutive blind speech separation,” in *Blind speech separation*, Springer, 2007, pp. 169–192.

[46] N. Ono, “Blind Source Separation On iPhone In Real Enviroment,” in *2013 European Signal Processing Conference, (EUSIPCO)*, 2013.

[47] ——, “Stable and fast update rules for independent vector analysis based on auxiliary function technique,” pp. 189–192, 2011.

[48] I. Lee, T. Kim, and T.-W. Lee, “Fast fixed-point independent vector analysis algorithms for convolutive blind source separation,” *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[49] H. Kuttruff, *Room acoustics.* CRC Press, 2009.

[50] M. R. Schroeder, “New method of measuring reverberation time,” *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[51] C. Brown, *Matlab central, t60.m*, `http://de.mathworks.com/matlabcentral/fileexchange/1212-t60-m`, Accessed: 10-06-2015, 2002.

[52] E. ISO, “3382-2: 2008,” *Acoustics. Measurements of room acoustics parameters. Part*, vol. 2, pp. 3382–2,

[53] J. Allen and D. Berkley, “Image method for efficiently simulating small-room acoustics,” *Journal of the Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[54] B. G. Shinn-Cunningham, N. Kopco, and T. J Martin, “Localizing nearby sound sources in a classroom: Binaural room impulse responses,” *Journal of the Acoustic Society of America*, vol. 117, p. 3100, 2005.

[55] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorländer, “Uncertainties of measurements in room acoustics,” *Acta Acustica united with Acustica*, vol. 81, no. 4, pp. 344–355, 1995.

[56] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.

[57] C. Févotte, R. Gribonval, and E. Vincent, “BSS EVAL toolbox user guide,” IRISA Technical Report 1706, Rennes, France, Tech. Rep. 1706, 2005.

[58] W. Baumann, B.-U. Kohler, D. Kolossa, and R. Orglmeister, “Real time separation of convolutive mixtures,” in *Proceedings of ICA 2001*, 2001, pp. 65–69.

[59] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, “Acoustic source separation of convolutive mixtures based on intensity vector statistics,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 748–756, 2008.

[60] I. Jolliffe, *Principal component analysis.* Wiley Online Library, 2002.

[61] S. Naqvi, Y. Zhang, and J. Chambers, “Multimodal Blind Source Separation for Moving Sources,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, IEEE, 2009, pp. 125–128.

[62] S. Naqvi, W. Wang, M. Khan, M. Barnard, and J. Chambers, "Multimodal (Audiovisual) Source Separation Exploiting Multi-Speaker Tracking, Robust Beamforming and Time-Frequency Masking," *IET Signal Processing*, vol. 6, no. 5, pp. 466–477, 2012.

[63] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 601–616, 2007.

[64] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for acoustic source localization," *IEEE Transactions on Speech Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[65] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of Multiple Moving Speakers with Multiple Microphone Arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[66] B. Rivet, L. Girin, and C. Jutten, "Visual voice activity detection as a help for speech source separation from convolutive mixtures," *Speech Communication*, vol. 49, no. 7-8, pp. 667–677, 2006.

[67] J. Harris, S. Naqvi, B. Rivet, J. Chambers, and C. Jutten, "Visual based reference for enhanced audio-visual source extraction," in *9th IMA International Conference on Mathematics in Signal Processing*, 2012.

[68] M. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, “A survey of convolutive blind source separation methods,” *Multichannel Speech Processing Handbook*, pp. 1065–1084, 2007.

[69] I. Cohen, “Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 684–699, 2003, ISSN: 1063-6676. DOI: 10.1109/TSA.2003.818105.

[70] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[71] B. Van Veen and K. Buckley, “Beamforming: A versatile approach to spatial filtering,” *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4–24, 1988, ISSN: 0740-7467. DOI: 10.1109/53.665.

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

[73] N. Forsyth, J. Chambers, and P. Naylor, “Alternating fixed-point algorithm for stereophonic acoustic echo cancellation,” *IEE Proceedings on Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 1–9, 2002.

[74] G. Golub and C. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 2012, vol. 4.

[75] A. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of computation*, vol. 27, no. 123, pp. 579–594, 1973.

[76] B. Widrow, J. R. Glover Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong Jr, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[77] B. Widrow and S. D. Stearns, *Adaptive signal processing*. Prentice Hall, 1985.

[78] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2002.

[79] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 7, pp. 1431–1438, 2010.

[80] J. Hao, I. Lee, T. Lee, and T. Sejnowski, "Independent vector analysis for source separation using a mixture of Gaussians prior," *Neural computation*, vol. 22, no. 6, pp. 1646–1673, 2010.

[81] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, IEEE, 2014, pp. 107–111.

[82] S. Ding, J. Huang, D. Wei, and A. Cichocki, "A near real-time approach for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 1, pp. 114–128, 2006.

[83] L. Oliva-Moreno, J. Moreno-Cadenas, L. Flores-Nava, and F. Gomez-Castaneda, “DSP implementation of extended infomax ICA algorithm for blind source separation,” in *3rd International Conference on Electrical and Electronics Engineering*, IEEE, 2006, pp. 1–4.

[84] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Real-time blind source separation for moving speakers using blockwise ICA and residual crosstalk subtraction,” in *Proceedings of ICA*, 2003, pp. 975–980.

[85] L. Parra and C. Spence, “On-line convolutive blind source separation of non-stationary signals,” *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 26, no. 1-2, pp. 39–46, 2000.

[86] J. Anemüller and T. Gramss, “On-line blind separation of moving sound sources,” in *Proceedings of ICA*, 1998.

[87] S. Gazor and Z. Wei, “Speech probability distribution,” *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003, ISSN: 1070-9908. DOI: 10.1109/LSP.2003.813679.

[88] R. Martin and C. Breithaupt, “Speech enhancement in the DFT domain using Laplacian speech priors,” in *2003 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2003.

[89] I. Cohen, “Speech enhancement using super-gaussian speech models and noncausal a priori SNR estimation,” *Speech communication*, vol. 47, no. 3, pp. 336–350, 2005.

[90] Y. Liang, G. Chen, S. Naqvi, and J. Chambers, “Independent vector analysis with multivariate Student’s t-distribution source

prior for speech separation," *Electronics Letters*, vol. 49, no. 16, 2013.

[91] R. Chassaing and D. Reay, *Digital Signal Processing and Applications with the TMS320C6713 and TMS320C6416 DSK*, 2nd. John Wiley & Sons, 2008.

[92] D. M. Ritchie and B. W. Kernighan, *The C programming language.* Bell Laboratories, 1975.

[93] Texas Instruments, *TMS320C6000 programmer's guide, SPRU187K*, Dallas, TX, 2002.

[94] ——, *TMS320C67x DSP library programmer's reference guide, SPRU657C*, Dallas, TX, 2010.

[95] J. F. Blinn, "Floating-point tricks," *IEEE Computer Graphics and Applications*, vol. 17, no. 4, pp. 80–84, 1997, ISSN: 0272-1716. DOI: 10.1109/38.595279.

[96] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2915–2929, 2005.

[97] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 87, no. 8, pp. 1941–1948, 2004.

[98] K. Reindl, Y. Zheng, and W. Kellermann, “Speech enhancement for binaural hearing aids based on blind source separation,” in *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, IEEE, 2010, pp. 1–6.

[99] H. Sheikhzadeh, R. Brennan, and H. Sameti, “Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications,” in *1995 International Conference on Acoustics, Speech, and Signal Processing ,(ICASSP)*, vol. 1, 1995, 808–811 vol.1. DOI: 10.1109/ICASSP.1995.479817.

[100] L. C. Parra and C. V. Alvino, “Geometric source separation: Merging convolutive source separation with geometric beamforming,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[101] M. Z. Ikram and D. R Morgan, “A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, IEEE, vol. 1, 2002, pp. I–881.

[102] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, “Real-time foreground–background segmentation using codebook model,” *Real-time imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[103] R. Y. Tsai, “A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses,” *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[104] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, vol. 1, 2001, pp. I–511.

[105] Y. Liang, J. Harris, G. Chen, S. Naqvi, C. Jutten, and J. Chambers, "Auxiliary function based Independent Vector Analysis using a source prior exploiting fourth order relationships," in *2013 European Signal Processing Conference, (EUSIPCO)*, 2013.

[106] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Adaptive step size independent vector analysis for blind source separation," in *17th International Conference on Digital Signal Processing (DSP)*, IEEE, 2011, pp. 1–6.

[107] S. Cecchi, A. Primavera, F. Piazza, and A. Carini, "An adaptive multiple position room response equalizer," 2011.

[108] S. Elliott and P. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *Journal of the Audio Engineering Society*, vol. 37, no. 11, pp. 899–907, 1989.