Multimodal Methods for Blind Source Separation of Audio Sources

Thesis submitted to Loughborough University in candidature for the degree of Doctor of Philosophy

Syed Mohsen Raza Naqvi

Advanced Signal Processing Group
Loughborough University
2009

ABSTRACT

The enhancement of the performance of frequency domain convolutive blind source separation (FDCBSS) techniques when applied to the problem of separating audio sources recorded in a room environment is the focus of this thesis. This challenging application is termed the cocktail party problem and the ultimate aim would be to build a machine which matches the ability of a human being to solve this task. Human beings exploit both their eyes and their ears in solving this task and hence they adopt a multimodal approach, i.e. they exploit both audio and video modalities. New multimodal methods for blind source separation of audio sources are therefore proposed in this work as a step towards realizing such a machine.

The geometry of the room environment is initially exploited to improve the separation performance of a FDCBSS algorithm. The positions of the human speakers are monitored by video cameras and this information is incorporated within the FDCBSS algorithm in the form of constraints added to the underlying cross-power spectral density matrix-based cost function which measures separation performance. Both objective, signal-to-interference ratio and performance index, and subjective, mean opinion score, performance measures are used to confirm the improved separation performance achieved by this multimodal method.

Further improvement is next achieved through the adoption of a fast fixed-point independent component analysis (FastICA) algorithm carefully designed for complex signals within the multimodal FDCBSS framework. An intelligent initialization strategy is developed for the FastICA algorithm on the basis of the geometric information. This method is shown to yield more robust solutions for audio separation even when the human sources are moving with a step-wise motion. In the situation where there are more microphone measurements than human sources, it is also shown that it is possible to exploit multiple combinations of microphone sensors to improve the quality of separation.

Finally, in the situation where the human sources are moving significantly within the room environment, it is proposed that FDCBSS must be combined with beamforming to provide a new solution to audio moving source separation. A Markov Chain Monte Carlo particle filter is used for tracking the true three-dimensional position of the centre of the head of the target human speakers on the basis of a state evolution model of position and velocity. The improved separation performance of the schemes is confirmed through both objective and subjective measures.

To my parents.

CONTENTS

ACKNOWLEDGEMENTS

I would like to give my immense thanks to Prof. Jonathon A. Chambers for his greatness as a leader, guide, and enthusiastic supervisor throughout this work. I may have left this work incomplete if he had not been my supervisor. Prof. Jonathon A. Chambers has kindly spent much time working with me regarding the problems of my research. Discussions that we had during our meetings were never one sided as he allowed me to express my thoughts about any aspect of my studies. He provided me great support to attend international conferences and meet the professionals in the field of signal processing all over the globe. His advice and comments on the publications and the thesis drafts have been invaluable. Without them, I would not have finished this thesis.

I wish to give special thanks to Doctor Raza Sammar for partial financial support for this research work. Additional thanks to Prof. Jonathon A. Chambers for further financial support to complete this study.

I wish to thank Doctors Saeid Sanei, Sangarapillai Lambotharan, Julia Hicks and Yonggang Zhang for their valuable discussions and comments on my work and for having been co-authors of the research papers.

I must express my gratitude to my family for their unconditional love & prayers for my success and to Allah who gives nothing to those

who keep their arms crossed.

Lastly, all members of the advanced signal processing group (ASPG) deserve thanks for making the ASPG laboratory a constructive environment for research work. Our discussions during the PhD period have provided me with much needed information, such intention has been invaluable.

STATEMENT OF ORIGINALITY

The original contributions are on exploiting visual aspects of speech with audio for source separation. The novelty of the contributions are supported by one full journal paper submitted after review, one complete book chapter in press, and, six published conference papers.

In Chapter 3, a novel geometrically constrained multimodal method for second-order statistics (SOS) based convolutive blind source separation is presented. Audio-visual information is integrated through a penalty function-based formulation to improve the BSS algorithm. A comparative study of the proposed method with the two emerging frequency domain convolutive blind source separation (FDCBSS) techniques is also presented in this chapter. The results of this method have been published in:

- **S. M. Naqvi**, Y. Zhang, T. Tsalaile, S. Sanei and J. A. Chambers "Evaluation of emerging frequency domain convolutive blind source separation algorithms based on real room recordings," *in Proc. Int. Conf. IEEE SAM*, 2008, Darmstadt, Germany.

- S. Sanei, **S. M. Naqvi**, J. A. Chambers and Y. Hicks "A geometrically constrained multimodal approach for convolutive blind

source separation," *in Proc. Int. Conf. ICASSP*, 2007, Hawaii, USA.

In Chapter 4, a novel multimodal method for higher-order statistics (HOS) based independent component analysis (ICA) of complex valued frequency domain signals is presented which utilizes video information to provide geometrical description of both the speakers and the microphones. This geometric information is intelligently incorporated into the initialization of the complex FastICA algorithm, which not only solves the inherent permutation problem in frequency domain CBSS but also improves the rate of convergence for static sources. In this chapter, this multimodal method is also improved for separation of step-wise moving sources. The results of both methods have been published in:

- **S. M. Naqvi**, Y. Zhang, T. Tsalaile, S. Sanei and J. A. Chambers "A multimodal approach for frequency domain independent component analysis with geometrically-based initialization," *in Proc. Int. Conf. EUSIPCO*, 2008, Lausanne, Switzerland.

- **S. M. Naqvi**, Y. Zhang and J. A. Chambers "A multimodal approach for frequency domain blind source separation for moving sources in a room," *in Proc. Int. Conf. IAPR CIP*, 2008, Santorini, Greece.

In Chapter 5, based on the multimodal method presented in Chapter 4, a new approach to overdetermined frequency domain blind source separation (BSS) of speech signals which exploits all combinations of observations and hence varying inter-microphone spacings is presented.

The observations are divided into subgroups so that conventional frequency domain BSS algorithms can be used.

In Chapter 6, a novel multimodal solution is proposed for the problem of blind source separation (BSS) of moving sources. In the proposed approach, the visual modality is utilized to facilitate the separation for both stationary and moving sources. To obtain the positions and velocities of the sources, a full 3-D visual tracker based on a Markov Chain Monte Carlo particle filter (MCMC-PF) is implemented. The complete BSS solution is formed by integrating a frequency domain blind source separation algorithm and beamforming. The proposed method not only improves the performance of the BSS algorithm and mitigates the permutation problem for stationary sources, but also provides a good BSS performance for moving sources. The results of the complete solution are presented in:

- **S. M. Naqvi**, Y. Zhang and J. A. Chambers "Multimodal Blind Source Separation for Moving Sources," *in Proc. Int. Conf. ICASSP*, 2009, Taipei, Taiwan.

- **S. M. Naqvi**, Y. Zhang, M. Yu and J. A. Chambers "A Multimodal Approach to Blind Source Separation of Moving Sources," *submitted after review, IEEE Transactions on Multimedia.*

- **S. M. Naqvi** and Editors, *Machine Audition: Principles, Algorithms and Systems: A Multimodal Solution to Blind Source Separation of Moving Sources.* Will be published by the IGI Global, USA.

The results of another contribution which is related to the general approaches adopted in this work is published in:

- T. Tsalaile, **S. M. Naqvi**, K. Nazarpour, S. Sanei and J. A. Chambers "Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound," *in Proc. Int. Conf. ICASSP*, 2008, Las Vegas, USA.

List of Acronyms

ABF	Adaptive Beamformer
CBSS	Convolutive Blind Source Separation
CPP	*Cocktail Party Problem*
DFT	Discrete Fourier Transform
DOA	Direction Of Arrival
FDCBSS	Frequency Domain CBSS
FIR	Finite Impulse Response
HOS	Higher-Order Statistics
ICA	Independent Component Analysis
IDFT	Inverse DFT
IIFastICA	Intelligently Initialized FastICA
KF	Kalman Filter
LCMV	Linearly Constrained Minimum Variance
LS	Least-Squares
MCMC-PF	Markov Chain Monte Carlo Particle Filter

MH	Metropolis-Hastings
MIMO	Multi-Input-Multi-Output
MOS	Mean Opinion Score
MSE	Mean Square Error
MVDR	Minimum Variance Distortionless Response
PCA	Principal Component Analysis
pdf	Probability Density Function
PI	Performance Index
PSD	Power Spectral Density
RT	Reverberation Time
s.t	subject to
SIR	Signal-to-Interference Ratio
SISO	Single-Input-Single-Output
SOS	Second-Order Statistics
SPL	Sound Pressure Level
w.r.t	with respect to

LIST OF SYMBOLS

$\|.\|$	Absolute value
$\|\|.\|\|_2$	Euclidean norm
$\|\|.\|\|_F$	Frobenius norm
$(.)^T$	Transpose operator
$(.)^H$	Hermitian transpose operator
$(.)^\dagger$	Pseudo-inverse
Λ	Diagonal matrix
$diag(\mathbf{d})$	Diagonal matrix with vector $\boldsymbol{d}$ on its main diagonal
$E\{\cdot\}$	Statistical expectation operator
$\mathbf{I}$	Identity matrix
$kurt(.)$	Kurtosis
N	Number of mixtures
M	Number of sources
$Neg(.)$	Negentropy
ω	Angular frequency

R	Covariance matrix
$\mathbf{S}$	Source matrix
T	FFT length
$J(.)$	Cost function
$\mathbf{Q}$	Whitening matrix
Q	Length of unmixing filter
$Var(.)$	Variance
$G(.)$	Nonlinear contrast function
$g(.)$	First derivative of nonlinear contrast function
$\acute{g}(.)$	Second derivative of nonlinear contrast function
$\mathbf{h}_i$	ith column vector of mixing matrix
$\mathbf{H}$	Mixing matrix
$\mathbf{w}$	Estimated unmixing vector
$\mathbf{W}$	Estimated unmixing matrix
$\mathbf{P}$	Permutation matrix
$\mathbf{G}$	Overall system matrix
P	Length of mixing filer

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 The cocktail party problem

The term *cocktail party* implies a gathering of people in a room where several people are participating in a conversation, some might be moving while talking and background music may also be being played. Colin Cherry, in 1953 [1], first defined the *cocktail party problem* and Cherry and Taylor in 1954 [2] further worked on this problem, which is defined as:

"How do we recognise what one person is saying when others are speaking at the same time (the "cocktail party problem")?" - Colin Cherry 1954 [2]

Simon Haykin in [3] explained that tackling the *cocktail party problem* is underpinned by a psychoacoustic phenomenon which refers to the remarkable human ability to selectively attend and recognize one speech source amongst the many competing speech sources, all of which are usually assumed to be independent of each other. Over half a century after the first definition of the *cocktail party problem* by Cherry, a complete understanding of the psychoacoustic phenomenon exploited by a human is still missing.

The long term research aim of scientists working on the *cocktail*

party problem is, by exploiting advanced computing and signal-processing technologies, to build an intelligent machine which can mimic the ability of the human auditory system to solve the *cocktail party problem.*

During the last two decades, the increase in computing power has motivated researchers to attempt to produce a real time solution to the *cocktail party problem* as represented in Figure 1.1.

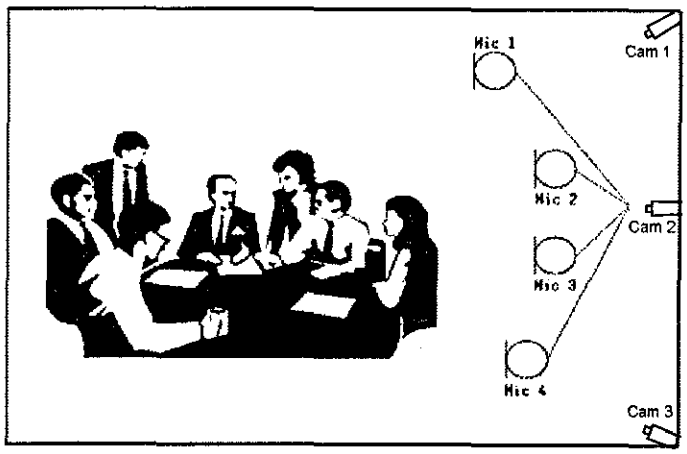

Figure 1.1. *Machine cocktail party problem*: to build an intelligent machine which can duplicate some aspects of the human auditory system to solve the *cocktail party problem* through microphones and video measurements.

Attempts to solve the *machine cocktail party problem* have come from the signal processing community in the form of blind source separation (BSS) and generally from the computer science community in the form of computational auditory scene analysis [4]. The work in this thesis through exploiting geometric information is a combination of the two approaches.

The purpose of BSS is to recover unobserved source signals from observed mixtures exploiting only the assumption of mutual independence between the source signals [5], and this is next explained in more detail.

1.2 Blind source separation

In the field of signal processing the most popular approach to solve the *cocktail party problem* is blind source separation (BSS). This is a statistical approach by which individual sources can be estimated from measurements containing mixtures of sources. The term "blind" refers to the fact that BSS relies on only weak assumptions on the sources and the mixing process. These weak assumptions enable the approach to be potentially used in a wide variety of applications, including teleconferencing, bio-medicine, financial time series analysis, security surveillance or as a pre-processing step for speech recognition [6, 7].

Many methods have been proposed to attempt to solve the BSS problem. Herault and Jutten in 1985 [8] seem to be the first who addressed the problem of blind source separation. In the standard BSS problem, the mixtures are assumed to be instantaneous. Instantaneous means that there is essentially no delay between the sources and the sensors i.e. the source takes a direct and zero delay path to the sensor. Common in 1994 [9] formulated the problem of separating measurements formed from an instantaneous linear mixing model and clearly defined the term independent component analysis. He also presented an algorithm that measures independence by capturing higher-order statistics (HOS) of the sources.

However, the instantaneous model is not useful for solving the *cocktail party problem* as it does not represent most real-world environments, where acoustic sources take multiple paths to the microphone sensor measurements. The convolutive model on the other hand is representative of the practical situation. Two types of mixing model exist in the convolutive case, anechoic and echoic. Anechoic mixing simply

represents the transmission delays between the sources and sensors, whereas echoic mixing as shown in Figure 1.2 represents the source to sensor delays and also reverberations (echoes) of the sources. The literature is generally concerned with the echoic mixture model as it is the most useful for representing a reverberant environment and it contains the anechoic form as a special case. In echoic or convolutive BSS, each element of the mixing filter $\mathbf{H}$ is in fact a linear filter to simulate multipaths from sources to sensors.

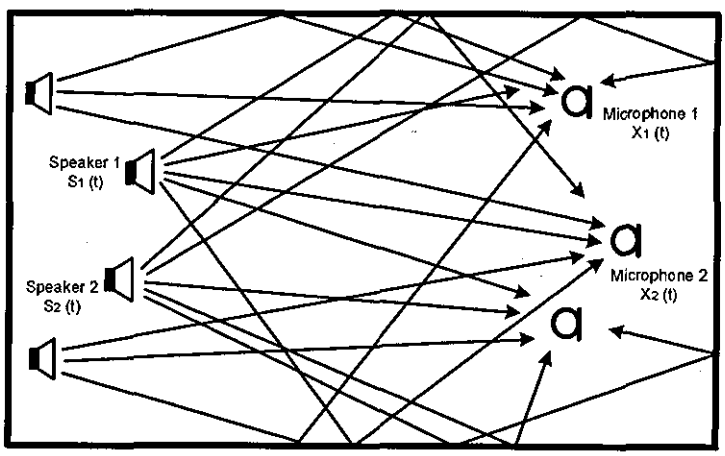

Figure 1.2. A convolutive mixing environment with two Sources, Speakers 1 & 2, and two Sensors, Microphones 1 & 2, together with other interferences.

During the past decade there has been considerable research performed in the field of convolutive blind source separation (CBSS) e.g. [10–26]. Most existing CBSS algorithms, initially, assume that the sources are physically stationary i.e. the mixing filters are fixed, which is not always true in the *cocktail party problem* where sources may be moving and secondly, are uni-modal, relying solely on audio information. Fundamentally, it is very difficult to separate convolutive mixed signals by utilizing the statistical information only extracted from audio signals, and this is not the manner in which humans solve the prob-

lem [3] since they generally use both their ears and eyes. These are the main motivations for the multimodal methods proposed in this thesis.

1.3 Why multimodal methods ?

Colin Cherry, in 1953, mentioned that it is best to combine and exploit the information provided by audio and visual measurements to replicate the ability of a human to solve the *machine cocktail party problem.* In [27] it has been shown that a speaker's face in a noisy environment greatly improves the intelligibility of that person's voice. The McGurk [28] effect also highlights the relationship between the audio and visual aspects of speech and how humans perceive speech. Visual cues, for example, are used to determine who is being addressed. Human speech is inherently bimodal, with both audio and visual components [4]. Colin Cherry observed that the human approach to solve the cocktail party problem exploits visual cues [1,2] and Simon Haykin in [3] also highlighted the importance of visual information to solve the *machine cocktail party problem.*

There is no significant literature in the area of bi-modal or video assisted CBSS. However, audio-visual information is already used in speech acquisition, speaker tracking, source localization, and speech enhancement [29–40]. In the context of CBSS, in unimodal systems there are generally no priori assumptions on the source statistics or the mixing system. On the other hand, in a multimodal approach the video system can capture the positions of the speakers and the directions they face [41]. The video information can thereby help to estimate the mixing filters $\mathbf{H}$ more accurately and ultimately increase the separation performance. Following this idea, the first objective of this work is to

use efficiently such information in the enhancement of the separation of stationary sources.

Most existing BSS algorithms are based on statistical information, second-order statistics (SOS) and/or higher-order statistics (HOS), extracted from the recorded data. Such methods are generally not applicable in CBSS of moving sources. The challenge of CBSS for moving sources is that the mixing filters are time varying, thus the unmixing filters should also be time varying, which are difficult to calculate in real time on the basis of only audio information [34,42,43]. To fulfil the second and last objective, in the proposed methods, the visual modality is utilized to facilitate the separation for both stationary and moving sources.

This thesis develops methods to answer the following important questions:

- How can visual information be integrated in the existing CBSS algorithms with audio to improve the CBSS of stationary sources?
- How can multiple moving sources be best detected and tracked by utilizing audio-visual information?
- How can audio-visual information be incorporated to solve the BSS of multiple moving sources?

The first question is answered in the initial part of this thesis. The last two questions are answered in detail, in the final part of this thesis.

The organization of this thesis with brief overview of each chapter is presented next.

1.4 Organization of the thesis

This thesis is organized as follows:

- Chapter 2 lays the foundation of the thesis. The main objective of this chapter is to introduce the fundamentals of CBSS, limitations of the key techniques in CBSS, and performance measures required for objective and subjective evaluation of CBSS. A key component of multimodal methods for CBSS of moving sources is tracking of speakers therefore an overview of nonlinear Bayesian filtering (particle filters) is also provided in this chapter.

- Chapter 3 improves BSS based on SOS by a novel constrained multimodal approach for CBSS. Audio-visual information is integrated through a penalty function-based formulation to solve the permutation problem and enhance the source separation. The separation is performed in the frequency domain and the geometrical model which provides the geometrical positions of the sources and the sensors is also described in this chapter.

- Chapter 4 presents a novel multimodal method for higher-order statistics (HOS) based independent component analysis (ICA) of complex valued frequency domain signals, which utilizes video information to provide geometrical description of both the speakers and the microphones. This geometric information is incorporated into the initialization of the complex FastICA algorithm for each frequency bin, which not only solves the inherent permutation problem in the frequency domain CBSS (with complex valued signals) but also improves the rate of convergence for static sources. In this chapter, this multimodal method is also improved by ex-

ploiting the permutation free unmixing matrix of the previous block together with the whitening matrix of the mixtures of the current block, to initialize intelligently FastICA for separation of step-wise moving sources.

- Chapter 5, based on the multimodal method presented in Chapter 4, provides a new approach to overdetermined frequency domain blind source separation (BSS) of speech signals which exploits all combinations of observations and hence varying inter microphone spacings. A conventional scheme using only one microphone group and an existing overdetermined frequency domain BSS algorithm are also compared in this chapter.

- Chapter 6, explains a novel multimodal solution for the problem of blind source separation (BSS) of moving sources. In the proposed method, the visual modality is utilized to facilitate the separation for both stationary and moving sources. To obtain the positions and velocities of the sources, a full 3-D visual tracker based on a Markov Chain Monte Carlo particle filter (MCMC-PF) is implemented, which results in high sampling efficiency. The complete BSS solution is formed by integrating a frequency domain blind source separation algorithm and beamforming: on the basis of velocity obtained from the 3-D visual tracker, if the sources are identified as stationary for a certain minimum period, a frequency domain BSS algorithm is implemented. Once the sources are moving, a beamforming algorithm which requires no prior statistical knowledge is used to perform real time speech enhancement and provide separation of the sources. The proposed

method not only improves the performance of the BSS algorithm and mitigates the permutation problem for stationary sources, but also provides a good BSS performance for moving sources.

- Chapter 7, concludes the thesis and includes suggestions for future work.

Chapter 2

FUNDAMENTALS OF BLIND SOURCE SEPARATION

The main work presented in this thesis is based on blind source separation (BSS) therefore the problem statement, techniques, limitations, and, performance measures of BSS are discussed. A key component of the proposed multimodal solution to BSS of moving sources is tracking of speakers therefore a brief overview of nonlinear Bayesian filtering (particle filters) is also presented in this chapter.

2.1 Problem statement

The BSS problem is to recover M unobserved source signals contained in $\mathbf{s}(t)$ from the N observed mixture signals contained in $\mathbf{x}(t)$ with minimum assumptions about the mixing medium and the underlying sources.

A classical instantaneous generative model can be described as:

$$x_i(t) = \sum_{j=1}^{M} h_{ij} s_j(t) \qquad i = 1, ..., N \tag{2.1.1}$$

where $x_i(t)$ denotes the i-th element of mixture column vector $\mathbf{x}(t) \in \mathbb{R}^N$, $s_j(t)$ denotes the j-th element of source column vector $\mathbf{s}(t) \in \mathbb{R}^M$,

t denotes the discrete time index, and h_{ij} is the attenuation element of the mixing matrix $\mathbf{H}$ corresponding to its i-th row and j-th column.

In convolutive blind source separation (CBSS) the sources are assumed to be convolved with a linear model of the physical medium (mixing matrix) which can be represented in the form of a multichannel (in the real world e.g. a room, the speech signals recorded by the microphones are affected by reverberations [20, 44]) FIR filter $\mathbf{H}(p)$, $p = 0, ..., P-1$ to produce N sensor signals. In this thesis to demonstrate the proposed methods the exactly determined CBSS problem i.e. $N = M = 2$, is considered except in Chapter 5 (here the noise free case is also assumed to simplify the formulation)

$$x_i(t) = \sum_{j=1}^{M} \sum_{p=0}^{P-1} h_{ij}(p) s_j(t-p) \qquad i = 1, ..., N \tag{2.1.2}$$

where $h_{ij}(p)$, $p = 0, \ldots, P-1$, is the P-tap impulse response from source j to microphone i and the p-th slice of the FIR filter $\mathbf{H}(p)$ is:

$$\mathbf{H}(p) = \begin{bmatrix} h_{11}(p) & \cdots & h_{1M}(p) \\ \vdots & \ddots & \vdots \\ h_{N1}(p) & \cdots & h_{NM}(p) \end{bmatrix} \tag{2.1.3}$$

In instantaneous BSS the sources are estimated by:

$$y_j(t) = \sum_{i=1}^{N} w_{ji} x_i(t) \qquad j = 1, ..., M \tag{2.1.4}$$

where $y_j(t)$ denotes the j-th element of the estimated source column vector $\mathbf{y}(t)$, and w_{ji} is the gain element of the so-called separating or unmixing matrix $\mathbf{W}$ corresponding to its j-th row and i-th column.

In time domain CBSS, the sources are estimated using a set of unmixing filters $\mathbf{W}(q)$, $q = 0, \ldots, Q-1$ such that

$$y_j(t) = \sum_{i=1}^{N} \sum_{q=0}^{Q-1} w_{ji}(q) x_i(t-q) \qquad j = 1, ..., M \tag{2.1.5}$$

where the q-th slice of the unmixing filter $\mathbf{W}(q)$ is:

$$\mathbf{W}(q) = \begin{bmatrix} w_{11}(q) & \cdots & w_{1N}(q) \\ \vdots & \ddots & \vdots \\ w_{M1}(q) & \cdots & w_{MN}(q) \end{bmatrix} \tag{2.1.6}$$

Using a T-point windowed discrete Fourier transformation (DFT), the time domain signals $x_i(t)$, $i = 1, ..., N$, can be converted into time-frequency domain signals $x_i(\omega, t_k)$ where ω is a normalized frequency index and t_k, $k = 1, ..., K$, is a discrete time index (K represents the total number of data blocks and is only required in Chapter 3 for second order statistics (SOS) based BSS discussed in the sequel). For each frequency bin it can be written as:

$$\mathbf{x}(\omega, t_k) = \mathbf{H}(\omega)\mathbf{s}(\omega, t_k) \tag{2.1.7}$$

where $\mathbf{s}(\omega, t_k) = [s_1(\omega, t_k), ..., s_N(\omega, t_k)]^H$ and $\mathbf{x}(\omega, t_k) = [x_1(\omega, t_k),$ $..., x_N(\omega, t_k)]^H$, where $(\cdot)^H$ denotes Hermitian transpose, are the time-frequency representations of the source signals and the observed signals, $\mathbf{H}(\omega)$ is an N-by-M matrix composed of $h_{ij}(\omega)$, which is the frequency representation for the mixing impulse response $h_{ij}(p)$. It is assumed that $\mathbf{H}(\omega)$ is invertible when $N = M$, and does not depend on time

and can be represented as:

$$\mathbf{H}(\omega) = \begin{bmatrix} h_{11}(\omega) & \cdots & h_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ h_{N1}(\omega) & \cdots & h_{NM}(\omega) \end{bmatrix} \tag{2.1.8}$$

The separation can be completed by an M-by-N unmixing matrix $\mathbf{W}(\omega)$ at a frequency bin ω

$$\mathbf{y}(\omega, t_k) = \mathbf{W}(\omega)\mathbf{x}(\omega, t_k) \tag{2.1.9}$$

$$\mathbf{W}(\omega) = \begin{bmatrix} w_{11}(\omega) & \cdots & w_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ w_{M1}(\omega) & \cdots & w_{MN}(\omega) \end{bmatrix} \tag{2.1.10}$$

where $\mathbf{y}(\omega, t_k) = [y_1(\omega, t_k), ..., y_M(\omega, t_k)]^H$ is the time-frequency representation of the estimated source signals and $\mathbf{W}(\omega)$ is the frequency representation of the unmixing matrix. $\mathbf{W}(\omega)$ is determined so that $\hat{s}_i(\omega, t_k) = y_i(\omega, t_k)$, $i = 1, ..., M$, become as mutually independent as possible.

The time domain separated signals $\hat{s}_i(t) = y_i(t)$, $i = 1, ..., M$, can then be obtained by using an inverse DFT (IDFT) operation, provided the scale and permutation ambiguities (discussed in the sequel) are mitigated.

2.2 Techniques for BSS

This section briefly overviews two techniques, second-order statistics (SOS) based BSS and higher-order statistics (HOS) based BSS known as independent component analysis (ICA).

2.2.1 Second-order statistics based BSS

In SOS based separation algorithms the sources are separated on the basis of decorrelation rather than independence. These methods assume that the sources are statistically non-stationary or have a minimum phase mixing system [45–52]. However, second order statistics are not sufficient for separation of stationary sources and the required conditions are presented in [53, 54], but the main advantage of SOS is that they require shorter data lengths for accurate estimation [47, 53, 55, 56].

Statistical non-stationarity

For time scales beyond 10ms speech signals can be considered statistically non-stationary [57, 58]. The algorithms which exploits such non-stationarity were proposed in [19, 59]. Parra and Spence [19] proposed frequency domain algorithm which jointly diagonalizes the unmixing matrix $\mathbf{W}(\omega)$ for all frequency bins by minimizing the sum squared error (as the sum of off diagonal elements of the covariance matrix of the estimated sources) using the gradient descent algorithm [60]. Since the performance of this technique is improved in Chapter 3 by using a multimodal method it is important to overview the important concepts within [19] in this chapter. In [19] the unmixing matrix $\mathbf{W}(\omega)$ is found

across all frequency bins from

$$\begin{aligned}\mathbf{R}_Y(\omega, t_k) &= \mathbf{W}(\omega)\mathbf{R}_x(\omega, t_k)\mathbf{W}^H(\omega) \\ &= \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{\Lambda}_s(\omega, t_k)\mathbf{H}^H(\omega)\mathbf{W}^H(\omega)\end{aligned}$$

where $\mathbf{\Lambda}_s(\omega, t_k)$ is a diagonal covariance matrix describing the source signals and is a different diagonal matrix for each time block t_k, and $\mathbf{R}_x(\omega, t_k)$ is the covariance matrix of $\mathbf{X}(\omega, t_k)$. The covariance matrices are estimated using an averaged cross-power spectrum

$$\hat{\mathbf{R}}_x(\omega, t_k) = \frac{1}{L}\sum_{n=1}^{L-1}\mathbf{X}(\omega, t_k + nT)\mathbf{X}^H(\omega, t_k + nT) \tag{2.2.1}$$

where T is the block length of the FFT. The cost function J_m based on the off-diagonal elements of $\mathbf{R}_Y(\omega, t_k)$ estimated at $t_k = kTL$, $k = 1, 2, \ldots, K$, with K being the number of matrices to diagonalize, is

$$J_m = \sum_{\omega=1}^{T}\sum_{k=1}^{K}\|E(\omega, t_k)]\|_F^2 \tag{2.2.2}$$

where $E(\omega, t_k) = \{\mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t_k)\mathbf{W}^H(\omega)\} - \mathbf{\Lambda}_s(\omega, t_k)$, and $\|.\|_F^2$ is the squared Frobenius norm. To avoid the trivial $\mathbf{W}(\omega) = \mathbf{0}\ \forall\omega$ solutions, the constant $diag(\mathbf{W}(\omega)) = I\ \forall\omega$ is applied. To minimize (2.2.2) the method of steepest descent [60] is applied to yield

$$\frac{\partial J_m}{\partial \mathbf{W}(\omega)} = 2\sum_{k=1}^{K}E(\omega, t_k)\mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t_k) \tag{2.2.3}$$

and the update equation for $\mathbf{W}(\omega)$ becomes

$$\mathbf{W}_{j+1}(\omega) = \mathbf{W}_j(\omega) - \mu\sum_{k=1}^{K}E(\omega, t_k)\mathbf{W}_j(\omega)\mathbf{R}_x(\omega, t_k) \tag{2.2.4}$$

where j and μ are the iteration index and learning rate respectively. The unmixing filter matrix $\mathbf{W}(\omega)$ is updated for all the frequency bins. The source covariance matrix can be estimated at each iteration by $\hat{\mathbf{\Lambda}}_s(\omega, t_k) = diag\{\mathbf{W}(\omega)\mathbf{R}_x(\omega, t_k)\mathbf{W}^H(\omega)\}$.

Another common approach to solving BSS problem is to use a particular kind of objective function in the context of independent component analysis (ICA) [6,61] presented next.

2.2.2 Independent component analysis

ICA is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA. ICA is superficially related to principal component analysis and factor analysis. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classical methods fail completely. For ICA to work some assumptions [6,61] must be taken.

- The sources are assumed to be statistically independent of each other. Mathematically, independence implies that the joint probability density function $p(s(t))$ of the sources can be factorized

as:

$$p(s(t)) = \prod_{j=1}^{m} p_j(s_j(t)) \tag{2.2.5}$$

where $p_j(s_j(t))$ is the marginal distribution of the j-th source.

- All but one of the sources must have *non-Gaussian* distributions.
- The unknown mixing matrix is usually assumed to be square and invertible. In other words it is assumed that the number of sources is equal to the number of mixtures, i.e. an exactly determined problem.
- Methods to realize ICA are more sensitive to data length than the methods based on SOS.

In ICA the statistical independence of the sources implies the uncorrelatedness of the sources, but the reverse is not necessarily true. As a pre-processing step, most ICA algorithms decorrelate (pre-whiten) the mixtures via spatial whitening, before optimizing their separating objective contrast or cost functions. This spatial whitening is achieved by employing the well-known principal component analysis (PCA).

Principal component analysis

In the context of BSS, principal component analysis (PCA) seeks to remove the cross-correlation between the observed signals, and ensures that they have unit variance [6]. PCA operates by finding the projections of the mixture data in orthogonal directions of maximum variance. A zero mean vector $\mathbf{z}$ containing observations from spatially distinct locations is said to be spatially white if

$$E\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I} \tag{2.2.6}$$

where $E\{.\}$ is the statistical expectation operator, $(.)^T$ is the transposed operator, and $\mathbf{I}$ is the identity matrix. The unmixing matrix, $\mathbf{W}$, can be decomposed into two components as:

$$\mathbf{W} = \mathbf{UQ} \tag{2.2.7}$$

where $\mathbf{Q}$ denotes the whitening matrix and $\mathbf{U}$ is the rotation matrix [6]. For $m = n$, there are n^2 unknown parameters in $\mathbf{W}$. PCA requires the n diagonal elements of the whitened data covariance matrix $\mathbf{C_z}$ to be unity, and due to the symmetry property of $\mathbf{C_z}$, it suffices that only $(n^2 - n)/2$ of its off-diagonal terms be zero. Therefore, spatial whiteness imposes $n(n+1)/2$ constraints. This leaves $n(n-1)/2$ unknown parameters. The whitening matrix $\mathbf{Q}$ can be formulated as:

$$\mathbf{Q} = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T \tag{2.2.8}$$

where $\mathbf{E}$ is the matrix whose columns are the unit-norm eigenvectors of the covariance matrix $\mathbf{C_x} = \mathbf{EDE}^T$ and $\mathbf{D}$ is the diagonal matrix of the eigenvalues of $\mathbf{C_x}$. $\mathbf{D}^{-\frac{1}{2}}$ plays a vital role in $E\{\mathbf{zz}^T\} = I$ and it is also important to note that the whitening matrix $\mathbf{Q}$ is not unique because it can be pre-multiplied by an orthogonal matrix to obtain another version of $\mathbf{Q}$.

Actually, decorrelation or "no correlation" deals with only SOS and independence is a stronger concept because it deals with HOS. Cardoso in [5] explained that *"prewhitening only does half of the BSS job"*.

Fundamentally, ICA relies on two factors: 1) A statistical criterion expressed in terms of a cost/contrast function $J(\mathbf{y}(t))$ which requires to be either minimized or to be maximized, 2) An optimization technique

to carry out the minimization or maximization of the cost function.

Independent components can be found by nonlinear, non-stationary, or time delay decorrelation. In depth studies of ICA theories conclude that the nonlinear decorrelation is a satisfactory way to separate the independent components and the algorithms are based on the minimization of mutual information, maximization of non-Gaussianity, or maximization of likelihood. This class of ICA algorithms is derived from an information theoretic perspective in [62,63]. In most classical statistical theories random variables are assumed to have a Gaussian distribution. In the theory of ICA, as mentioned above, random variables are assumed to have a non-Gaussian distribution. Since generally speech has super-Gaussian distribution or is leptokurtic therefore it satisfies the ICA theory and independent components in this work are calculated on the basis of maximization of non-Gaussianity.

2.3 Limitations of SOS/ICA

During the past decade there has been considerable research performed in the field of convolutive blind source separation (CBSS) [10–26]. Initially, research was aimed at solutions based in the time domain [16,64,65]. As shown in Figure 1.2, recordings taken in a real room (convolutive environment) where the impulse response of the room is on the order of 1000's of samples in length, a time domain algorithm would be computationally very expensive to separate the sources [22]. To overcome this problem, a solution in the frequency domain was proposed. As convolution in the time domain is equivalent to multiplication in the frequency domain, then transforming into the frequency domain simplifies the convolutive mixing problem to that of independent complex

instantaneous mixing operations at each frequency bin. In realization, care is necessary to overcome circular convolution effects [6]. Transferring into the frequency domain provides two advantages. Initially, the computational complexity is reduced and secondly, ICA can be applied at each frequency bin as an instantaneous BSS problem, but the two indeterminacies, namely the scaling and permutation ambiguities, inherent to BSS, become more severe.

The main limitations of SOS/ICA are as follows:

- Permutation problem: The order in which the uncorrelated/independent components are recovered cannot be determined due to the "blindness" of the problem, i.e. both the mixing matrix and the sources are unknown [11]. Thus, a change in the order of the recovered sources also implies a permutation of the corresponding colums of the mixing matrix.

- Scaling ambiguity: It is not possible to determine the energy of the original uncorrelated/independent components e.g.

$$\mathbf{x}(t) = (\frac{1}{c_k}\mathbf{h}_k)(c_k s_k(t)) + \sum_{j \neq k} \mathbf{h}_j s_j(t) \qquad (2.3.1)$$

  It is clear that the multiplying factor c_k to the kth source can be cancelled out by dividing the kth column of the mixing matrix $\mathbf{H}$ by the same factor c_k. This demonstrates that the sources can be estimated only up to a scaling constant. However, some researchers force the estimated sources to have unit variance [11].

- Data length sensitivity: Since SOS/ICA methods are based on statistical information, they are sensitive to data length, and

therefore such approaches are unlikely to be applicable to solve the problem of BSS of moving sources.

The first two ambiguities show that it is not necessary for the separating matrix $\mathbf{W}$ to extract exactly the inverse of the mixing matrix $\mathbf{H}$. Instead,

$$\mathbf{W} = \mathbf{P}\mathbf{\Lambda}\mathbf{H}^{-1} \tag{2.3.2}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is a diagonal matrix to convey the scaling ambiguity.

In summary, in FDCBSS based on SOS/ICA, perfect separation cannot be achieved without additional information. This is another motivation for the proposed multimodal methods which not only mitigate the above problems but also provide BSS for moving sources, a substantial step forward towards the solution of the real *cocktail party problem.*

In the next subsections several recent attempts to overcome the severe permutation problem in FDCBSS are discussed.

2.3.1 Solutions to the permutation problem

As already mentioned, by transferring the BSS problem from the time domain to the frequency domain results in more severe permutation ambiguity. The amplitude (scale) ambiguity can be managed by matrix normalization [66]. As mentioned above a popular approach to FDCBSS problem is based on SOS. One of the more effective frequency domain BSS algorithms based on SOS by Parra and Spence is mentioned in Subsection 2.2.1. Parra and Spence utilized second order statistics by exploiting the non-stationarity of speech and provided a

solution to the permutation problem. They performed the separation in the frequency domain and used a multiple decorrelation approach based upon a gradient descent algorithm [60] to estimate the mixing/unmixing matrix. The non-stationarity of speech is exploited to diagonalize simultaneously the covariance matrices estimated at each time interval. The solution to the permutation problem is proposed by imposing a smoothness constraint on the unmixing filters. The smoothing essentially forces the frequency bins to align and is achieved by constraining the filter length in the time domain to be much less than the size of the DFT (discrete Fourier transform). Restricting the length of the filter in the time domain forces the solutions in the frequency domain to be continuous or smooth. However, it has been shown in [67,68] that Parra's method failed to align all the permutations when used in a realistic environment.

Sawada et al. [69] proposed a method to solve the permutation problem by integrating two known approaches, direction of arrival (DOA) and inter-frequency correlation of signal envelopes. They used a combination of the natural gradient and information maximization algorithms to perform the initial speech separation then aligned the permutations in two stages; first to fix the permutations at those frequencies where the confidence of the DOA approach is high and secondly to decide the permutations for the remaining frequencies based on neighboring correlations without changing those fixed by the DOA method. A method for DOA estimation for more than two sources is also proposed, and by exploiting the harmonic structure of the signal they were able to align the permutations at low frequencies where it is difficult to estimate the DOAs [69].

Ikram and Morgan [68] provided an in-depth discussion of permutation inconsistency. They used prior knowledge of the mixing filters to derive ideal benchmarks of signal-to-interference ratio (SIR) improvements by comparing the SIR of individual sources and decide whether or not to manually rearrange the permutations based on the comparison. Based on the solution to the permutation problem suggested by Parra and Spence [19], Ikram and Morgan [68] show that as the length of the unmixing filter increases to represent real room conditions, the SIR becomes worse. A solution to overcome this drawback is proposed in the form of a multistage algorithm where the separation is carried out in multiple stages. The initial mixing stage is followed by several unmixing stages, with the length of the unmixing filter increasing at each stage, where the final values of the unmixing matrix obtained at the previous stage are used as initial values of the next stage. It was found that the majority of the permutations aligned in the early stages retained their order during later stages, and there was no overall significant increase in computational complexity as the optimum number of stages was found to be two.

In multimodal methods presented in this thesis, visual information is exploited not only to solve the permutation problem in SOS and ICA based BSS algorithms but also provide the BSS solution for moving sources.

Now SOS and ICA based BSS methods, including the limitations with existing solutions, have been reviewed, objective and subjective performance measures employed to evaluate BSS in this thesis are defined in the next section.

2.4 Performance measures used in this study

Objective and subjective performance measures for evaluation of separation are used in this work. In BSS, the objective evaluation is possible only if true system parameters are known. The performance index (PI) is a dominant performance measure in BSS. PI and its variants (also applicable to the under-determined case) measure the quality of either the estimated separating matrix or the estimated mixing matrix. The PI measure in FDCBSS provides the performance at each frequency bin level. Another classical measure is the mean square error (MSE) and can be employed to measure the quality of the separation in terms of the estimated sources. A lower value of these measures indicates good separation.

In the context of audio BSS another recently proposed objective measure [70] is signal-to-interference ratio (SIR) and in contrast to other measures, increases proportionality with quality of source separation. The main limitation of all the above objective evaluations is the requirement of the true system parameters; which fundamentally, are not available for use in real BSS applications. Therefore subjective measures are also proposed. Mean opinion score (MOS) tests for voice specified by the ITU-T recommendation P.800 provides such a subjective measure.

2.4.1 Signal-to-interference ratio

The SIR proposed by Vincent et al. [70] is considered as an objective evaluation measurement in this thesis.

$$SIR = 10\log_{10}\frac{\|s_{target}\|_2^2}{\|e_{intf}\|_2^2} \tag{2.4.1}$$

where

$$s_{target} = \langle \hat{s}_i, s_i \rangle s_i / \|s_i\|_2^2 \tag{2.4.2}$$

$$e_{intf} = \sum_{i \neq j} \langle \hat{s}_i, s_j \rangle s_i / \|s_j\|_2^2 \tag{2.4.3}$$

and s_{target} and s_{intf} represent respectively the source of interest and interference introduced by the other sources, $\langle \hat{s}_i, s_j \rangle = \sum_{t=1}^{T} \hat{s}_i(t) s_j(t)$. The above formulation can also be written as:

$$SIR = 10 \log_{10} \frac{\Sigma_i \Sigma_\omega |H_{ii}(\omega)|^2 \langle |s_i(\omega)|^2 \rangle}{\Sigma_i \Sigma_{i \neq j} \Sigma_\omega |H_{ij}(\omega)|^2 \langle |s_j(\omega)|^2 \rangle} \tag{2.4.4}$$

where H_{ii} and H_{ij} represent respectively the diagonal and off-diagonal elements of the frequency domain mixing filter, and $s_i(\omega)$ is the frequency domain representation of the source of interest. It is important to be noted that if the sources are mutually orthogonal than it leads to a large value of SIR, provided good estimates of the sources can be achieved.

2.4.2 Performance index

The PI [11, 71] as a function of the overall system matrix $\mathbf{G} = \mathbf{WH}$ can be formulated as:

$$\begin{aligned} PI(G) = & \Big[\frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{k=1}^{m} \frac{abs(G_{ik})}{max_k abs(G_{ik})} - 1 \Big) \Big] \\ & + \Big[\frac{1}{m} \sum_{k=1}^{m} \Big(\sum_{i=1}^{n} \frac{abs(G_{ik})}{max_i abs(G_{ik})} - 1 \Big) \Big] \end{aligned} \tag{2.4.5}$$

where G_{ik} is the ik-th element of $\mathbf{G}$. It is assumed that the number of sources equals to the number of mixtures. $\mathbf{G}$ can be utilized in a measure, which will be insensitive to permutations and scaling ambi-

guities. The lower bound value for PI is zero, while the upper bound value depends on the normalization factor. The lower bound value of zero means best separation. The motivation for selecting this criterion is the evaluation of performance at each frequency bin that highlights the in depth evaluation of FDCBSS algorithms. All algorithms in this thesis are also evaluated with this criterion.

2.4.3 Evaluation of permutation

Since the performance index calculated by (2.4.5) is insensitive to permutation, another criterion is introduced for the two sources case which is sensitive to permutation and shown for the real case for convenience, i.e. in the case of no permutation, $\mathbf{H} = \mathbf{W} = I$ or $\mathbf{H} = \mathbf{W} = [0, 1; 1, 0]$ then $\mathbf{G} = I$ and in the case of permutation if $\mathbf{H} = [0, 1; 1, 0]$ then $\mathbf{W} = I$ and vice versa, therefore, $\mathbf{G} = [0, 1; 1, 0]$. Hence for a permutation free FDCBSS $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$. It is highlighted that the criterion is only tested for the exactly determined case N=M=2.

2.4.4 Mean opinion score

In voice and video communication, whether the experience is a good or bad one is evaluated on the basis of perceived quality. There is a numerical method of expressing voice and video quality known as mean opinion score (MOS). Mean opinion score tests for voice is specified by the ITU-T recommendation P.800 and the listening-quality scale is shown in Table 2.1.

Table 2.1. Listening-quality scale.

Quality of the speech	Mean Opinion Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

The scores do not need to be whole numbers. Certain thresholds and limits are often expressed in decimal values. For instance, a value of 4.5 to 5.0 is referred to as a quality which provides complete satisfaction. To perform the measure a certain number of people hear the separation results. Each one of them gives a rating from within 1 to 5. Then an arithmetic mean (average) is calculated which provides Mean Opinion Score.

A key component in the proposed multimodal solution to BSS of moving sources is tracking of speakers. Now a brief overview of nonlinear Bayesian filtering (particle filtering) is presented next.

2.5 Overview of nonlinear Bayesian filtering (particle filters)

A classical problem in nonlinear filtering theory is to estimate recursively the state sequence $\{\mathbf{x}_k, k \in K\}$ of a system, from a noisy observation sequence $\{\mathbf{z}_k, k \in K\}$ made on the system.

Let $\mathbf{x}_k(t)$ evolve according to the dynamic model:

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, k) + \mathbf{v}_{k-1} \tag{2.5.1}$$

and the observation sequence $\mathbf{z}_k$ be related to the state sequence via the observation model:

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, k) + \mathbf{r}_k \tag{2.5.2}$$

where $\mathbf{f}_k(.)$ represents the state evolution function and $\mathbf{h}_k(.)$ is the observation function that represents the relationship between the state and observation sequences. The vectors $\mathbf{v}_{k-1}$ and $\mathbf{r}_k$ are the system and observation noises respectively.

The state sequence $\mathbf{x}_k$ is characterized by its probability density function estimated from a sequence of observations $\mathbf{z}_k$. In the sequential Bayesian filtering framework, the conditional density of the state sequence given the observations is propagated through prediction and update stages;

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1} \tag{2.5.3}$$

The probabilistic model of the state evolution $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is defined in (2.5.1) with noise vector $\mathbf{v}_{k-1}$. With the availability of measurement $\mathbf{z}_k$ at time k, the prior $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$ is corrected (updated) via Bayes' rule

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \tag{2.5.4}$$

with the normalization constant

$$p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})d\mathbf{x}_k \tag{2.5.5}$$

Note the likelihood function $p(\mathbf{z}_k|\mathbf{x}_k)$ is defined by the measurement model (2.5.2) with noise vector $\mathbf{r}_k$. The measurement $\mathbf{z}_k$ is used for the modification of the prior density to obtain the required posterior density of the current state and the posterior density $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ enables the

computation of an optimal estimate by recursive relations (2.5.3) and (2.5.4). It is noteworthy, if the state-space formulation is non-Gaussian then Kalman filter with its general forms is not usually applicable and particle filters can provide the state estimation [72].

The basic idea of particle filtering is to estimate recursively the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k-1})$ as in (2.5.3) by a set of samples (particles) $\{\mathbf{x}_k^n, n = 1, ..., N_p\}$, and their associated weights $\{w_k^n, n = 1, ..., N_p\}$. The posterior distribution of the state can be represented in a non-parametric way by using particles drawn from the distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ as:

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \frac{1}{N_p} \sum_{n=1}^{N_p} \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^n) \tag{2.5.6}$$

where $(.)^n$ refers to the n_{th} particle, $\delta(.)$ is the Dirac delta function, N_p is the number of particles, and have a discrete approximation of the true posterior. As N_p approaches infinity, this discrete formulation will converge to the true posterior distribution depending on the samples. However, practically this is impossible, since the posterior distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ is to be estimated and hence is unknown. In practice, the particles are sampled from a known proposal distribution $q(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ called the importance density and the concept is known as importance sampling [72] and then $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ is estimated. The distribution $p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k})$ can be formulated as:

$$p(\mathbf{x}_{0:k}|\mathbf{z}_{1:k}) \approx \sum_{n=1}^{N_p} \omega_k^n \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^n) \tag{2.5.7}$$

where

$$\omega_k^n \propto \frac{p(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k})} \tag{2.5.8}$$

and is normalized so that $\sum_i \omega_k^i = 1$.

Before state k, if $p(\mathbf{x}_{0:k-1}^n|\mathbf{z}_{1:k-1})$ can be approximated from the available samples $\mathbf{z}_{1:k-1}^n$, then with the arrival of measurement $\mathbf{z}_k$ at state k, $p(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k})$ can be approximated with a new set of samples. The importance density can be factorized as:

$$q(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k}) = q(\mathbf{x}_k^n|\mathbf{x}_{0:k-1}^n, \mathbf{z}_{1:k})q(\mathbf{x}_{0:k-1}^n|\mathbf{z}_{1:k-1}) \tag{2.5.9}$$

then new samples $\mathbf{x}_{0:k}^n$ can be obtained from $q(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k})$ by augmenting each of the old samples $\mathbf{x}_{0:k-1}^n$ from $q(\mathbf{x}_{0:k-1}^n|\mathbf{z}_{1:k-1})$ with the new state $\mathbf{x}_k^n$ from $q(\mathbf{x}_k^n|\mathbf{x}_{0:k-1}^n, \mathbf{z}_{1:k})$.

The pdf $p(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k})$ can be simplified [72] to:

$$p(\mathbf{x}_{0:k}^n|\mathbf{z}_{1:k}) \propto p(\mathbf{z}_k|\mathbf{x}_k^n)p(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n)p(\mathbf{x}_{0:k-1}^n|\mathbf{z}_{1:k-1}) \tag{2.5.10}$$

By placing (2.5.9) and (2.5.10) in (2.5.8) and if $q(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n, \mathbf{z}_k) = q(\mathbf{x}_k^n|\mathbf{x}_{0:k-1}^n, \mathbf{z}_{1:k})$ the weight update equation can be written as:

$$\omega_k^n = \omega_{k-1}^n \frac{p(\mathbf{z}_k|\mathbf{x}_k^n)p(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n)}{q(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n, \mathbf{z}_k)} \tag{2.5.11}$$

The choice of importance density function is one of the critical issues in the design of a particle filter and plays a critical role in the performance [72]. The function should have the same support as the probability distribution to be approximated, and the approximation will be better if the importance function is closer to the distribution. The above mentioned assumption $q(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n, \mathbf{z}_k) = q(\mathbf{x}_k^n|\mathbf{x}_{0:k-1}^n, \mathbf{z}_{1:k})$ means that the importance density depends only on the previous state $\mathbf{x}_{k-1}^n$ and current measurement $\mathbf{z}_k$, and the path $\mathbf{x}_{0:k-1}^n$ and history of obser-

vations $\mathbf{z}_{1:k-1}$ will be discarded. The most popular choice for the prior importance function is given by

$$q(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n, \mathbf{z}_k) = p(\mathbf{x}_k^n|\mathbf{x}_{k-1}^n) \tag{2.5.12}$$

and this particular importance density is applied at every time index to simplify the weight update equation to:

$$\omega_k^n \propto \omega_{k-1}^n p(\mathbf{z}_k|\mathbf{x}_k^n) \tag{2.5.13}$$

The importance sampling weight indicates the level of importance of the corresponding particle. In the above mentioned sequential importance sampling algorithm, after a few iterations, all but one particle will have very small weight, this is known as the degeneracy phenomenon. With a relatively small weight mean the particle is ineffective in calculation of the posterior distribution. To overcome the degeneracy, residual resampling can be used for example, which is a scheme that eliminates the particles with small weights and replicates those with large weights accordingly [72].

The implementation of the particle filter can then be divided into two steps:

1. Sampling step: N particles are sampled from the proposal density formulated by (2.5.12) according to (2.5.1).

2. Computing the particle weights according to (2.5.13), and resampling the particles if necessary.

Based on the weights the conditional mean of $\mathbf{x}_k$ can then be calculated.

2.6 Summary

In this chapter an overview of the main BSS techniques has been provided. Both the generic mixing and separating models have been provided to demonstrate instantaneous and convolutive BSS. Two main statistical techniques of BSS i.e. SOS and ICA has been discussed in Section 2.2. In Subsection 2.2.1 it has been highlighted that a SOS based bench marked algorithm [19] which exploits the non-stationarity of speech can be improved by adopting a multimodal approach. This is a subject of the next chapter. In Subsection 2.2.2 an HOS based technique, i.e. ICA has been discussed in detail and a preprocessing step for ICA has also formulated. Advantages of FDCBSS were mentioned in Section 2.3 and indeterminacies of SOS/ICA were also discussed in detail. The solution to the permutation ambiguity with improved convergence for ICA based FDCBSS, by using a multimodal method, was presented in Chapter 4. In Section 2.3 another limitation of the existing BSS techniques (SOS/ICA) i.e. data length requirement, to extract the statistical information from observed mixtures was highlighted. This limits SOS/ICA to solve the BSS for moving sources (real *cocktail party problem*). A multimodal method which overcomes the data length limitation of SOS/ICA and provides an acceptable level of separation for moving sources is presented in Chapter 6. A key component of this method is a 3-D tracker and therefore an overview of nonlinear Bayesian filtering was provided in Section 2.5.

Chapter 3

A GEOMETRICALLY CONSTRAINED MULTIMODAL METHOD FOR FREQUENCY DOMAIN CONVOLUTIVE BLIND SOURCE SEPARATION

In this chapter, a novel constrained multimodal method for convolutive blind source separation is presented which incorporates video information related to geometrical position of both the speakers and the microphones, and the directionality of the speakers, into the separation algorithm. The separation is performed in the frequency domain and the constraints are incorporated through a penalty function-based formulation. The separation results show a considerable improvement over traditional frequency domain convolutive BSS systems such as that developed by Parra and Spence. Importantly, the inherent permutation problem in the frequency domain BSS is essentially solved.

3.1 Introduction

Many methods have been proposed for convolutive blind source separation (CBSS) [10–26]. As mentioned in Chapter 2, in frequency domain CBSS (FDCBSS) the permutation problem [13,67–69,73] increases exponentially and is therefore more severe and destructive than for time domain schemes. In such systems there are generally no priori assumptions on the source statistics or the mixing system. On the other hand, in a multimodal method the video system can capture the positions of the speakers and the directions they face [41]. The video information can thereby help to estimate the mixing matrix more accurately and ultimately increase the separation performance. Following this idea, the objective of this work is to use efficiently such information in the enhancement of the separation results.

The solution to the permutation problem in [19] is proposed by imposing a smoothness constraint on the unmixing filters in the frequency domain and is achieved by limiting the filter length in the time domain which is itself a constraint. This method has been found only to have limited success [19]; therefore, in order to solve the above mentioned problem and also improve the separation, spatial information is used in this work, indicating the positions and directions of the sources using the "data" acquired simultaneously by a number of video cameras. The comparison between the original Parra and Spence [19], Wang et al. algorithm [74], and the proposed multimodal constrained FDCBSS algorithms will be presented at the end of the chapter. In the proposed method the constraints are incorporated through a penalty function-based formulation.

The rest of the chapter is organized as: in Section 3.2 the geometri-

cal model is presented, in Sections 3.3 & 3.4 the constrained FDCBSS method is formulated, in Section 3.5 experimental results are discussed, and in Section 3.6 the chapter is summarized.

3.2 The geometrical model

The geometrical model is based on visual information captured by the video cameras. The use of visual speech information is also popular in current research into speech recognition and speech enhancement methods [27, 29]. In [27] it has been shown that a speaker's face in a noisy environment greatly improves the intelligibility of that person's voice. Visual cues, for example, are used to determine who is being addressed. Colin Cherry's observed that the human approach to solve the *cocktail party problem* exploits visual cues [1, 2] and Simon Haykin in [3] also highlighted the importance of visual information to solve the cocktail party problem. The following procedure is used to find the visual information which will be used in the geometrical model explained in the sequel.

3.2.1 Video camera calibration

Two colour video cameras are used to determine the approximate positions of the speakers. Both static video cameras were calibrated off line by the Tsai calibration (non-coplanar) technique [75]. The method for camera calibration recovers the interior orientation (principle distance f), the exterior orientation (relationship between a scene-centered coordinates system and a camera-centered coordinate system, the transformation from scene to camera consists of a rotation matrix R and translation vector $\mathbf{t}$), the power series coefficients for lenes distortion

κ, and image scale factor $\wp$ (uncertainty scale factor, due to TV camera scanning and acquisition timing error, for more detail see [75]). The transformation from real world to image coordinates used in the calibration process is presented in the following section.

Transformation of coordinates

Real world coordinates $\mathbf{z} = [z_x, z_y, z_z]^T$ can be projected into image coordinates $\iota \in \mathbb{R}^2$. Initially, the corresponding vector in camera coordinates $\mathbf{c} = [c_x, c_y, c_z]^T$ is calculated as:

$$\mathbf{c} = \mathbf{R}\mathbf{z} + \mathbf{t} \tag{3.2.1}$$

where $\mathbf{R}$ is the rotation matrix and $\mathbf{t} = [t_x, t_y, t_z]^T$ is a translation vector.

Transformation from 3-D camera coordinates to ideal (undistorted) image coordinates $\mathbf{u} = [u_x, u_y]^T$ using perspective projection with pinhole geometry is achieved with

$$\begin{aligned} u_x &= \frac{f}{c_z} c_x \\ u_y &= \frac{f}{c_z} c_y \end{aligned} \tag{3.2.2}$$

where f is the effective focal length in the appropriate units [75].

Real optical systems suffer from a number of inevitable geometric distortions. In optical systems made of spherical surfaces, with centers along the optical axis, a geometric distortion occurs in the radial direction. Electro-optical systems typically have larger distortions than optical systems made of glass. They also suffer from tangential distortion, which is at a right angle to the vector from the center of the image.

As for radial distortion, tangential distortion grows with distance from the center of distortion. For industrial machine vision application, only radial distortion needs to be considered [75].

The radial lens distortion terms D_x and D_y formulate the distorted coordinates in the following manner

$$\begin{aligned} x_d &= u_x - D_x \\ y_d &= u_y - D_y \end{aligned} \tag{3.2.3}$$

where x_d and y_d are the distorted coordinates on the image plane, and termed as the true image coordinates in [75].
Where

$$\begin{aligned} D_x &= x_d(\kappa_1 r^2 + \kappa_2 r^4 + ...) \\ D_y &= y_d(\kappa_1 r^2 + \kappa_2 r^4 + ...) \\ r &= \sqrt{x_d^2 + y_d^2} \end{aligned} \tag{3.2.4}$$

as shown in [75].

In practice only one term is needed in the expansions of D_x and D_y, therefore only coefficient κ_1 must be found as part of the calibration process, any more modeling would cause instability [75].

Finally, the 2-D image coordinates $\iota = [\iota_x, \iota_y]^T$ take the form:

$$\begin{aligned} \iota_x &= x_d \wp / \delta_{px} + x_0 \\ \iota_y &= x_d / \delta_{py} + y_0 \end{aligned} \tag{3.2.5}$$

where x_0 and y_0 define the center of the video frame and δ_{px}, and δ_{py} define the physical size of the sensor element of the cameras.

The extracted face of each speaker in the images of both cameras is transformed to find the real world position of each speaker. The face extraction in an image is explained in the next subsection.

3.2.2 Face extraction

The face of each speaker is extracted in the images of both cameras to find the position of each speaker. In each image frame, the face is extracted on the basis of a skin pixel model and a face model, briefly the procedure is (for further detail see [76] and [77])

Off line formulation of the skin pixel model:

- A training set of skin regions for different people with varying skin tones is obtained by manual extraction of the facial regions within a number of measured frames.
- Each skin region is converted from the RGB colour space into the normalized r-g colour space $r, g = R, G/(R + G + B)$, and a corresponding pixel in r-g colour space, i.e. a two-dimensional vector, is denoted by D.
- A two-dimensional histogram of all the D vectors from the training set of skin regions is produced. Parameters a_j are calculated which correspond to the relative occurrence of each vector D_j within the training set.
- Vectors which have a value of a_j less than some threshold are considered to correspond to noisy pixels. Such pixels are removed from the training set.

- The remaining unique vectors $D_1, ..., D_l$ with their respective a_j where $j = 1, ..., l$ are next used to formulate a skin pixel model.

- Skin pixel model: A skin pixel model $\Phi = (D; \Psi, \Lambda)$ is defined as

$$\Phi(D) = [D - \Psi]^T \Lambda^{-1} [D - \Psi] \tag{3.2.6}$$

where $(.)^T$ denotes vector transpose and the parameters Ψ and Λ can be calculated as

$$\Psi = \frac{1}{l} \sum_{j=1}^{l} D_j \tag{3.2.7}$$

$$\Lambda = \frac{1}{l} \sum_{j=1}^{l} a_j (D_j - \mu)(D_j - \mu)^T \tag{3.2.8}$$

$$\mu = \frac{1}{l} \sum_{j=1}^{l} a_j D_j \tag{3.2.9}$$

- Given threshold θ_{thresh} and r-g vector D of a pixel, D is classified as skin chrominance if $\Phi(D) < \theta_{thresh}$ and as non-skin chrominance otherwise.

Face extraction

- Given a measured frame converted to normalized r-g colour space all candidate face pixels are extracted according to the above skin pixel model on the basis of a threshold θ_{thresh}.

- Each significant cluster of candidate face pixels is cross-correlated with a standard face (which is an averaging of front views of 16 male and female faces wearing no glasses and having no facial hair), appropriate processing is used to align the relative sizes of the two images.

- The cross-correlation value between the standard face template and every skin region is calculated. The region which yields maximum cross-correlation is chosen as the desired face.

The center of the resultant face region is determined as the approximate position of the lips of the speaker in image coordinates ι. The image coordinates of each speaker with the calibration parameters f, R, $\mathbf{t}$, κ, $\wp$ of each camera are passed to the next section to calculate the position of the speakers in 3-D world coordinates.

This relatively simple approach to face extraction is replaced by the state-of-the-art Viola-Jones [78] scheme in the final contribution contained in Chapter 6. In this chapter further details of this scheme are provided.

3.2.3 Source position in the real world

With the help of the above calibration parameters the approximated 2-D image information of the same speaker in two different camera views $\iota = [\iota_x, \iota_y]^T$ is transformed to true (distorted) sensor plane coordinates $\mathbf{d} = [x_d, x_d]^T$ as:

$$\mathbf{d} = [\frac{\delta_{px}(\iota_x - x_o)}{\wp}, \delta_{py}(\iota_y - y_o)]^T \tag{3.2.10}$$

The ideal (undistorted) sensor plane coordinates $\mathbf{u} = [u_x, u_y]^T$ are calculated as:

$$\mathbf{u} = \mathbf{d}(1 + \kappa_1|\mathbf{d}|^2) \tag{3.2.11}$$

In 3-D space each point in each camera frame defines a ray. Intersection of both rays is found by optimization methods. By having undistorted sensor plane coordinates $\mathbf{u}$ of each speaker in both cam-

eras with calibration parameters of each camera, the positions of the speakers $\mathbf{z}$ are transformed into 3-D real world coordinates (for detail see [79]).

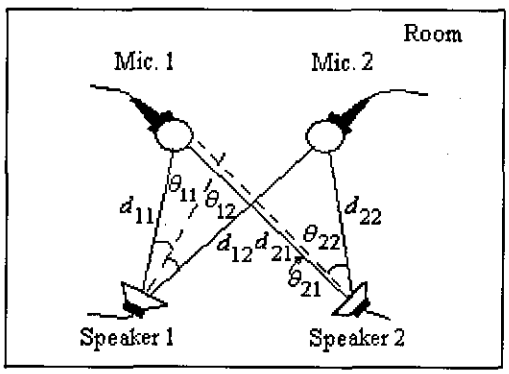

Figure 3.1. A two-speaker two-microphone setup for recording within a reverberant (room) environment; only distances and angles between sources and microphones are shown.

On the basis of the above calculated 3-D positions of the speakers and the microphones, the distances between the i-th microphone and the j-th speaker d_{ij}, and also the associated signal propagation times τ_{ij}, can be calculated (see Figure 3.1 for a simple two-speaker two-microphone case). Accordingly, in a homogenous medium such as air, the attenuation is related to the distances via

$$\alpha_{ij} = \frac{\kappa}{d_{ij}^2} \tag{3.2.12}$$

where κ is a constant representing the attenuation per unit length in a homogenous medium. Similarly, τ_{ij} in terms of the number of samples, is proportional to the sampling frequency f_s, sound velocity C in air, and the distance d_{ij} as:

$$\tau_{ij} = \frac{f_s}{C} d_{ij} \tag{3.2.13}$$

which is independent of the directionality. Both f_s and C are consid-

ered constant within each observation interval for a block-based BSS system. However, in practical situations the speakers' directions introduce another variable into the attenuation measurement. In the case of electronic loudspeakers (not humans) the directionality pattern depends on the type of loudspeaker. Here, this pattern is approximated as $\cos(\theta_{ij}/r)$ where $r > 2$, and has a smaller value for highly directional speakers and vice versa (an accurate profile can be easily measured using a sound pressure level (SPL) meter). Therefore, the attenuation parameters become

$$\alpha_{ij} = \frac{\kappa}{d_{ij}^2} \cos(\theta_{ij}/r) \tag{3.2.14}$$

If, for simplicity, only the direct path is considered the mixing filter is expected to have the form :

$$\hat{H}(t) = \begin{bmatrix} \alpha_{11}\delta(t-\tau_{11}) & \alpha_{12}\delta(t-\tau_{12}) \\ \alpha_{21}\delta(t-\tau_{21}) & \alpha_{22}\delta(t-\tau_{22}) \end{bmatrix} \tag{3.2.15}$$

where $\hat{(\cdot)}$ denotes that this is an estimate/approximation. In the frequency domain and z-domain the above filter has the forms

$$\hat{\mathbf{H}}(\omega) = \begin{bmatrix} \alpha_{11}e^{-j\omega\tau_{11}} & \alpha_{12}e^{-j\omega\tau_{12}} \\ \alpha_{21}e^{-j\omega\tau_{21}} & \alpha_{22}e^{-j\omega\tau_{22}} \end{bmatrix} \tag{3.2.16}$$

$$\hat{\mathbf{H}}(z) = \begin{bmatrix} \alpha_{11}z^{-\tau_{11}} & \alpha_{12}z^{-\tau_{12}} \\ \alpha_{21}z^{-\tau_{21}} & \alpha_{22}z^{-\tau_{22}} \end{bmatrix} \tag{3.2.17}$$

Although in reality the actual mixing matrix includes the reverberation terms related to the reflection of sounds by the obstacles and

walls, in such a room environment it will generally contain the direct path components as in the above equations. Therefore, $\hat{\mathbf{H}}(\omega)$ can be considered as a crude, albeit biased, estimate of the frequency domain mixing filter matrix.

3.3 The constrained problem

In order to improve separation performance the above visual information is integrated into a BSS algorithm [19] in the form of a constraint, which represents the squared Frobenius norm distance between the unmixing filter $\mathbf{W}(\omega)$ and the permuted mixing filter $\hat{\mathbf{H}}(\omega)$, i.e.

$$J_c = \|\mathbf{W}(\omega) - \mathbf{P}(\omega)\hat{\mathbf{H}}^{-1}(\omega)\|_F^2 = \|vec(\mathbf{W}(\omega) - \mathbf{P}(\omega)\hat{\mathbf{H}}^{-1}(\omega))\|_2^2 \quad (3.3.1)$$

where $\|.\|_2^2$ represent respectively, the squared Euclidean norm, $vec(.)$ converts a matrix argument column-wise into a column vector, and $\mathbf{P}(\omega)$ is the permutation matrix. Ultimately, the cost function J_c has to be minimized with respect to both $\mathbf{W}(\omega)$ and $\mathbf{P}(\omega)$.

3.4 The overall constrained BSS

In order to achieve the above goal, it is required to minimize jointly J_m and J_c with respect to $\mathbf{W}(\omega)$, and also minimise J_c with respect to the permutation matrix $\mathbf{P}(\omega)$. The constrained optimization problem can be changed to an unconstrained one using a penalty function as in [74]. In this case

$$J(\mathbf{W}(\omega)) = J_m(\mathbf{W}(\omega)) + \lambda J_c(\mathbf{W}(\omega)) \quad (3.4.1)$$

where λ is a weighting parameter. $\mathbf{W}(\omega)$ and $\mathbf{P}(\omega)$ are then found by minimizing the gradients of J and J_c respectively with respect to $\mathbf{W}(\omega)$ and $\mathbf{P}(\omega)$, i.e.

$$\mathbf{W}_{opt}(\omega) = \arg\min_{\mathbf{W}}\{J_m(\mathbf{W}(\omega)) + \lambda J_c(\mathbf{W}(\omega))\} \tag{3.4.2}$$

and

$$\mathbf{P}_{opt}(\omega) = \arg\min_{\mathbf{P}}\{J_c(\mathbf{W}(\omega))\} \tag{3.4.3}$$

Therefore, at each frequency bin ω the estimated sources will be aligned with the input source signals; as one of the major advantages of this algorithm there will not generally remain any permutation problem. Consequently, the update equations are obtained as:

$$\mathbf{W}_{j+1}(\omega) = \mathbf{W}_j(\omega) - \mu_j \nabla_{\mathbf{W}}(J(\mathbf{W}_j(\omega))) \tag{3.4.4}$$

$$\mathbf{P}_{j+1}(\omega) = \mathbf{P}_j(\omega) - \eta_j \nabla_{\mathbf{P}}(J_c(\mathbf{W}_j(\omega))) \tag{3.4.5}$$

where j is the iteration index, μ and η are the learning rates, and

$$\begin{aligned}\nabla_{\mathbf{W}^*}(J(\mathbf{W})) &= \nabla_{\mathbf{W}^*}(J_m(\mathbf{W})) + \lambda\nabla_{\mathbf{W}^*}(J_c(\mathbf{W})) \\ &= 2\sum_{k=1}^{K} E(\omega,k)\mathbf{W}(\omega)R_x(\omega,k) \\ &\quad +2\lambda[\mathbf{W}(\omega) - \mathbf{P}(\omega)\hat{\mathbf{H}}^{-1}(\omega)]\end{aligned} \tag{3.4.6}$$

and

$$\nabla_{\mathbf{P}}(J_c(\mathbf{W})) = -2\hat{\mathbf{H}}^{-1}(\omega)[\mathbf{W}(\omega) - \mathbf{P}(\omega)\hat{\mathbf{H}}^{-1}(\omega)] \tag{3.4.7}$$

Before starting the update process $\hat{\mathbf{H}}^{-1}(\omega)$ is normalized once using $\hat{\mathbf{H}}^{-1}(\omega) \leftarrow \tilde{\mathbf{H}}^{-1}(\omega)/\|\hat{\mathbf{H}}^{-1}(\omega)\|_F$ where $\|.\|_F$ denotes the Frobenius norm and after each iteration $\mathbf{W}(\omega)$ is also normalized. In the case of fractional filters where the distances between the speakers and the microphones are not integer multiples of the sampling interval then [80, 81] can be used to estimate firstly the fractional delay and then perform the BSS process.

Summary Table: Implementation steps for the proposed Multimodal FDCBSS Method

1. *Initialize parameters,* N, M, k, T, Q, λ, γ, $\mathbf{W_1}(\omega)$, f_s, C, r, κ, *iter. count.*

2. *Read input mixtures, i.e., time samples* $\mathbf{x}(t_k)$.

3. *Calculate the distances* d_{ij} *and angle of arrivals* θ_{ij} *between the speakers and the microphones on the basis of video information.*

4. *Calculate the propagation time* τ_{ij} *using (3.2.13) and attenuation* α_{ij} *using (3.2.14), on the basis of* d_{ij}, r, κ, θ_{ij}, f_s, C.

5. *Find the estimate of mixing filter* $\hat{\mathbf{H}}(\omega)$ *using (3.2.3).*

6. *Normalized* $\hat{\mathbf{H}}(\omega) \leftarrow \hat{\mathbf{H}}(\omega)/\|\hat{\mathbf{H}}(\omega)\|_F$

7. *Calculate the cross-power spectrum matrix:*

   - *Convert* $\mathbf{x}(t_k)$ *to* $\mathbf{x}(\omega, t_k)$.
   - *Calculate* $\mathbf{R}_{\mathbf{x}}(\omega, t_k)$ *using (2.2.1).*

8. *Calculate the cost function and update gradient:*

- *FOR* $i = 1$ *to iter. count.*
  - *Update* μ_i *and* η_i *using (3.5).*
  - *Update* $\mathbf{W}_{i+1}(\omega) = \mathbf{W}_i(\omega) - \mu_i \nabla_{\mathbf{W}}(J(\mathbf{W}_i(\omega)))$ *using (3.4.6) and* $\mathbf{P}_{i+1}(\omega) = \mathbf{P}_i(\omega) - \eta_i \nabla_{\mathbf{P}}(J_c(\mathbf{W}_i(\omega)))$ *using (3.4.7).*
  - *Update* $J_i(\mathbf{W}(\omega)) = J_m(\mathbf{W}(\omega)) + \lambda J_c(\mathbf{W}(\omega))$ *using (2.2.2) and (3.3.1).*
  - *if* $(J_i(\mathbf{W}(\omega)) > J_{i-1}(\mathbf{W}(\omega)))$ *break.*
- *END FOR.*

9. *Calculate* $\mathbf{y}(\omega, t_k)$ *according to (2.1.9).*
10. *Reconstruct the time domain signals* $\mathbf{y}(t_k) = IDFT(\mathbf{Y}(\omega, t_k))$.
11. *Calculate Performance Index (PI) (2.4.2) and Evaluate Permutation* $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$.
12. *Calculate the Signal-to-Interference Ratio (SIR) (2.4.4).*
13. *End.*

3.5 Experimental results and discussions

The objective of this section is to evaluate the proposed method in terms of performance and solution to permutation. The main reason for the work in frequency domain blind source separation is the convolutive nature of the real world problems. In the proposed method visual information, presented in the geometric model, is used which is limited to the direct path between the sensors and the sources. Therefore three experiments are performed with different positions of the microphones and the sensors to evaluate the effect of visual modality in the proposed method. The proposed method is not only evaluated but also the comparison with two SOS based BSS algorithms [19] and [66] is also provided. The simulations are performed on speech signals generated for a room geometry as illustrated in Figure 3.2. In this chapter, initially, the algorithms are objectively evaluated based on the generated room impulse responses and the observed mixture signals are obtained by convolving the source signals with these room impulse responses. Finally, the performance on the real room recordings of the same room geometry are confirmed subjectively by listening tests and mean opinion score (MOS) is provided at the end of the section.

The important variables were selected as: FFT length $T = 1024$ and filter length $Q = 512$ (half of T), $r = 4$, the sampling frequency was 8kHz, size of the room was 5x5x5 m^3, and the room impulse duration was 130ms, $\lambda = 0.15$ and the learning rates μ and η were gradually decreased with respect to the iteration index j

$$\mu_j = \eta_j = \gamma \frac{0.02}{1 - (0.98)^j} \tag{3.5.1}$$

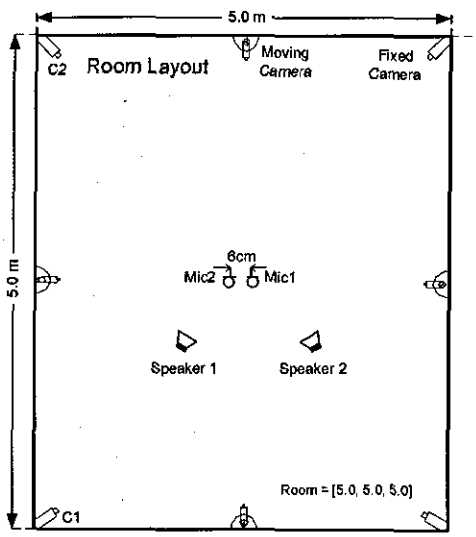

Figure 3.2. Plan view of a two-speaker two-microphone layout for recordings within a reverberant (room) environment. Height of the room = 5.0m. The objective evaluation of BSS requires the mixing filter therefore the audio signals are convolved with real room impulse responses recorded in the illustrated room geometry and real recordings of the same room geometry were also separated and evaluated subjectively by listening tests. Room impulse response length is 130 ms.

where γ is a constant with $\gamma = 0.01$.

In the first experiment the positions of the sensors and speakers are Mic1 $= [2.47, 2.50, 1.5]$, Mic2 $= [2.53, 2.50, 1.5]$, Speaker1 $= [1.0, 2.0, 1.5]$ and Speaker2 $= [3.5, 2.0, 1.5]$. The resulting performance indices are shown in Figure 3.3. At higher frequency bins there is less energy in the mixtures; therefore performance in those bins deteriorates. The results of calculating the criterion $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ for evaluation of permutation for the first experiment are shown in Figure 3.4.

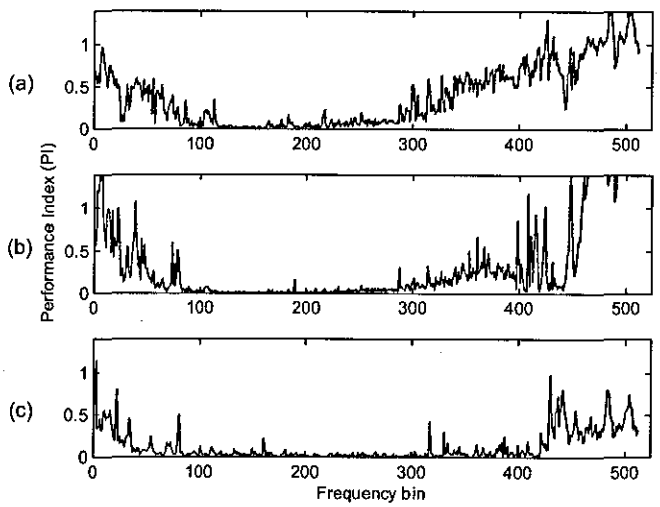

Figure 3.3. Performance index at each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. A lower PI refers to a superior method.

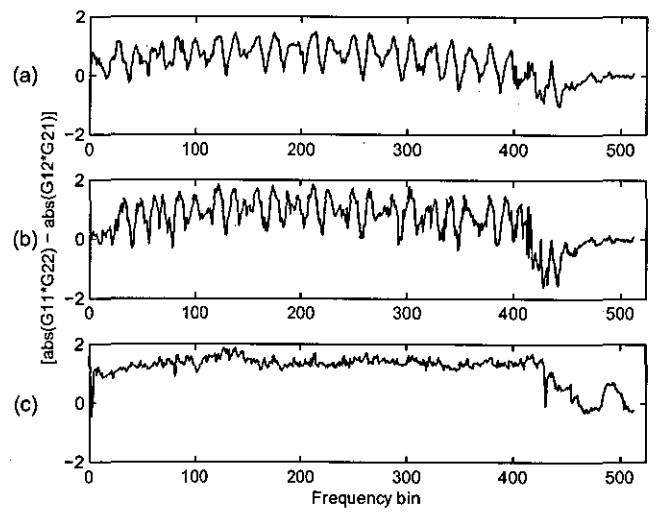

Figure 3.4. Evaluation of permutation in each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

In the second experiment only the positions of the speakers are changed, the Euclidean distance between the speakers and the center of the microphones is reduced. Mic1 = $[2.47, 2.50, 1.5]$, Mic2 = $[2.53, 2.50, 1.5]$, Speaker1 = $[2.00, 1.20, 1.5]$ and Speaker2 = $[3.25, 1.20, 1.5]$. The resulting performance indices are shown in Figure 3.5, and for this experiment the permutation is also evaluated on the basis of the criterion mentioned above and the results are shown in Figure 3.6.

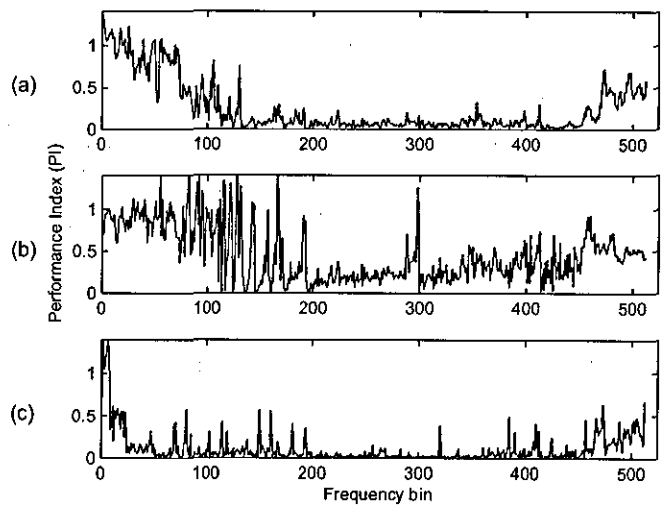

Figure 3.5. Performance index at each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. A lower PI refers to a superior method.

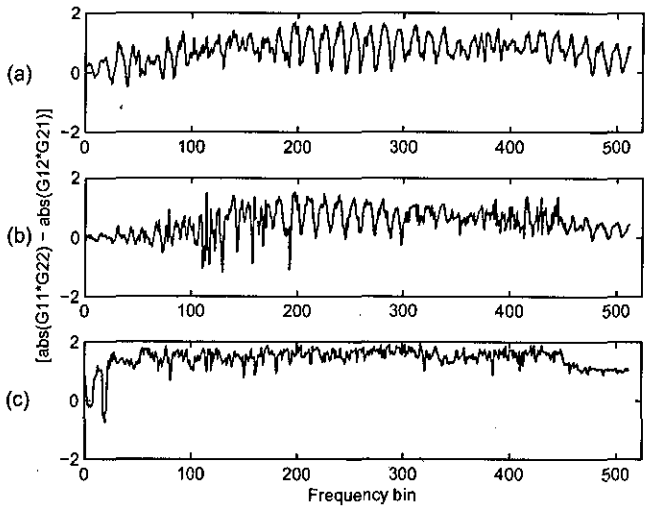

Figure 3.6. Evaluation of permutation in each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

In the third and last experiment the distance between the microphones is reduced to 4cm. Mic1 = $[2.48, 2.50, 1.5]$, Mic2 = $[2.52, 2.50, 1.5]$, Speaker1 = $[2.00, 1.20, 1.5]$ and Speaker2 = $[3.25, 1.20, 1.5]$. The resulting performance indices are shown in Figure 3.7, and the evaluation for permutation on the basis of the criterion mentioned above is shown in Figure 3.8. Justification for the experiment is that the distance between adjacent microphones should be as large as possible so as not to yield poor performance at low frequencies and should also be smaller than half of the minimum wavelength in order to avoid the spatial aliasing effect. The distance between adjacent microphones should be $round(8.5/2) = 4$cm, when the sampling frequency is 8kHz. In Figure 3.7 the effect of inter-element spacing is suppressed by the reduced energy in the recorded signals at higher frequency bins.

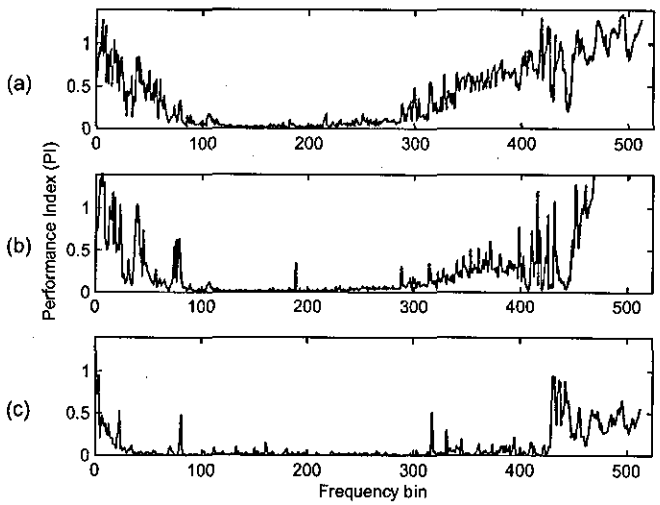

Figure 3.7. Performance index at each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. A lower PI refers to a superior method.

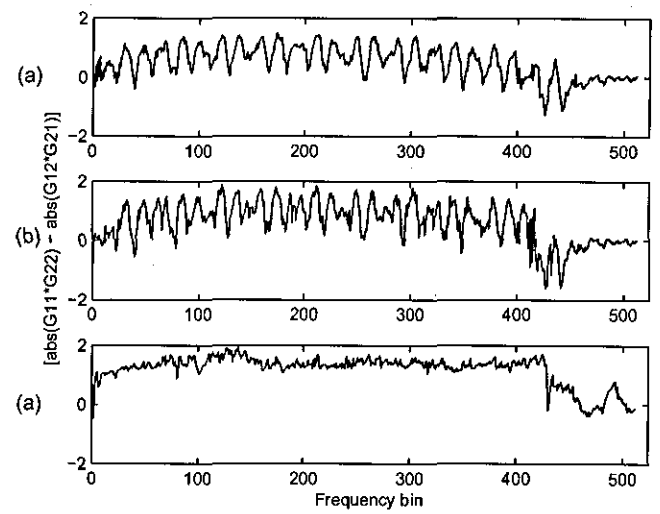

Figure 3.8. Evaluation of permutation in each frequency bin for the observed mixture signals obtained by convolving the source signals with the room impulse responses. (a) Parra and Spence algorithm [19], (b) Wang et al. algorithm [74], and (c) multimodal FDCBSS algorithm [66]. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Figures 3.3(c), 3.5(c) & 3.7(c) show good performance i.e. close to zero across the majority of the frequency bins since this is due to the multimodal method. Figures 3.4(c), 3.6(c) & 3.8(c) show that the multimodal FDCBSS method mitigates permutation. Actually, in unimodal BSS no prior assumptions are typically made on the source statistics or the mixing system. On the other hand, in a multimodal method a video system can capture the approximate positions of the speakers and the directions. Such video information helps to estimate the unmixing matrices more accurately and ultimately increases the separation performance. It is highlighted that the convergence time of the proposed method and [74] is higher than [19], typically around 20-30% larger.

In all the above simulation results it is obvious that by changing the inter-element spacing between sensors and the speakers can cause ill-conditioning in the mixing matrix; therefore for a certain frequency bin, the BSS performance is poor in one experiment and good in another experiment, and is more obvious in the simulation results based on [19] and [74]. The proposed multimodal method provides better performance than [19] and [74] but the effect is still obvious in some frequency bins. Therefore an over-determined BSS method is proposed in Chapter 5 to enhance the separation.

Finally, the SIR (2.4.4) is calculated and results are shown in Table-3.1, and the results have been confirmed subjectively by listening tests. Six people participated in the listening tests and MOS results are shown in Table-3.2.

Table 3.1. Comparison of SIR-Improvement between algorithms for different sets of mixtures.

Algorithms	SIR-Improvement/dB
Parra's Method [19]	6.8
Wang et al. Method [74]	9.2
Multimodal FDCBSS Method	9.8

Table 3.2. Subjective evaluation: Mean Opinion Score (MOS) for separation of real room recordings.

Algorithms	Mean Opinion Score
Parra's Method [19]	3.6
Wang et al. Method [74]	4.0
Multimodal FDCBSS Method	4.1

3.6 Summary

In this chapter a key SOS based FDCBSS algorithm [19] has been modified by accommodating geometrical information about the sources in a multimodal BSS method. The location and direction information has been obtained using a number of cameras and the geometric model. Camera calibration, face detection, and 2-D image coordinates to 3-D real world coordinates transformation have been presented in Sections 3.2 & 3.4. The constrained problem has been partially changed to an unconstrained problem using Lagrange multipliers in Section 3.3. The results in Section 3.5 show that the modified CBSS system enhances the performance of the traditional FDCBSS system both objectively and subjectively. The outcome of this multimodal method paves the way for establishing a multimodal audio-video system for separation of speech and music signals.

Chapter 4

A MULTIMODAL METHOD FOR FREQUENCY DOMAIN INDEPENDENT COMPONENT ANALYSIS AND SEPARATION OF STATIONARY AND STEP-WISE MOVING SOURCES

In this chapter, a novel multimodal method for independent component analysis (ICA) of complex valued frequency domain signals is presented which utilizes video information to provide geometrical description of both the speakers and the microphones. This geometric information is incorporated into the initialization of the complex ICA algorithm for

each frequency bin. The algorithm is also updated for stepwise moving sources which exploits the geometric information and permutation free unmixing matrix of the mixtures of the current block, to initialize FastICA for separation of stepwise moving sources. The advantages of this work are that no extra processing is required to solve the permutation problem separately in the frequency domain BSS nor is postprocessing required. The separation results show significant improvement in the performance of the resulting intelligently initialized FastICA method over conventional FastICA, and also confirm that the proposed algorithm is robust and potentially suitable for real time implementation for static and stepwise moving sources. It is also highlighted that certain fixed point algorithms proposed by Hyvärinen et al., or their constrained versions, are not valid for complex valued signals.

4.1 Introduction

BSS of moving sources is a more challenging aspect of solving the cocktail party problem [1] and only a few papers have been presented in this area [82–84]. In [82] BSS for moving sources is performed by using frequency domain ICA as a blockwise batch algorithm in the first stage, and the separated signals are refined by post processing using crosstalk component estimation and non stationary spectral subtraction in the second stage. The permutation problem is solved by an algorithm based on analytical calculation of null directions. Thus in [82] source separation is proposed in four stages 1) ICA based block wise batch algorithm, 2) analytical calculation to solve the permutation, 3) non stationary spectral subtraction, and 4) cross talk component estimation. Fundamentally, the algorithm is not applicable to moving

sources because ICA is based on fourth order statistics and requires sufficient data length to obtain accurate HOS estimates, which means the mixing filter should be approximately fixed in the batch of the data, therefore it is only applicable to stepwise moving sources. The proposed method is also applicable to stepwise moving sources and has less computational complexity and better performance. The proposed method does not require a separate process to solve the permutation problem or postprocessing for speech enhancement. The proposed method is justified theoretically as well as practically in this chapter.

The major problem for the moving sources case is the time variant mixing model which becomes more complicated when the environment is reverberant. The established unimodal approaches are not suitable to solve the problem, therefore a more cognitive approach is required [3] and therefore a multimodal method is exploited in which FastICA is initialized in an intelligent way. The permutation problem inherent to frequency domain blind source separation (FDCBSS) is automatically solved. Video information can help to estimate the unmixing matrices more accurately. Following this idea, the objective of this chapter is to use efficiently such video information to mitigate the permutation problem and ultimately increase the separation performance. The scaling problem in CBSS is easily solved by matrix normalization [66, 85]. In this method BSS become semiblind by initially exploiting the above mentioned prior geometrical information in initialization of FastICA to make the process robust and permutation free, and later on with the help of the unmixing matrix of the previous time block and using the whitening matrix of the current time block again the FastICA is initialized in an intelligent way to enhance the convergence properties of

BSS so that it is potentially suitable for real-time implementation for stepwise moving sources. As such, the separation matrix is updated for each time block $B_j = \{t : (j-1)T_b \leq t < jT_b)\}$, where T_b is the time block size, and j represent the block index ($j \geq 1$). This intelligent initialization based FastICA algorithm is more suitable when T_b is small, i.e. reduced change in the unmixing matrix will provide a less biased estimate for initialization, however reduction in T_b is limited by the data length required for FastICA to converge. Therefore this method is more robust in the case of stepwise moving sources because the mixing filter will be fixed at each step and in this chapter the performance is presented when sources are moving stepwise in a teleconference-like scenario.

In the following section a fast fixed-point algorithm for complex valued signals, for which the choice of contrast function is carefully motivated is discussed. In Section 4.3 the use of spatial information indicating the positions and directions of the sources using data acquired by a number of video cameras is examined and the proposed intelligent initialization based FastICA algorithm is discussed. In Section 4.4 the simulation results, for stationary and stepwise moving sources, of real world data confirm the usefulness of the algorithm. Finally, in Section 4.5 the chapter is summarized.

4.2 A fast fixed-point algorithm for ICA of complex valued signals

Recently, ICA has become one of the central tools for BSS [6,11,86–92] (in Chapter 2 ICA is discussed in detail). Actually, in ICA a set of estimated source signals $\hat{s}(\omega)$ in (2.1.9) are retrieved from their mixtures based on the assumption of their mutual statistical independence [9,15]. It is important to mention that ICA algorithms are commonly limited to separate instantaneous linear mixtures but most real world problems have a convolutive nature. This limitation can be avoided in FDCBSS because a T-point windowed discrete Fourier transformation (DFT) converts the real value, broadband, time domain signal into a set of complex valued, narrowband, frequency domain signals. Therefore at each frequency bin ICA is applicable. Hyvärinen and Oja [6,93] presented a fast fixed point algorithm (FastICA) for the separation of linearly mixed independent source signals. Unfortunately, these algorithms are not suitable for complex valued signals.

Algorithms for independent component analysis of complex valued signals are also presented in [94–96]; the first two algorithms are computationally more intensive than the last, and no proofs of consistency are given in either of the references. The use of algorithm [96] in this work is due to four main reasons: its suitability for complex signals, the proof of the local consistency of the estimator, more robustness against outliers and capability of deflationary separation of the independent component signals. In deflationary separation the components tend to separate in the order of decreasing non-Gaussianity. In exactly determined separation, however, it is generally better to use a symmetric scheme rather than a deflationary approach. In [96] the basic concept of complex random variables is also provided and the fixed point algo-

rithm for one unit is derived, and for ease of derivation the algorithm updates the real and imaginary parts of $\mathbf{w}(\omega)$ separately. The vector $\mathbf{w}(\omega)$ represents one row of $\mathbf{W}(\omega)$ used to extract a single source.

According to the Lagrange conditions [97], the optima of $E\{G(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2)\}$ under the constraint $E\{|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2\} = \|\mathbf{w}(\omega)\|^2 = 1$ are obtained [96] at points where

$$\nabla E\{G(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2)\} - \beta \nabla E\{|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2\} = 0 \qquad (4.2.1)$$

where $\beta \in \mathbb{R}$, $E\{.\}$ denotes the statistical expectation, $(.)^H$ Hermitian transpose, $\|.\|$ Euclidian norm, $|.|$ absolute value, $G(.)$ is a nonlinear contrast function, and the gradient denoted by ∇, is computed with respect to the real and imaginary parts of $\mathbf{w}(\omega)$ separately.

The Newton method is used to solve this equation for which the total approximate Jacobian is [96]

$$J = 2(E\{g(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2) + |\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2 \acute{g}(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2)\} - \beta)I \qquad (4.2.2)$$

which is diagonal and therefore easily invertible, where I denotes the identity matrix and $g(.)$ and $\acute{g}(.)$ denote the first and second derivatives of the contrast function. Bingham and Hyvärinen obtained the following approximate Newton iteration:

$$\begin{aligned} \mathbf{w}(\omega) &\leftarrow \mathbf{w}(\omega) \\ &\quad - \frac{E\{\mathbf{x}(\omega)(\mathbf{w}^H(\omega)\mathbf{x}(\omega))^* g(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2)\} - \beta\mathbf{w}}{E\{g(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2) + |\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2 \acute{g}(|\mathbf{w}^H(\omega)\mathbf{x}(\omega)|^2)\} - \beta} \\ \mathbf{w}(\omega) &\leftarrow \frac{\mathbf{w}(\omega)}{\|\mathbf{w}(\omega)\|} \end{aligned} \qquad (4.2.3)$$

where $(.)^*$ denotes the complex conjugate. In the experiments the statistical expectation is realized as a sample average.

4.2.1 Robustness of contrast function

A good contrast function is one for which the estimator given by the contrast function is more robust to outliers in the sample values. This means that single highly erroneous observations do not have much influence on the estimator. Using a simple ICA estimation method to measure non-Gaussianity by kurtosis has a drawback in practice. The main problem is that the kurtosis is very sensitive to outliers e.g. if a sample of 1000 values of a random variable, of zero mean and unit variance, even contains only one sample equal to 10, then the kurtosis of that sample will be at least $10^4/1000 - 3 = 7$ which clearly indicates the single value is likely to make kurtosis large [6]. For zero mean complex random variables the kurtosis in [95] is defined as:

$$\begin{aligned} kurt(y(\omega)) = E\{|y(\omega)|^4\} &- E\{y(\omega)y^*(\omega)\}E\{y(\omega)y^*(\omega)\} \\ &- E\{y(\omega)y(\omega)\}E\{y^*(\omega)y^*(\omega)\} \\ &- E\{y(\omega)y^*(\omega)\}E\{y^*(\omega)y(\omega)\} \end{aligned} \quad (4.2.4)$$

however there are in total 2^4 ways to define the kurtosis [98]. Kurtosis defined in [99] is used in the referred paper [96], which is defined as:

$$\begin{aligned} kurt(y(\omega)) &= E\{|y(\omega)|^4\} - 2(E\{|y(\omega)|^2\})^2 - |E\{y^2(\omega)\}|^2 \\ &= E\{|y(\omega)|^4\} - 3 \end{aligned} \quad (4.2.5)$$

under the condition that $y(\omega) = \mathbf{w}^H(\omega)\mathbf{x}(\omega)$ is white i.e. the real and imaginary parts are uncorrelated and their variances are equal. The contrast function defined in [96] is:

$$J_G(\mathbf{w}(\omega)) = E\{G(|y(\omega)|^2)\} \tag{4.2.6}$$

The function used in the experiments of the proposed method $G(y(\omega))$ $= \log(b + y(\omega))$, is the same as that used in [96] and its derivative is $g(y(\omega)) = 1/(b + y(\omega))$, where b in an arbitrary positive constant, empirically $b \approx 0.1$ is a reasonable value. The robustness of the estimator is captured in the slow growth of G, as its argument increases [6].

4.3 Proposed intelligent initialization based FastICA algorithm

4.3.1 Initialization for stationary sources

In the geometrical model (presented in Section 3.2) the actual mixing matrix $\mathbf{H}(\omega)$ should include the reverberation terms related to the reflection of sounds by the obstacles and walls, which are not possible to capture with the video system. However in such a room environment it will generally always contain the direct path components as in the equations in Section 3.2. Therefore, $\hat{\mathbf{H}}(\omega)$ is considered as a crude biased estimate of the frequency domain mixing filter matrix, but one which provides the learning algorithm with a good initialization *whilst importantly avoiding the bias in learning when used as a constraint within the FDCBSS algorithm as in Section 3.4.*

The position and direction information obtained from the video cameras equipped with a speaker tracking algorithm is automatically passed to (3.2.13) and (3.2.14) to estimate the $\hat{\mathbf{H}}(\omega)$. At the starting

point or the point when speakers move greater than the maximum step size, $\hat{\mathbf{H}}_1(\omega)$ is used to initialize the fixed point algorithm [96] for each frequency bin

$$\mathbf{W}_1(\omega) = \mathbf{Q}_1(\omega)\hat{\mathbf{H}}_1(\omega) \tag{4.3.1}$$

where $\mathbf{Q}_1(\omega)$ is the whitening matrix [6] of the mixtures and suffix 1 stands for the starting point or the point when speakers move greater than the maximum step size.

Before starting the process $\hat{\mathbf{H}}_1(\omega)$ is normalized once using $\hat{\mathbf{H}}_1(\omega) \leftarrow \hat{\mathbf{H}}_1(\omega)/\|\hat{\mathbf{H}}_1(\omega)\|_F$ where $\|.\|_F$ denotes the Frobenius norm.

The algorithm convergence depends on the estimate of $\hat{\mathbf{H}}_1(\omega)$, to improve accuracy. As will be shown by later simulations, an estimate of $\hat{\mathbf{H}}_1(\omega)$ obtained from Section 3.2 can result in a good performance for the proposed algorithm in a moderate reverberant environment.

4.3.2 Initialization when sources are moving in short steps

Since the unmixing matrix calculated by geometrically based initialized ICA (IIFastICA, initialization described in the above section) is permutation free, therefore $\mathbf{H} \approx \mathbf{W}^{-1}$. Scaling is not a major issue, and normalization during learning [66] can mitigate its effect.

For the real room recordings it is practically verified that when the sources move in small steps, then the inverse of the unmixing matrix obtained from IIFastICA at previous step j can be considered as a crude, albeit biased, estimate i.e. $\hat{\mathbf{H}}_{j+1} \approx \mathbf{W}_j^{-1}$, for current step $j+1$, and with the whitening matrix calculated from the recorded mixtures

at current step will provide the intelligent initialization as

$$\mathbf{W}_{j+1}(\omega) = \mathbf{Q}_{j+1}(\omega)\hat{\mathbf{H}}_{j+1}(\omega) \tag{4.3.2}$$

The equivalence between frequency domain blind source separation and frequency domain adaptive beamforming is already confirmed in [100]. It is highlighted that the whitening matrix $\mathbf{Q}(\omega)$ has strong impact in such smart initializations.

The above initializations increase the separation performance together with mitigating the permutation problem. Crucially, in the proposed Intelligently Initialized FastICA (IIFastICA) method, since the algorithm essentially fixes the permutation at each frequency bin, there will be no problem while aligning the estimated sources for reconstruction in the time domain.

As an initial step, it is usual in ICA approaches to sphere or whiten the data

$$\mathbf{z}(\omega) = \mathbf{Q}(\omega)\mathbf{x}(\omega) \tag{4.3.3}$$

$$\mathbf{Q}(\omega) = \mathbf{D}^{-\frac{1}{2}}(\omega)\mathbf{E}^{H}(\omega) \tag{4.3.4}$$

where $\mathbf{E}(\omega) = \{\mathbf{e}_1(\omega), ..., \mathbf{e}_1(\omega)\}$ is the matrix whose columns are the unit-norm eigenvectors of the covariance matrix $\mathbf{C}_{\mathbf{x}}(\omega) = E\{\mathbf{x}(\omega)\mathbf{x}^{H}(\omega)\} = \mathbf{E}(\omega)\mathbf{D}(\omega)\mathbf{E}^{H}(\omega)$ and $\mathbf{D}(\omega) = diag(d_1(\omega), ..., d_n(\omega))$ is the diagonal matrix of the eigenvalues of $\mathbf{C}_{\mathbf{x}}$. Since $\mathbf{z}$ is white, i.e., zero mean, unit variance and with uncorrelated real and imaginary parts of equal variances, therefore $E\{\mathbf{z}(\omega)\mathbf{z}^{T}(\omega)\} = \mathbf{o}$ and $E\{\mathbf{z}(\omega)\mathbf{z}^{H}(\omega)\} = I$, and $\mathbf{D}^{-\frac{1}{2}}(\omega)$ plays a vital role in $E\{\mathbf{z}(\omega)\mathbf{z}^{H}(\omega)\} = I$.

Next, when sources are at the starting point or the point when

speakers move greater than the maximum step size, the row vectors of $\mathbf{W}_1(\omega)$ obtained from (4.3.1) are used, one-by-one, to initialize the fixed point algorithm [96] for each frequency bin. Once the sources start moving within the limit of the maximum step size then similarly, the row vectors of $\mathbf{W}_{j+1}(\omega)$, obtained from (4.3.2) are used, one-by-one, to initialize [96] for each frequency bin.

The final update equation to be initialized by a row vector of $\mathbf{W}_1(\omega)$ or $\mathbf{W}_{j+1}(\omega)$, for each vector of each frequency bin, by applying (4.3.3), and by multiplying both sides of (4.2.3) by $\beta - E\{g(|\mathbf{w}^H\mathbf{x}|^2) + |\mathbf{w}^H\mathbf{x}|^2 \acute{g}|\mathbf{w}^H\mathbf{x}|^2)\}$, the fixed point algorithm simplifies as:

$$\begin{aligned}\mathbf{w}_1^+(\omega) \leftarrow & E\{\mathbf{z}(\omega)(\mathbf{w}_1(\omega)^H\mathbf{z}(\omega))^* g(|\mathbf{w}_1(\omega)^H\mathbf{z}(\omega)|^2)\} \\ & -E\{g(|\mathbf{w}_1(\omega)^H\mathbf{z}(\omega)|^2) + |\mathbf{w}_1(\omega)^H\mathbf{z}(\omega)|^2 \\ & \acute{g}(|\mathbf{w}_1(\omega)^H\mathbf{z}(\omega)|^2)\}\mathbf{w}_1(\omega) \qquad (4.3.5)\end{aligned}$$

$$\mathbf{w}_1(\omega) \leftarrow \frac{\mathbf{w}_1^+(\omega)}{\|\mathbf{w}_1^+(\omega)\|} \qquad (4.3.6)$$

which importantly eliminates the need to calculate β.

Since there are M independent components, the other separating vectors, i.e. $\mathbf{w}_i(\omega)$, $i = 2, \cdots, M$, are calculated in a similar manner and than decorrelated in a Gram-Schmidt-like [101] decorrelation i.e. deflationary orthogonalization scheme. The deflationary orthogonalization for the M-th separating vector [6] takes the form

$$\mathbf{w}_M(\omega) \leftarrow \mathbf{w}_M(\omega) - \sum_{j=1}^{M-1}\{\mathbf{w}_M^H(\omega)\mathbf{w}_j(\omega)\}\mathbf{w}_j(\omega) \qquad (4.3.7)$$

$$\mathbf{w}_M(\omega) \leftarrow \frac{\mathbf{w}_M(\omega)}{\|\mathbf{w}_M(\omega)\|} \qquad (4.3.8)$$

and $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), \cdots, \mathbf{w}_M(\omega)]^H$ is formulated after separating all vectors of each frequency bin.

Generally, for determined BSS problem, M independent components, i.e. $\mathbf{w}_i(\omega)$, $i = 2, \cdots, M$, are calculated in parallel to obtain $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), \cdots, \mathbf{w}_M(\omega)]^H$ for each frequency bin are decorrelated in a symmetric orthogonalization scheme which is more accurate than a deflationary orthogonalization in the exactly determined case addressed in this thesis. The symmetric orthogonalization takes the form [6]

$$\mathbf{W}(\omega) \leftarrow \mathbf{W}(\omega)\{\mathbf{W}^H(\omega)\mathbf{W}(\omega)\}^{-\frac{1}{2}} \tag{4.3.9}$$

$$\mathbf{W}(\omega) \leftarrow \frac{\mathbf{W}(\omega)}{\|\mathbf{W}(\omega)\|} \tag{4.3.10}$$

Summary Table: Implementation steps for the proposed IIFastICA algorithm

1. *Initialize parameters,* N, M, T, Q, γ, f_s, C, r, κ, b, T_b, *maximum step size, maximum count.*
2. *Read input mixtures, i.e., time samples* $\mathbf{x}(t)$.
3. *Calculate the distances* d_{ij} *and angle of arrivals* θ_{ij} *between the speakers and the microphones on the basis of video information.*
4. *Calculate the propagation time* τ_{ij} *using (3.2.13) and attenuation* α_{ij} *using (3.2.14) on the basis of* d_{ij}, r, κ, θ_{ij}, f_s, C.
5. *Find the estimate of mixing filter* $\hat{\mathbf{H}}(\omega)$ *using (3.2.3).*
6. *Normalized* $\hat{\mathbf{H}}(\omega) \leftarrow \hat{\mathbf{H}}(\omega)/\|\hat{\mathbf{H}}(\omega)\|_F$

7. *Calculate the Intelligent Initialization:*

   - *Calculate* $\mathbf{W}_1(\omega)$ *using (4.3.1) when sources are at starting point or speaker move greater than the maximum step size.*
   - *Calculate* $\mathbf{W}_{j+1}(\omega)$, $j \geq 1$ *using (4.3.2), when sources are moving in small steps.*

8. *Whiten the data after conversion into the frequency domain:*

   - *Convert* $\mathbf{x}(t)$ *to* $\mathbf{x}(\omega)$ *using the DFT.*
   - *Whiten the data* $\mathbf{z}(\omega) = \mathbf{Q}(\omega)\mathbf{x}(\omega)$ *using (4.3.3), calculate* $\mathbf{Q}(\omega)$ *from (4.3.4).*

9. *Define the non-linear function G and calculate its first and second derivative g and g' respectively.*

10. *Update unmixing matrix:*

    - *FOR* $i = 1$ *to* M.
      - *WHILE* $\{min(abs(\nabla \mathbf{w}_i(\omega)),\ maximum\ count - count) > 0.0001\}$
        - *Update each vector of each frequency bin* $\mathbf{w}_i^+(\omega) = E\{\mathbf{z}(\omega)(\mathbf{w}_i(\omega)^H\mathbf{z}(\omega))^* g(|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)|^2)\} - E\{g(|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)|^2) + |\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)|^2 \dot{g}(|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)|^2)\}\mathbf{w}_i(\omega)$ *using (4.3.2).*
        - *Normalize each vector* $\mathbf{w}_i(\omega) = \frac{\mathbf{w}_i^+(\omega)}{\|\mathbf{w}_i^+(\omega)\|}$, *using (4.3.6).*
        - *Do the deflationary orthogonalization of the first vector* $(i > 1)$ *using (4.3.7) and normalize each vector* $\mathbf{w}_i(\omega) = \frac{\mathbf{w}_i^+(\omega)}{\|\mathbf{w}_i^+(\omega)\|}$, *using (4.3.8).*
        - $counter = counter + 1$

- *END WHILE.*

- *END FOR.*

11. *After Separation of all vectors of each frequency bin formulate the unmixing filter* $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), \cdots, \mathbf{w}_M(\omega)]^H$.

12. *Calculate* $\mathbf{W}(\omega) = \mathbf{W}(\omega)\mathbf{Q}(\omega)$.

13. *Calculate* $\mathbf{y}(\omega)$ *according to (2.1.9).*

14. *Reconstruct the time domain signals* $\mathbf{y}(t) = IDFT(\mathbf{Y}(\omega))$.

15. *Calculate Performance Index (PI) (2.4.2) and Evaluate Permutation* $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$.

16. *Calculate the Signal-to-Interference Ratio (SIR) (2.4.4).*

17. *End.*

4.4 Experiments and results

In this section the proposed method is evaluated in two parts, initially, when the sources are physically stationary and secondly, when the sources are physically moving stepwise.

4.4.1 Stationary sources

In the experiments which correspond to the environment in Figure 3.2, the Bingham and Hyvärinen [96] algorithm and the proposed algorithm were tested for real room recordings. The objective evaluation of BSS requires the mixing filter therefore the audio signals were convolved with real room impulse responses recorded in the room. The separation of the real recorded signals is also evaluated subjectively by listening tests and mean opinion score (MOS) is also provided. The length of the audio signals was 10sec. The other important variables were selected as: FFT length $T = 1024$ and filter length $Q = 512$ (half of T), $r = 4$, the sampling frequency for the recordings was 8KHz, size of the room was 5x5x5 m^3 and the room impulse duration was 130ms. In the proposed algorithm $G(y) = \log(b + y)$, with $b = 0.1$.

First the performance is evaluated on the basis of performance index (PI) (2.4.2). The resulting performance indices are shown in Figure 4.1 which show good performance for the proposed algorithm i.e. close to zero across the majority of the frequency bins. This is due to geometrical information used in the initialization. Both algorithms were tested at a fixed iteration count of seven, as the proposed algorithm has converged in this number of iterations. The visual modality therefore renders the BSS algorithm semiblind and thereby much improves the resulting performance and rate of convergence.

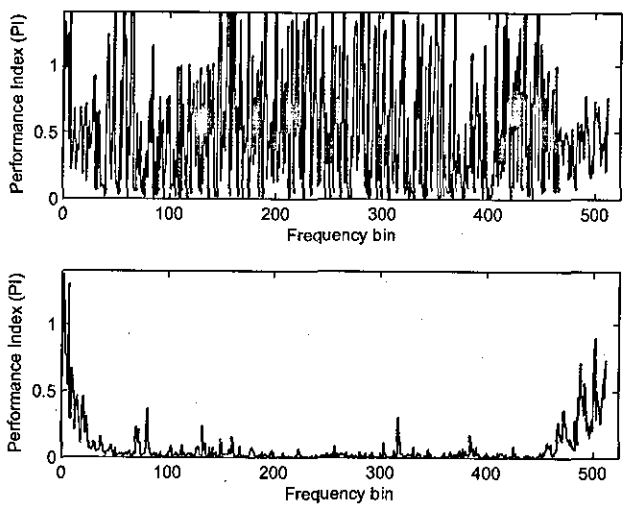

Figure 4.1. Performance index at each frequency bin for the Bingham and Hyvärinen algorithm on the top [96] and the proposed algorithm at the bottom, audio signals of length 10sec were convolved with recorded real room impulse responses, iteration count = 7 was fixed. A lower PI refers to a superior method.

The permutation performance is also evaluated on the basis of the criterion mentioned in Subsection 2.4.3. In Figure 4.2 the results confirm that the proposed algorithm automatically mitigates the permutation at each frequency bin. Since in Figure 4.2 (bottom) $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ for all frequency bins therefore the multimodal method provides appropriate solution to the permutation problem.

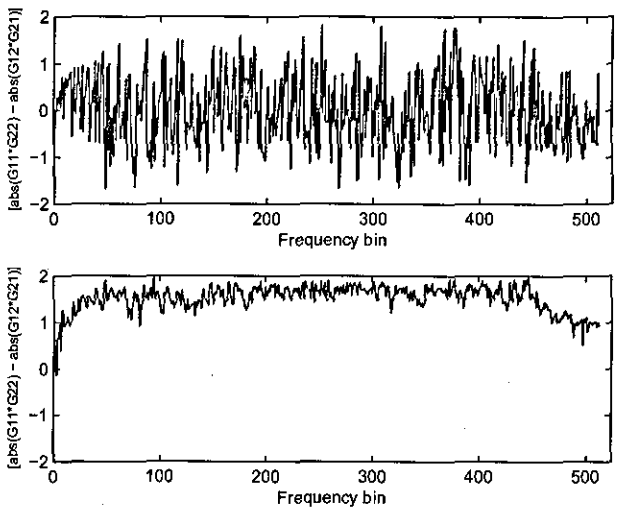

Figure 4.2. Evaluation of permutation in each frequency bin for the Bingham and Hyvärinen algorithm at the top [96] and the proposed algorithm at the bottom, audio signals of length 10sec were convolved with recorded real room impulse responses, iteration count = 7 was fixed. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

In contrast, the performance indices and evaluation of permutation by the original FastICA algorithm [96] (MATLAB code available online) with random initialization, on the recorded mixtures are shown in Figure 4.3. It is highlighted that thirty-five iterations are required for the performance level achieved in Figure 4.3(a) with no solution for permutation as shown in Figure 4.3(b). The permutation problem in frequency domain BSS degraded the SIR to approximately zero on the recorded mixtures.

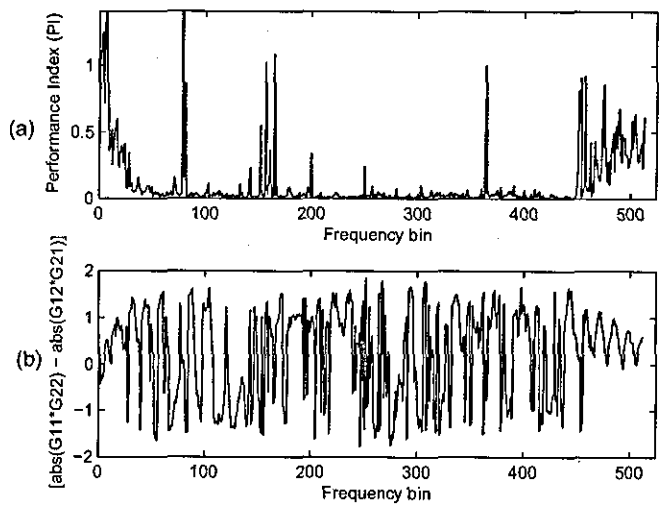

Figure 4.3. (a) Performance index at each frequency bin and (b) Evaluation of permutation in each frequency bin for Bingham and Hyvärinen FastICA algorithm [96], audio signals of length 10sec were convolved with recorded real room impulse responses, iteration count = 35 was fixed. A lower PI refers to a better separation and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Figure 4.4 confirms the convergence of the underlying cost, i.e. $E\{G(|\mathbf{w}^H\mathbf{x}|^2)\}$, within seven iterations for the proposed algorithm. The results are averaged over all frequency bins. The convergence within seven iterations with solution for permutation confirms that the proposed algorithm is likely to be robust and suitable for real-time implementation.

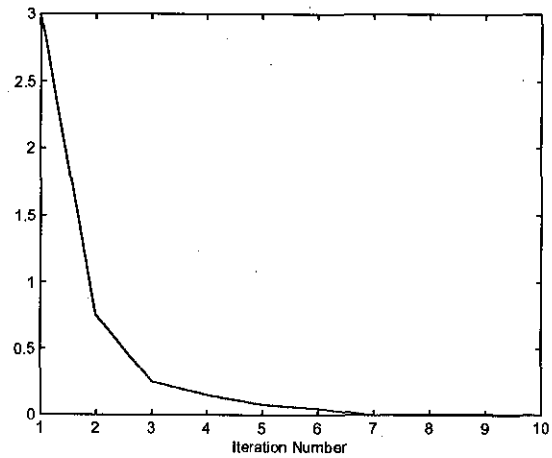

Figure 4.4. The convergence graph of the cost function of the proposed algorithm for the audio signals of length 5sec convolved with recorded real room impulse responses, using contrast function $G(y) = \log(b+y)$; the results are averaged over all vectors of all frequency bins.

The signal-to-interference ratio (SIR) (2.4.4) was calculated and comparison of SIR-Improvement between algorithms and the proposed method for different sets of mixtures is shown in Table 4.1. The proposed IIFastICA provides 4.3dB and 5.1dB SIR-Improvement than the Wenwu et al. Method [74] and Parra's Method [19] respectively.

Table 4.1. Comparison of SIR-Improvement between algorithms and the proposed method for different sets of mixtures.

Algorithms	SIR-Improvement/dB
Parra's Method [19]	6.8
FDCBSS [66]	9.4
Wenwu et al. Method [74]	10.2
IIFastICA	14.5

Finally, the proposed method was also evaluated subjectively by listening, with eight people participated in the tests and MOS is 4.7, which is very high quality.

4.4.2 Stepwise moving sources

In this section the proposed IIFastICA technique for moving sources is evaluated. The simulations were performed on real recorded speech signals generated for a room geometry as illustrated in Figure 4.5. The audio signals were recorded, the estimate of $\hat{\mathbf{H}}_1(\omega)$ was calculated on the basis of geometrical information obtained from video cameras, and real room impulse responses were also recorded, when speaker 1 was at position A and speaker 2 was at position C. The recorded mixtures were separated with intelligent initialization for starting point (4.3.1) and PI with solution to permutation was evaluated with the recorded impulse responses. When the speakers started movement then the positions of the speakers after each two second interval were marked (room impulse responses at marked points were calculated for PI) and the same procedure was repeated with the initialization for stepwise moving sources (4.3.2). The other important parameters were: block length T_b = 2sec, FFT length $T = 1024$, filter length $Q = 512$ half of T and 50% overlapping was used. The room impulse response duration was 130 ms. Speaker1 moved from A to B i.e. 60 degrees counterclockwise and Speaker2 moved from C to E via D in a back and forth motion i.e. 30 degrees in total at a speed of 5 deg/sec. The maximum step size was 10 degrees. This could correspond to moving around a circular table in a tele-conferencing context. To reduce the complexity of the tracker circular motion is assumed in this work. In the proposed algorithm $G(y) = \log(b + y)$, with $b = 0.1$.

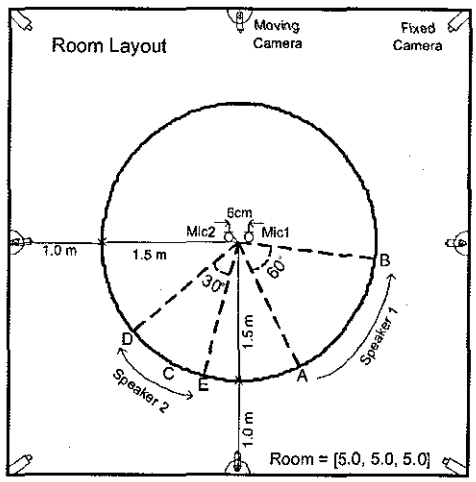

Figure 4.5. A two-speaker two-microphone layout for recording within a reverberant (room) environment. Speakers move with the speed of 5 deg/sec. Room impulse response length is 130 ms.

Initially, PI (2.4.2) was calculated and the resulting performance indices are shown in Figure 4.6. Figure 4.6(a) shows good performance i.e. close to zero across the majority of the frequency range, since this is due to the geometrical based initialization (4.3.1). In Figures 4.6(b) and 4.6(c) the performance is again good but slightly degraded because the estimates for initialization are slightly more biased as explained in Subsection 4.3.2.

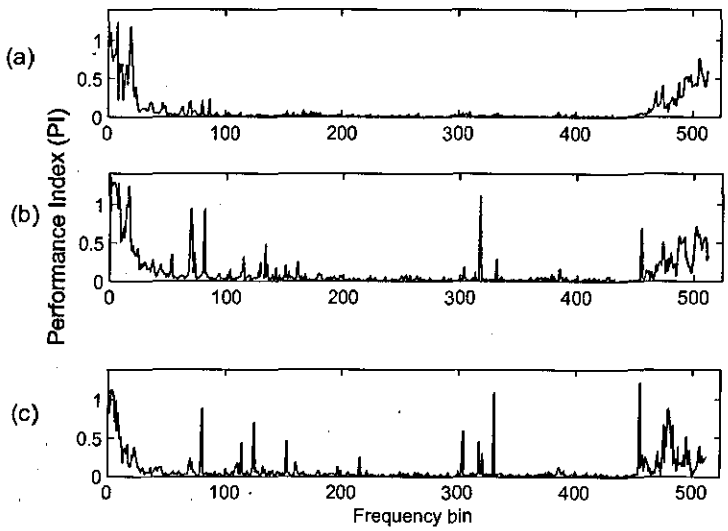

Figure 4.6. Performance Index at each frequency bin when (a) Both sources are static i.e. speaker 1 at position A and speaker 2 at position C, (b) One source moved i.e. speaker 1 at position A and speaker 2 moved 10 degrees counterclockwise from position C, and (c) Both sources moved i.e. speaker 1 moved 10 degrees counterclockwise from position A and speaker 2 moved 5 degrees clockwise from position C. A lower PI refers to a better separation.

Secondly, the permutation performance was also evaluated. Figure 4.7 confirmed that the proposed algorithm automatically mitigates the permutations due to the intelligent initializations mentioned in Sections 4.3.1 & 4.3.2, and therefore no additional processing is required. Figures 4.7(a) and 4.7(b) show improved results over 4.7(c) because when both sources are moving there is more variation in the mixing environment, in spite of the strong impact of $Q(\omega)$ as mentioned in section 4.3.2 the initialization is slightly more biased, therefore as explained in the next paragraph, the algorithm generally requires more iterations to converge when sources are moving.

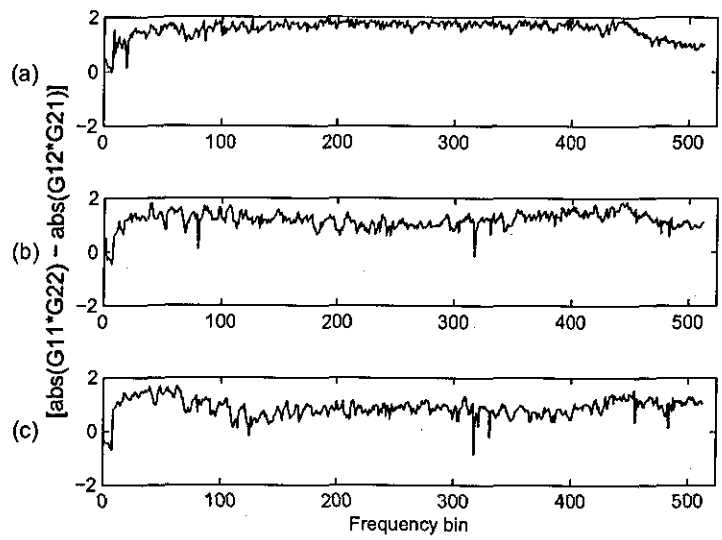

Figure 4.7. Evaluation of permutation in each frequency bin when (a) Both sources are static i.e. speaker 1 at position A and speaker 2 at position C, (b) One source moved i.e. speaker 1 at position A and speaker 2 moved 10 degrees counterclockwise from position C, and (c) Both sources moved i.e. speaker 1 moved 10 degrees counterclockwise from position A and speaker 2 moved 5 degrees clockwise from position C. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Since the convergence rate of any algorithm has a vital rule for a real time system. The number of iterations required for the convergence of the underlying cost, in the proposed IIFastICA algorithm at different conditions of the sources is shown in Table 4.2. Algorithm [96] is not applicable for moving sources. The maximum of seven iterations when both sources are moving confirms that the proposed algorithm is more suitable for a real-time system.

Finally, SIR (2.4.4) was calculated. The separation was performed at $T_b = 2$ sec, and average SIR-Improvement when both speakers were stationary was 14 dB, when one speaker was moving 13.5 dB and when both speakers were moving 12.8 dB. The minimum 12.8 dB SIR-Improvement again confirmed that no additional postprocessing is

Table 4.2. Number of iterations required for convergence in the proposed IIFastICA algorithm averaged over all frequency bins under different conditions of the sources.

Sources Condition	Iterations (IIFastICA)	Iterations ([96])
Both sources are static	7	35
One source moved	9	-
Both sources moved	12	-

required which has been confirmed subjectively by listening tests, with six people participated in the tests and MOS is 4.2.

4.5 Summary

A new multimodal method for FDCBSS, with intelligent initialization for FastICA, for moving sources has been presented in this chapter. The advantage of the proposed algorithm was confirmed in simulations from a real room environment. The location and direction information were obtained using a number of cameras and this information was used in the initialization of the proposed algorithm. The proposed multimodal method is block-based and the initialization is performed based on either the geometrical information obtained from tracking or the BSS results from the previous block. The separation was evaluated objectively by the performance indices with solution for permutation at frequency bin level and overall SIR-Improvement at different conditions of sources, and also confirmed subjectively by listening tests. The outcome of this method is a step towards solving the cocktail party problem for moving sources by using a cognitive approach.

Chapter 5

EXPLOITING ALL COMBINATIONS OF MICROPHONE SENSORS IN OVERDETERMINED FREQUENCY DOMAIN BLIND SOURCE SEPARATION OF SPEECH SIGNALS

In this chapter, based on the multimodal method presented in Chapter 4, a new approach to overdetermined frequency domain blind source separation (BSS) of speech signals which exploits all combinations of observations and hence varying inter-microphone spacings is presented. The observations are divided into subgroups so that conventional fre-

quency domain BSS algorithms can be used. By evaluating the separation performance obtained from each group on the basis of approximately measuring the independence of separated signals, the output of the group that has the best performance amongst all groups on a frequency-by-frequency basis is chosen as the overall output. The separated signals of the overall system are then obtained by transforming their frequency domain representations into the time domain. Simulation results based on speech signals confirm that the presented approach has better performance based on the performance index (PI) as compared with a conventional scheme using only one microphone group and an existing overdetermined frequency domain BSS algorithm.

5.1 Introduction

BSS algorithms are designed to recover unobservable source signals from observed mixtures with the assumption that the sources are mutually independent. Convolutive BSS algorithms have received much attention recently since they have more practical applications as compared with instantaneous BSS algorithms [102]. Frequency domain approaches simplify the convolutive BSS problem into the instantaneous but complex BSS problem at each frequency bin, thus retaining the advantages of mathematical simplicity, reduction of the computational complexity, and fast convergence.

For most classical frequency domain BSS algorithms, N observed mixture signals are required and sufficient to recover N source signals. For conventional convolutive frequency domain BSS algorithms, and it is also observed in the simulation results of Chapters 3, 4 & 6, in some frequency bins, the performance is not good, for example, due to the

ill-conditioning of the mixing matrix, which is related to the positions of the sources and inter-element microphone spacings (the problem is more obvious in the experiments of Chapter 3 when positions of the sources and the sensors were changed). Assuming that a microphone array which contains N microphones where $M > N$, i.e., the overdetermined case, may want to use all these observed signals to attain a better BSS performance rather than only use N observed signals. Several overdetermined BSS algorithms have been proposed in recent years. It is shown in [103] that the unmixing filters of overdetermined BSS contain much simpler structure, thus are easier extract. Rather than utilizing all the observed signals simultaneously in the overdetermined BSS algorithms, a principal component analysis (PCA) approach is performed as a preprocessing in [104] so that the output of PCA can then be used to extract the source signals by using conventional BSS algorithms. In the method proposed in [105] different sensors are utilized for different frequency bins: sensors with wide sensor spacings are used for low frequency bins, and sensors with narrow sensor spacings are used for a high frequency range. As discussed in [106] for low frequency bins the methods proposed in [104] and [105] have similar performance, while the method proposed in [104] is preferred for high frequency bins. However, the technique proposed in [104] utilizes the PCA preprocessing which is not suitable for online BSS, while the method proposed in [105] needs carefully designed microphone arrays, which may limit their application.

Another attractive overdetermined BSS algorithm can be seen in [107], in which the observations are divided into subgroups and each group contains N observed signals. Conventional BSS algorithms can

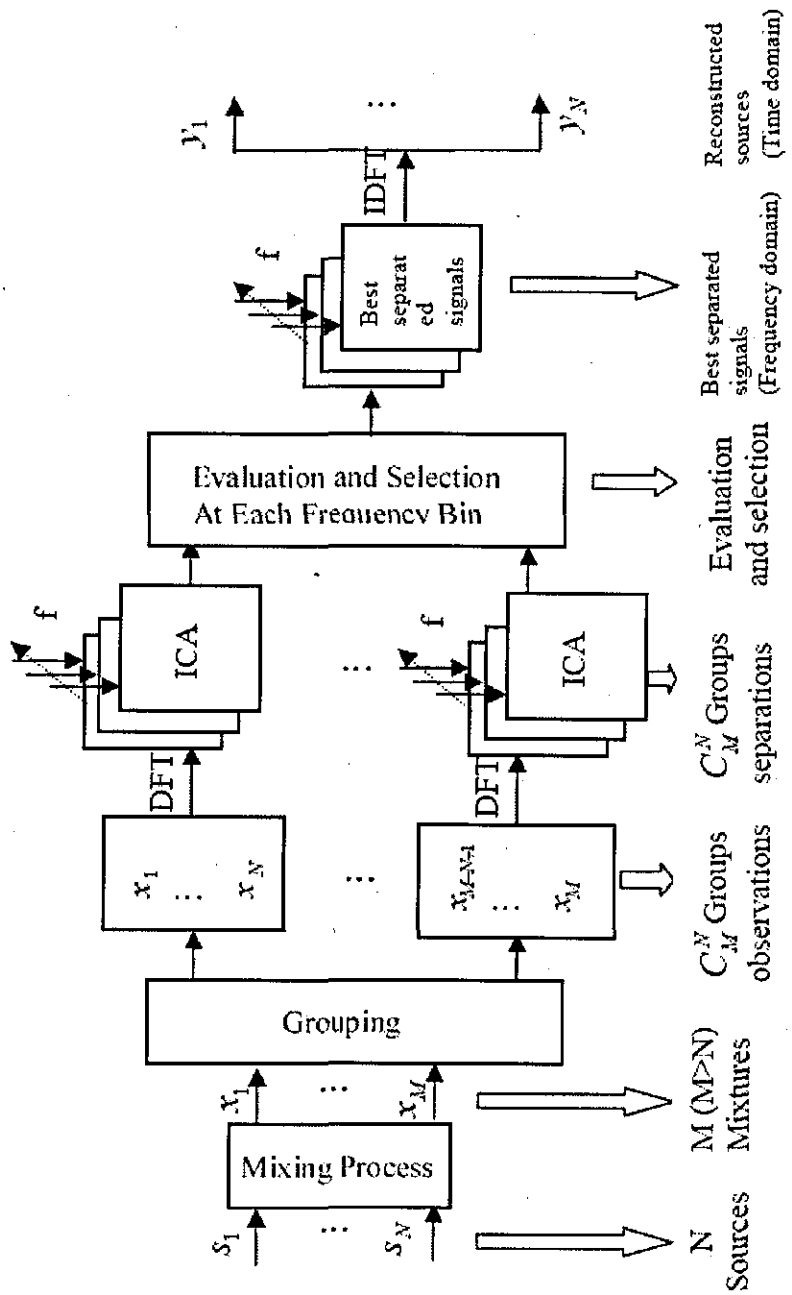

Figure 5.1. Illustration of the proposed grouping approach in overdetermined frequency domain blind source separation.

then be used for each group. The overall output of the algorithm is the average of outputs of all groups. In this chapter therefore a new overdetermined BSS approach is presented, where groups of observed

signals obtained from different microphones, with different positions and hence different inter-element spacings can be obtained for the BSS solution. Unlike the averaging operation performed in [107], an approximate evaluation of the independence of the separated signals for each group is performed frequency-by-frequency, and the separated signals of the group with the best performance are chosen as the overall output of the system. The separated signals of the overall system are obtained by transforming their frequency domain representations into the time domain. The implementation of the presented approach is illustrated in Fig. 5.1, the elements of which will be explained further in the sequel. By utilizing such an approach, the ill-conditioning problem for an individual group of observed signals can be avoided. As will be shown in the later simulations, the proposed approach has better performance based on the measurement of the performance index (PI) as compared with a conventional scheme using only one microphone group and the averaging approach proposed in [107].

In the following section a fast fixed-point algorithm for complex valued signals based on over determined BSS algorithms is presented. In Section 5.3 the simulation results confirm the usefulness of the algorithm. Finally, in Section 5.4 the chapter is summarized.

5.2 Algorithm formulation

The time domain mixing (or generative) model and separation model are shown in (2.1.2) and (2.1.5) respectively. Using a T-point windowed discrete Fourier transformation (DFT), the time domain signal can be converted into the frequency domain signal as shown in (2.1.7) and (2.1.9) respectively.

Many methods have been proposed to separate the observed signals in the frequency domain. In this chapter, the IIFastICA algorithm presented in Chapter 4 is used due to its robust convergence properties and solution to permutation. The update rule for the unmixing matrix and the separated signals is the same at each frequency bin. Initially, the mixtures are whitened for each frequency bin (4.3.3). In the IIFastICA algorithm vectors of the unmixing matrix $\mathbf{W}(\omega)$ are updated row by row as in (4.3.2) and after the first vector all other vectors are decorrelated by a deflation scheme based on a Gram-Schmidt-like decorrelation (4.3.7).

The nonlinear function $G(y) = \log(b+y)$ is used in this work due to its robust property [108], b is a small positive value and chosen as 0.1. Similar to all the other frequency domain convolutive BSS algorithms, the scale ambiguity problem appears at different frequency bins. This scale ambiguity problem is mitigated by the normalization approach as discussed in Chapter 3 & 4. In this chapter the performance index (PI) (2.4.2) which is a measurement of separation performance is also used.

In the IIFastICA method, N source signals can be recovered from N observed signals, i.e., an exactly determined approach. If the observed signals satisfy $M > N$, i.e., an overdetermined problem, then C_M^N groups of observed signals can be obtained, which can all be utilized to recover the original source signals. The performance of BSS obtained from these groups may be different due to different positions and inter-element spacings of the microphones. For a certain frequency bin, the BSS performance obtained from some groups may be poor, and good for other groups. The idea of the proposed approach is to evaluate the BSS performance of each group at every frequency bin, and choose the

output of the group with the best performance at each frequency bin as the overall system output. Assuming $K = C_M^N$, i.e., K groups of observed signals are obtained, the assessment of the BSS performance of each group can be obtained by approximately measuring the independence between the separated signals. Obviously, the measurement of the independence of the separated signals is a function of the separation matrix $\mathbf{W}(\omega)$. By denoting the unmixing matrix obtained from the kth group of observed signals as $\mathbf{W}_k(\omega)$ and $J(\mathbf{W}_k(\omega)), k = 1, ..., K$ as the measures of independence, the group with the minimum value of $J(\mathbf{W}_k(\omega))$ can be identified as

$$l = \arg\min_k J(\mathbf{W}_k(\omega)) \tag{5.2.1}$$

The output of the overall system at each frequency bin can then be described as

$$\mathbf{y}(\omega) = \mathbf{W}_l(\omega)\mathbf{x}_l(\omega) \tag{5.2.2}$$

where $\mathbf{x}_l(\omega)$ is the input vector of the lth group. As an example, if there are two source signals to be recovered, an approximate assessment of the independence of two separated signals can be performed by measuring their coherence function. Assuming at frequency bin ω two separated signal sequences are obtained $y_1(\omega)$ and $y_2(\omega)$

$$J(\mathbf{W}_k(\omega)) = \frac{|\langle y_1(\omega)y_2(\omega)\rangle|_n}{\sqrt{\langle|y_1(\omega)|^2\rangle_n\langle|y_2(\omega)|^2\rangle_n}} \tag{5.2.3}$$

where $\langle\cdot\rangle_n$ denotes the sample average over index n. Second order statistics are used to approximately measure independence as it is robust to the small sample numbers available in frequency domain BSS.

It is important to mention that for the proposed approach, the computational complexity will be high, especially for large number of sources and mixtures, since the number of observed signal groups will be large. For this case, all combinations of observations should not be used. A simple solution to reduce the computational complexity and retain the advantages of the proposed approach is to only choose the groups with different inter-microphone spacings.

In the next section simulations will be performed based on a two source two observation BSS problem, and the proposed approach with the criterion formulated in (5.2.3) will be compared with the complex FastICA algorithm and the approach proposed in [107] to show its advantages.

Summary Table: Implementation steps for the proposed algorithm

1. *Initialize parameters,* N, M, T, Q, γ, f_s, C, r, κ, b, T_b, *maximum count.*

2. *Calculate the number of groups* $K = C_M^N$.

3. *Read input mixtures for each group* k, $k = 1, ..., K$ *i.e., time samples* $\mathbf{x}_k(t)$.

4. *Calculate unmixing matrix for each group:*

   - *FOR* $i = 1$ *to* K.

     - *Calculate* $\mathbf{W}_i(\omega)$ *with IIFastICA.*
     - *Calculate* $\mathbf{y}_i(\omega)$ *according to (2.1.9).*
     - *Calculate* $J(\mathbf{W}_i(\omega)) = \frac{|\langle y_1(\omega) y_2(\omega) \rangle|_n}{\sqrt{\langle |y_1(\omega)|^2 \rangle_n \langle |y_2(\omega)|^2 \rangle_n}}$ *for each group by using (5.2.3).*

   - *END FOR.*

5. *Find the group l with minimum value of $J(\mathbf{W}_k(\omega))$ using (5.2.1).*

6. *Calculate $\mathbf{y}(\omega) = \mathbf{W}_l(\omega)\mathbf{x}_l(\omega)$ according to (5.2.2).*

7. *Reconstruct the time domain signals $\mathbf{y}(t) = IDFT(\mathbf{Y}(\omega))$.*

8. *Calculate Performance Index (PI) according to (2.4.2).*

5.3 Experiments and results

Two simulations are performed in this section. In the first simulation, a simulated room and its impulse responses $h_{ij}(p)$ between source j and sensor i are simulated by an image room [109]. The room size is set to be 5x5x5 meter^3 and the reflection coefficient is set to be 0.7 in approximate correspondence with the actual room. The reverberation time (RT) of this room is 130ms. Two anechoic 40 second male speech signals with a sampling frequency of 8kHz are utilized as source signals. Four sensors are utilized to show the advantages of the proposed approach. The IIFastICA algorithm is used with a DFT length of $T = 2048$ to implement the BSS algorithm and the criterion formulated in (5.2.3) is used to assess the BSS performance. The positions of these two sources are set to [4.00 3.50 1.50] and [4.00 1.00 1.50]. The positions of the four sensors are set to [2.00 2.00 1.50], [2.00 2.04 1.50], [2.00 2.08 1.50] and [2.00 2.12 1.50] respectively, i.e., the distance between neighboring sensors is 4cm. The set up of the room can be seen in Fig 5.2. The setup of the second simulation is the same as that of the first simulation except the RT is increased to 300ms by changing the reflection coefficient to 0.9. To evaluate the performance of different algorithms, the performance index (PI) (2.4.2) is measured at each frequency bin.

According to the set up of the simulation, C_4^2, i.e., 6 groups of observed signals are used to run the IIFastICA algorithm independently. The iteration number of the IIFastICA algorithm is set to be 7. The average PI measurements over all frequency bins obtained from these 6 groups for, the overdetermined BSS method in [107] and the proposed approach are shown in Table 1. To show clearly the advantages of

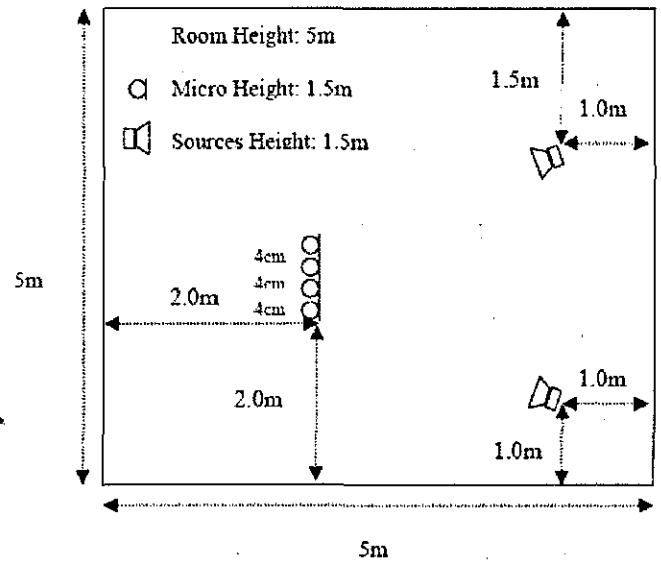

Figure 5.2. The room set up.

the proposed approach, PI for each frequency bin obtained from the overdetermined BSS method in [107] and from the proposed approach for both RT=130ms and RT=300ms room environments are plotted in Fig 5.3.

Table 1 shows that the averaged PI measurements over all frequency bins obtained from the proposed approach for both room environments with 130ms RT and 300ms RT are much smaller (a smaller value of PI indicates a better performance) as compared with those obtained from individual implementation of the FastICA algorithm and the method formulated in [107]. The advantages of the proposed approach can also be seen in Figure 5.3, in that for both room environments, the proposed approach has a smaller PI at nearly all frequency bins as compared with the method formulated in [107]. The simulation results indicate that a better performance of BSS is obtained by using the proposed approach.

Table 5.1. Averaged PI measurements of the complex FastICA algorithm, the method in [107] and the proposed approach

	Averaged PI(130ms)	Averaged PI(300ms)
Group1(x_1, x_2)	0.146	0.208
Group2(x_1, x_3)	0.164	0.184
Group3(x_1, x_4)	0.142	0.209
Group4(x_2, x_3)	0.190	0.222
Group5(x_2, x_4)	0.164	0.204
Group6(x_3, x_4)	0.153	0.178
Approach in [107]	0.160	0.201
Proposed approach	0.080	0.101

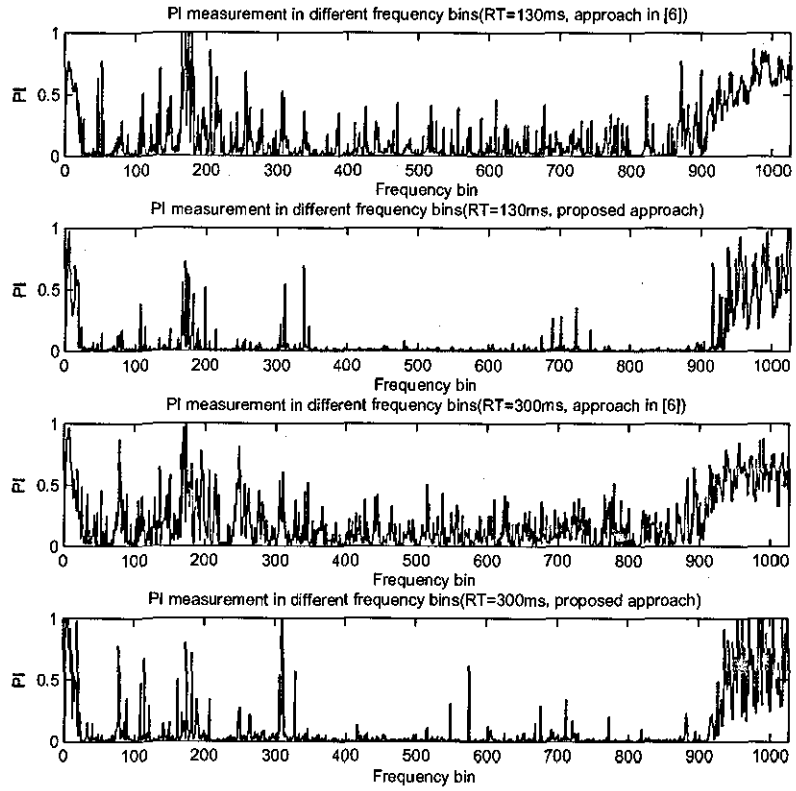

Figure 5.3. PI measurements of the method in [107] and the proposed approach at different frequency bins

5.4 Summary

A new frequency domain BSS approach exploiting non-uniform microphone spacings has been presented in this chapter. As has been shown by the simulation results, the proposed approach has better performance from objective measures as compared with a conventional scheme using only one microphone group and an existing overdetermined BSS method. This approach has the particular advantage that the best group is frequency dependent and therefore the best inter-element spacing is likely to be chosen.

Now that the initial part of the work on multimodal methods for BSS of physically stationary sources has been presented therefore the following questions remain:

- How can multiple moving sources be best detected and tracked by utilizing audio-visual information?
- How can audio-visual information be incorporated to solve the BSS of multiple moving sources?

These are answered in the next chapter to form a full BSS solution for stationary and moving sources.

Chapter 6

A MULTIMODAL SOLUTION TO BLIND SOURCE SEPARATION OF MOVING SOURCES

A novel multimodal solution is proposed for the problem of blind source separation (BSS) of moving sources. The challenge of BSS for moving sources is that the mixing filters are time varying, thus the unmixing filters should also be time varying, which are difficult to calculate in real time. In the proposed method, the visual modality is utilized to facilitate the separation for both stationary and moving sources. The movement of the sources is detected by a 3-D tracker based on video cameras. Positions and velocities of the sources are obtained from the 3-D tracker based on a Markov Chain Monte Carlo particle filter (MCMC-PF), which results in high sampling efficiency. The full BSS solution is formed by integrating a frequency domain blind source separation algorithm and beamforming: if the sources are identified as stationary for a certain minimum period, a frequency domain BSS algorithm is implemented with an initialization derived from the positions

of the source signals. Once the sources are moving, a beamforming algorithm which requires no prior statistical knowledge is used to perform real time speech enhancement and provide separation of the sources. Experimental results confirm that by utilizing the visual modality, the proposed algorithm not only improves the performance of the BSS algorithm and mitigates the permutation problem for stationary sources, but also provides a good BSS performance for moving sources in a low reverberant environment.

6.1 Introduction

Most existing BSS algorithms assume that the sources are physically stationary, i.e., the mixing filters are fixed. All these algorithms are based on statistical information extracted from the received mixed data [11, 19, 74]. However, in many real applications, the sources may be moving, for example, a presenter may walk around inside a room. In such applications, there will generally be insufficient data length available over which the sources are physically stationary, which limits the application of these algorithms. Thus BSS methods for moving sources are very important to solve the cocktail party problem in practice [1]. Only a few papers have been presented in this area [82–84, 110–112]. In [82], sources are separated by employing frequency domain ICA using a block-wise batch algorithm in the first stage, and the separated signals are refined by postprocessing in the second stage which constitutes crosstalk component estimation and spectral subtraction. In the case of [83], they used a framewise on-line algorithm in the time domain. However, both these two algorithms potentially assume that in a short period the sources are physically stationary, or the change of the mixing

filters is very slow, which are very strong constraints. In [84], BSS for time-variant mixing systems is performed by piecewise linear approximations. In [111], they used an online PCA algorithm to calculate the whitening matrix and another online algorithm to calculate the rotation matrix. However, both algorithms are designed only for instantaneous source separation, and cannot separate convolutive mixed signals. Fundamentally, it is very difficult to separate convolutively mixed signals by utilizing the statistical information only extracted from audio signals, and this is not the manner in which humans solve the problem [3] since they generally use both their ears and eyes.

In this work, a multimodal method is therefore proposed by utilizing not only received linearly mixed signals, but also the video information obtained from cameras. A video system can capture the approximate positions and velocities of the speakers, from which the directions and motions, i.e., stationary or moving, of the speakers can be identified. A source is identified as stationary if its velocity is approximately zero for a certain minimum period, so that enough data length can be obtained for frequency domain BSS algorithms. Furthermore, the direction of the source signals can also be obtained from the video cameras, and a geometrically based initialization can then be performed to improve the performance of the frequency domain BSS algorithm and mitigate the permutation problem [113]. If the velocity is larger than an upper bound value, the source is identified as moving. In this case, a beamforming method which does not need prior statistical information, in common with the fundamental assumptions in blind source separation, is used to enhance the signal from one source direction and reduce the energy received from another source direction, so that source separation

can be obtained. Although the beamforming approach can only reduce the signal from a certain direction and the reverberance of the interference still exists, it can obtain an acceptable separation performance in a low reverberation environment. Note that the beamforming approach only depends on the direction of the source signals, and no received audio data are required, thus an online real time source separation can be obtained [114].

The chapter is organized as follows: Section 6.2 presents the related work, Section 6.3 provides the system model, Section 6.4 explains the tracking process. Section 6.5 describes the source separation by combining frequency domain BSS and beamforming. Experimental results are provided in Section 6.6 based on real room recordings from our intelligent office. Finally, in Section 6.7 this chapter is summarized.

6.2 Related work

Most existing BSS algorithms are based on the statistical information, second order statistics (SOS)/ higher order statistics (HOS), extracted from the recorded data. Such methods are generally not applicable in CBSS of moving sources due to data length limitations and are therefore not included in simulation studies with moving sources. In the context of CBSS of moving sources in a moderate reverberant environment, with a reverberation time (RT) $<$ 130ms, it is believed that a multimodal method is necessary which exploits different processing techniques as a function of the velocity of the speakers. A key component in this method is the tracking of speakers. Many methods have been proposed for the tracking of speakers on the basis of audio information, visual information, or audio-visual fusion [32–40, 42, 43, 115–121]. Broadly

speaking, the differences among the existing approaches arise on the basis of single-person or multi-person tracking and the type of the sensor configuration used. In most of the works [32,35–37,40,115], on the basis of simple sensor configuration, either a single person is tracked in a single-person scene or the current active speaker is tracked in the multi-person scene. Multi-person tracking has been studied in [42,118–121] on the basis of only a single modality, either audio or video. In more recent works [33,38,39,116] the multi-person tracking problem has been studied by using the audio-visual sensor configuration. To the best of my knowledge, the most recent work on tracking, near to the requirement in proposed method, is proposed by Gatica-Perez et al. [39]. In this work a detect before track technique is applied, and a small microphone array with multiple uncalibrated cameras with non-overlapping field of view (FOV) is used for sensor configuration. For detection, audio observations are derived from a source localization algorithm and visual observations are based on models of the shape and spatial structure of human heads. For tracking, a 2-D tracker in the image plane is implemented with an MCMC-PF. In this case, for source separation, 3-D positions of the speakers are required to handle complicated human motions. Therefore, initially, video cameras should be calibrated [75] and have overlapping FOVs, because at least two cameras are required for conversion of 2-D image co-ordinates to 3-D real world coordinates. Secondly, it is computationally better to use one 3-D tracker rather than two 2-D trackers. Finally, audio localization is not effective due to the complexity in the case of multiple concurrent speakers. In [122] source localization based on binaural cues is proposed and cue selection is based on the results of a number of psychophysical studies. A

modeling mechanism is proposed and the implementation in real room environment is itself a task. Localization for a single active speaker based only on audio is also difficult because human speech is an intermittent signal and contains much of its energy in the low-frequency bins where spatial discrimination is imprecise, and locations estimated only by audio are also affected by noise and room reverberations [29]. In [29] the tracker proposed in [39] is implemented for speech enhancement and the simulation results confirm that for stationary speakers and overlapping speech utterances the audio-visual localization improves by $2cm$ and $3cm$ respectively as compared to using only visual information. McCowan in [30] proposed that any time when the distance between the tracked speaker location and the focus location of the beamformer exceeds $5cm$, the beamformer channel filters should be recalculated, so practically there is no significant improvement by integrating audio localization. In other recent works [42,43] only audio information is used. In [43] particle filtering is used for acoustic source localization and it is assumed that a single acoustic source with known speed of wave propagation is present in a reverberant environment. In [42] time difference of arrival (TDOA) estimation and localization of moving speakers is proposed which distinguish individual speakers in a multipath environment by associating one TDOA per frame to the predominant speaker. In the situation when speakers are simultaneously speaking and moving, both the above methods have limitations. In [123] joint acoustic source localization and orientation estimation using sequential monte carlo is presented and it is also highlighted in the paper that in a situation where only one microphone pair (sensor configuration used in this work) provides measurements then the performance is predictably

poorer. Therefore, in the proposed method the speakers are tracked by using only visual information motivated by Colin Cherry's observation that the human approach to solve the cocktail party problem exploits visual cues [1, 2]. In the application environment, an intelligent office, the cameras also benefit from being mounted above the height of a human and thereby make it easier to discriminate sources in close spatial proximity. In the proposed method, the source localization is performed by using the state-of-the-art Viola-Jones face detector [78]. The 3-D visual tracker is implemented with an MCMC-PF which results in high sampling efficiency. It is stressed that the domain of the proposed method in this chapter lies in system integration and the main contribution is to provide the proof of the concept for CBSS of moving sources. The areas of detection and tracking are disciplines in their own rights and in this chapter recent results from these fields to provide geometric information to facilitate a novel multimodal method to CBSS are simply exploited. The output of the tracking is position and velocity information, on the basis of which source separation is divided into two parts to provide the full BSS solution. As will be shown in later simulations, the proposed method can provide a reasonable BSS performance for moving sources in a low reverberant environment in which the RT is 130ms. The system model is described next.

6.3 The system model

The proposed method can be divided into two stages: human tracking to obtain position and velocity information; and source separation by utilizing the position and velocity information based on frequency domain BSS or beamforming. The schematic diagram of the system is

shown in Figure 6.1.

For the localization of the sources two fully calibrated colour video cameras are used to determine the approximate positions of the speakers. Both cameras are calibrated by the Tsai calibration (non-coplanar) technique [75] and synchronized by the external hardware trigger module and frames are captured at the rate of $f_v = 25$ frames/sec, which means $T_v = 1/25$sec. The face of each speaker is extracted in the images of both cameras to find the position of each speaker i at each state (time) k. In each image frame, the face can be extracted by the state-of-the-art Viola-Jones face detector [78]. It is highlighted that for this proof of concept work it is assumed that the full face of a speaker is clearly visible and a simple geometric visual cue, i.e. the center of the face is available. The machine cocktail party problem is very challenging and this work is only to approach the ability of a human to solve this task. It is easy to contrive situations where a human would fail in this task and these are beyond the scope of this work. Further details are in Subsection 6.4.2.

It is common in many science and engineering situations to estimate the hidden state of a system that changes over time using a sequence of noisy observations made on the system. Normally, the state-space approach, which focuses attention on the state vector of the system, is adopted for modeling a dynamic system. In this approach the 3-D location of each speaker is estimated by using the Bayesian multispeaker state space approach. The 3-D multispeaker observation is defined as $\mathbf{Z}_{1:k} = \{\mathbf{Z}_{1,1:k}, ..., \mathbf{Z}_{n,1:k}\}$ where $\mathbf{Z}_{i,1:k}$ represents the observations of speaker i and the multispeaker state configuration is defined as $\mathbf{X}_{1:k} = \{\mathbf{X}_{1,1:k}, ..., \mathbf{X}_{n,1:k}\}$. The filtering distribution of states given ob-

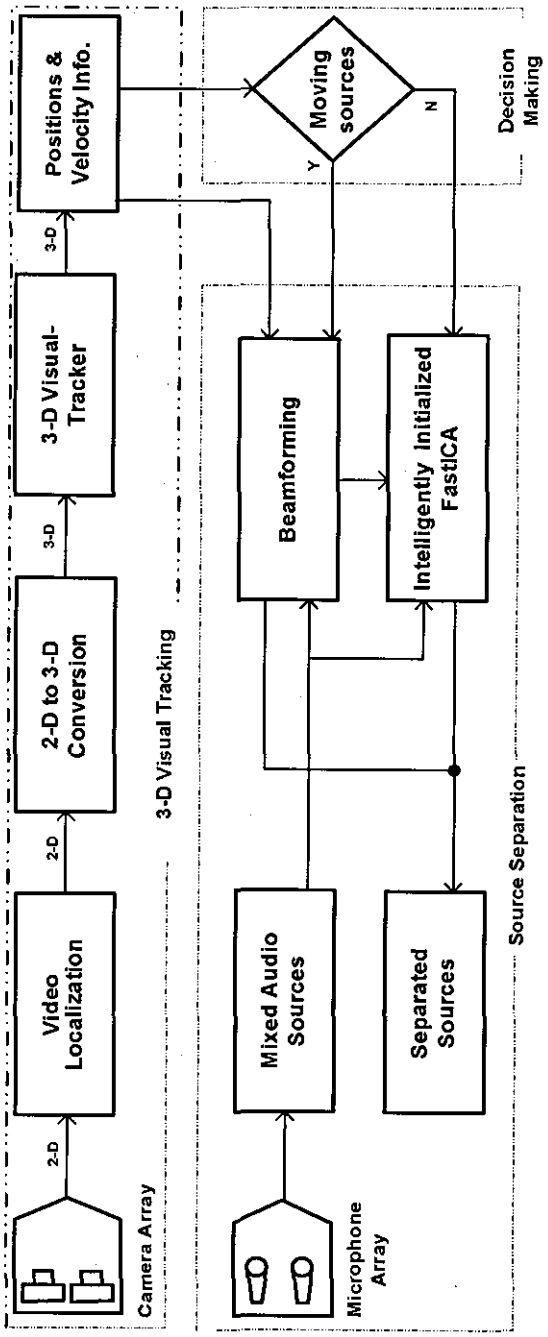

Figure 6.1. System block diagram: Video localization is based on state-of-the-art Viola-Jones face detector [78], two fully calibrated colour video cameras are used to determine the approximate 2-D positions of the speakers. The 2-D image information of the two video cameras is converted to 3-D world co-ordinates through the calibration parameters and optimization method. The approximated 3-D locations are fed to the visual-tracker, and on the basis of estimated 3-D real world position and velocity from the tracking, the sources are separated either by beamforming or by intelligently initializing the FastICA algorithm.

servations $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ is recursively approximated using a Markov Chain Monte Carlo (MCMC) particle filter and the algorithm is explained in Section 6.4.

After estimating the 3-D position of each speaker the velocity information is extracted, if the sources are physically stationary for a certain period T_k, then the positions of the speakers are incorporated within the Intelligently Initialized FastICA (IIFastICA) algorithm otherwise they are used within the beamformer to obtain the source separation for stationary or moving sources. The details of the beamformer and IIFastICA are explained in Section 6.5. The 3-D visual tracker including state model, measurement model and sampling mechanism is explained in the following section.

6.4 3-D visual tracker

The most suitable candidate for a 3-D multispeaker visual tracker is a particle filter because the probabilistic state-space formulation (non-Gaussian) and the requirement for the update of information on receipt of new measurements are ideally suitable for the Bayesian approach, which provides a rigorous general framework for dynamic state estimation problems. In the Bayesian approach to stochastic state estimation, the idea is to construct the posterior probability density function (pdf) of the state based on all the available information, including the received observations. Since such a pdf contains all the available statistical information, it can be considered to be the complete solution to the estimation problem.

For many problems, some sort of recursive processing is required in that at each time an observation is received, an estimate is required

based on that observation. This may be achieved by the use of a recursive filter. Essentially, such a filter comprises prediction and update stages. During the prediction stage, the state pdf is predicted using the state model. Since the state is usually subject to some unknown disturbances (modelled as random noise), prediction generally deforms the state pdf. The predicted pdf, resulting from the prediction stage, is modified by the latest observation during the update stage. The update operation is achieved through Bayes' rule. The advantage of this recursive filtering is that the received data can be processed sequentially rather than as a batch. The posterior density $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ is recursively calculated by Bayes' rule according to:

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) \propto p(\mathbf{Z}_k|\mathbf{X}_k) \int p(\mathbf{X}_k|\mathbf{X}_{k-1}) p(\mathbf{X}_{k-1}|\mathbf{Z}_{1:k-1}) d\mathbf{X}_{k-1} \quad (6.4.1)$$

where $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ denotes the multispeaker state model and $p(\mathbf{Z}_k|\mathbf{X}_k)$ represents the multispeaker measurement model. In general, no closed-form solution exists for (6.4.1) although these recursions can be approximated by Monte Carlo simulations of a set of particles having associated discrete probability mass and the generic particle filter is described in Section 2.5. A particle filter recursively approximates the filtering distribution $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ by a weighted set of N_p particles at time k, $\left\{\mathbf{X}_k^n, \omega_k^n\right\}_{n=1}^{N_p}$, by using the weighted particles at the previous time-step $k-1$, $\left\{\mathbf{X}_{k-1}^n, \omega_{k-1}^n\right\}_{n=1}^{N_p}$, and the new update will be

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) \approx K^{-1} p(\mathbf{Z}_k|\mathbf{X}_k) \sum_{n=1}^{N_p} \omega_{k-1}^n p(\mathbf{X}_k|\mathbf{X}_{k-1}^n) \quad (6.4.2)$$

where K is a normalization constant, $(.)^n$ refers to the n_{th} particle,

and N_p is the number of particles, so that a discrete approximation of the true posterior can be calculated. As N_p approaches infinity, this discrete approximation can converge to actual distribution depending on the sampling mechanism, discussed in the sequel.

The three important items of the probabilistic multispeaker 3-D visual tracker, the state model, the measuement likelihood model and the MCMC-sampling mechanism are formulated in the following three subsections.

6.4.1 State model

There are several state models that can be used to represent the state transition. In [124] the random walk model is used, another model which is shown to work well, to represent the time-varying location of a speaker in a typical room [124,125], is the Langevin model [126], also used in [34,40,43]. The motion of the speakers in each of the Cartesian coordinates is assumed to be independent in this state model. In the $x-$coordinate this motion is described as:

$$\begin{aligned} \dot{x}_k &= a_x \dot{x}_{k-1} + b_x F_x \\ x_k &= x_{k-1} + \Delta T \dot{x}_k \\ a_x &= e^{-\beta_x \Delta T} \\ b_x &= v_x \sqrt{1 - a_x^2} \end{aligned} \tag{6.4.3}$$

where the thermal excitation process F_x is a normally distributed random variable i.e. $\mathcal{N}(0,1)$, and $\Delta T = 1/f_v$. The other model parameters suggested by [34] are $\beta_x = 10 sec^{-1}$, and $v_x = 100 cm sec^{-1}$. The

dynamics and parameters for the other Cartesian coordinates are the same.

The above state model which includes independent single speaker dynamics is formulated for the multispeaker state model as:

$$p(\mathbf{X}_k|\mathbf{X}_{k-1}) \propto \prod_i p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1}) \tag{6.4.4}$$

where $p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1})$ denotes the dynamics for speaker i. It is highlighted that $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ can be factorized for individual speaker.

6.4.2 Measurement model

Visual measurements used in this work are based on the Viola-Jones face detector. The Viola-Jones face detector [78] yields good performance and detects faces extremely rapidly, by using a boosted cascade of features. It is a cascade of strong classifiers, each slightly more complex than the last. The input images are sub-sampled at multiple scales and locations to form the sub-windows for the faces to be detected. Face detection is performed in three stages. Initially, to minimize the effect of illuminations, the variance of all sub-windows are normalized. Secondly, the cascade of classifiers makes a decision based on the sub-window. Finally, to merge the overlapping face candidates around each face and output the final results the post processing method is used. A sub-window is detected as a face if it successfully passed all strong classifiers. If any classifier fails a sub-window then no further processing is required on that window, detailed formulation is available in [78].

The center of the detected face is determined as the approximate position of the lips of the speaker in image coordinates $\boldsymbol{\iota}^c_{i,k} = [x_{i,k}, y_{i,k}]^T$,

where c represents the number of cameras $c = 1, 2$. In 3-D space each point in each camera frame defines a ray. Intersection of both rays is found by optimization methods, which finally help in calculation of the positions $\mathbf{Z}_{i,k}$ of the speakers in 3-D real world coordinates [79].

The multispeaker measurement model can be factorized in terms of individual speakers as:

$$p(\mathbf{Z}_k|\mathbf{X}_k) = \prod_i p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}) \tag{6.4.5}$$

where $p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k})$ is the observation model for speaker i and is calculated as:

$$p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}) \propto \exp\left(-\frac{||\mathbf{Z}_{i,k} - \mathbf{X}'_{i,k}||^2}{2\sigma^2}\right) \tag{6.4.6}$$

where $\mathbf{X}'_{i,k}$ denotes a vector formed from the 3-D position components of the state vector and σ is a standard deviation parameter chosen empirically, typically unity.

6.4.3 MCMC-sampling mechanism

In the 1990s MCMC-based methods attracted great attention among researchers in the Bayesian community [127]. The advantage over alternative approaches is in the capacity to work with a high-dimensional space and complex models. It is computationally infeasible to track multiple objects in the high dimensional space by using an importance sampling [72] based traditional particle filter [128]. In tracking the MCMC sampling is a methodology for generating samples from a Markov chain whose stationary distribution corresponds to a filtering distribution. In order to efficiently place samples as close as possible to regions of high likelihood and approximate $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ in (6.4.2) with

MCMC techniques, it is important to specifically design a Metropolis-Hastings (MH) sampler (also known as MCMC sampler) at each time step [39, 129, 130]. After running the MCMC sampler for long enough at each time step the initial part of the run, called the burn in period, is discarded to achieve a stationary distribution [131]. The key to the efficiency of the MCMC algorithm rests in the proposal distribution (discussed in the sequel), in which the configuration of one single object is modified at each step of the Markov chain, and each move in the chain is accepted or rejected by the so called acceptance ratio α. The MCMC-based tracking algorithm is summarized as follows:

- Initialize the MCMC sampler: At time k predict the state of each speaker i for N_p particles i.e. $\left\{\mathbf{X}_{i,k}^n\right\}_{n=1}^{N_p}$ from the particle set at time $k-1$ i.e. $\left\{\mathbf{X}_{i,k-1}^n\right\}_{n=1}^{N_p}$ based on the factorized dynamic model $\prod_i p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1}^n)$.

- $B + N_p$ MCMC Sampling Steps: B and N_p denote the number of particles in the burn-in period and fair sample sets respectively.

  - Randomly select a speaker i from all speakers. This will be the speaker proposed to move.

  - Sample a new state $\mathbf{X}_{i,k}^*$ for only speaker i from the single speaker proposal density $Q(\mathbf{X}_{i,k}^*|\mathbf{X}_{i,k})$.

  - Compute the acceptance ratio which involves (6.4.1) for the evaluation of likelihood for only speaker i:

$$\alpha = min\left\{1, \frac{p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}^*)Q(\mathbf{X}_{i,k}|\mathbf{X}_{i,k}^*)p(\mathbf{X}_{i,k}^*|\mathbf{X}_{i,k-1}^n)}{p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k})Q(\mathbf{X}_{i,k}^*|\mathbf{X}_{i,k})p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1}^n)}\right\} \tag{6.4.7}$$

  - Draw $\mu \sim U(u|0,1)$.

- If $\alpha > \mu$ then accept the move for speaker i and change the $\mathbf{X}^*_{i,k}$ into $\mathbf{X}_k$. Otherwise, reject the move, do not change $\mathbf{X}_k$ and copy to the new sample set.

- Discard the first B samples to form the particle set, $\{\mathbf{X}^n_k\}_{n=1}^{N_p}$, at time step k.

The output of the 3-D tracker at each state k is the mean estimate for each speaker i and is calculated as the weighted sum over the associated particles as $\hat{\mathbf{x}}_{i,k} = \frac{\sum_{n=1}^{N_p} w^n_{i,k} \mathbf{x}^n_{i,k}}{\sum_{n=1}^{N_p} w^n_{i,k}}$ where in this work as in [39] $w^n_{i,k} = 1/N_p$. The change in the position of a speaker with respect to the previous state $k-1$ (known as velocity information) also plays a critical role to decide the method for source separation and is discussed next.

Discussion on algorithm choice

1. State model (6.4.4) and measurement model (6.4.5) are independent in terms of the speakers and therefore can be factorized into the product of the marginal models for each speaker. In tele-conferencing applications within an intelligent office, physical separation in speakers is always likely to be possible as the speakers are unlikely to embrace each other and speakers are also clearly separable in the 3-D real world coordinates used in this thesis due in part to the height of the cameras. Therefore, in this work, there is no requirement to incorporate interaction cues in the state model. However, the state model could be extended with an interaction term in future work as in [129].

2. Independent particle filters for all speakers and an MCMC-PF

are applicable for the requirement of the work in this thesis. Independent SIR-PF for each speaker is the most optimal choice and is already used in our work [124]. Joint particle filtering suffers from exponential complexity in the number of targets to be tracked [129]. In [130] it is also mentioned that the joint filter is not feasible in practice for more than three targets.

3. Due to the limitation of importance sampling in high dimensional state space, Markov Chain Monte Carlo (MCMC) methods are used. The MCMC method used in this work is based on [129,130] and has the appealing property that *"the filter behaves as a set of individual particle filters when the targets are not interacting, but efficiently deals with complicated interactions when targets approach each other"*. The design of proposal density plays an important role in the success of an MCMC algorithm. In [130] a *"One target at a time"* scheme is implemented. The proposal density used is also defined in [39] as

$$Q(\mathbf{X}_k^*|\mathbf{X}_k) = \sum_{i^*} Q(i^*)Q(\mathbf{X}_k^*|\mathbf{X}_k, i^*) \tag{6.4.8}$$

where a single speaker is first chosen with probability $Q(i^* = i)$ and a move is attempted on i (shown in the algorithm summary) and the rest of the multi-speaker configuration is left unchanged, where

$$Q(\mathbf{X}_k^*|\mathbf{X}_k, i^*) \propto \frac{1}{N_p} \sum_n p(\mathbf{X}_{i^*,k}^*|\mathbf{X}_{i^*,k-1}^n) \prod_{l \in m-\{i^*\}} \delta(\mathbf{X}_{l,k}^* - \mathbf{X}_{l,k}) \tag{6.4.9}$$

and $\mathbf{X}_k$ represents the whole state for all speakers m, and $\mathbf{X}_{i^*,k}$

denotes the sub state for one speaker.

4. The proposal density used in [39] appears not to be properly formulated since it is not properly normalized. This has thus been modified in (6.4.9). An alternative interpretation of the sampler using an auxiliary variables approach can fix the problem, and then does indeed lead to the acceptance ratio of (6.4.7). Further information can be found in Berzuini et al 1997, Golightly and Wilkinson 2005 and [132]. In realization of the MCMC-PF it is very important to avoid having degenerate sampling over the joint target distribution implied by the auxiliary variables approach. A simple additional Gibbs sampling step that moves just the previous state $\mathbf{X}_{k-1}$ can be added at each iteration, as described in [133], to ensure that degeneracy is avoided.

5. Tracking results for independent particle filters for all speakers and an MCMC-PF are provided in Section-6.6. The numbers of particles used are 1000 and 600+200 respectively for independent SIR-PFs and MCMC-PF. Results show no significant difference (based on Euclidean error) but MCMC-PF reduces the computational complexity in multispeaker tracking. More advanced algorithms with detailed discussion are presented in [132, 134].

6.5 Source separation

The audio mixtures from the microphone sensor array are separated with the help of visual information from the 3-D tracker. On the basis of this visual information it is decided either the sources should be separated as moving or stationary. The pseudo code to issue the command

for selecting the source separation methods is as follows.

Pseudo Code: Command for Selecting the Source Separation Methods

- *Reset the counter and set the threshold*
- *FOR* $j = 2 : k$

  -Find $d_{i,j} = ||\hat{\mathbf{X}}_{i,j} - \hat{\mathbf{X}}_{i,j-1}||_2$
- *IF* $d_{i,j} < threshold$

  -Update the counter

  **IF counter* $> T_k/T_v$

  -Command for the FastICA based method

  **END IF*

  -Set $\hat{\mathbf{X}}_{i,j} = \hat{\mathbf{X}}_{i,j-1}$
- *ELSE*

  -Command for the beamforming based method

  -Reset the counter
- *END IF*
- *END FOR*

  THIS CODE WOULD BE USED FOR EACH SPEAKER i.

---

where T_k represents the expected stationary period for the speakers, $T_v = 1/f_v$, $||.||_2$ denotes Euclidean norm, and *threshold* is the minimum distance required for the beamformer channel filters, which should be recalculated to separate the sources.

When the sources are physically stationary for a certain period T_k the sources are separated with IIFastICA. By changing the value of T_k

the expected required stationary period for the sources can be changed.

The other important parameter to be calculated before starting the source separation is the angle of arrival of each speaker to the sensor array. By having the position information of the microphones and the speakers at each state k from the 3-D visual tracker the angle of arrival $\boldsymbol{\theta}_{i,k}$ of speakers relative to the microphone sensor array can be easily calculated.

With $\boldsymbol{\theta}_{i,k}$ and the control command from the above decision criterion at each state k, the sources are separated either by beamforming or by IIFastICA as discussed in the following subsections.

6.5.1 Beamforming based separation

A simple two set beamforming system configuration is shown in Figure 6.2. The equivalence between frequency domain blind source separation and frequency domain adaptive beamforming (ABF) is already studied in [100]. In the case of a two microphone sensor array an ABF creates only one null towards the jammer. Since the aim is to separate two source signals s_1 and s_2 therefore two sets of ABFs are presented in Figure 6.2. An ABF by using filter coefficients $\mathbf{w}_{21}$ and $\mathbf{w}_{22}$ forms a null directive patterns towards source s_1 and by using filter coefficients $\mathbf{w}_{11}$ and $\mathbf{w}_{12}$ forms a null directive patterns towards source s_2. If two speakers are located at the same direction with different distances, it is not possible to separate the sources by phase difference. One of the other limitation for the blind source separation is acoustic environment with the long reverberations, in this work the reverberation time in the intelligent office is 130msec which will be considered as a fairly moderate reverberant environment.

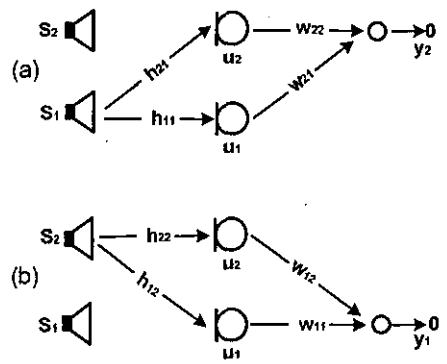

Figure 6.2. Two set beamforming system configuration: (a) Beamformer for target s_2 and jammer s_1 (b) Beamformer for target s_1 and jammer s_2

In the intelligent office where the recordings are taken the microphones used are uni-directional. By using a short-time discrete Fourier transform (DFT) the mixing process can be formulated as follows: having M statistically independent real sources $\mathbf{s}(\omega) = [s_1(\omega), \ldots, s_M(\omega)]^H$ where ω denotes discrete normalized frequency, a multichannel FIR filter $\mathbf{H}(\omega)$ producing N observed mixed signals $\mathbf{u}(\omega) = [u_1(\omega), \ldots, u_N(\omega)]^H$, where $(.)^H$ is Hermitian transpose, can be described as (it is assumed that there is no noise or noise can be deemed as a source signal in the model for simplicity)

$$\mathbf{u}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) \tag{6.5.1}$$

and the source separation can be described as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{u}(\omega) \tag{6.5.2}$$

$\mathbf{y}(\omega) = [y_1(\omega), \ldots, y_N(\omega)]^H$ contains the estimated sources, and $\mathbf{W}(w)$ is the unmixing filter matrix. An inverse short time Fourier transform is then used to find the estimated sources $\hat{\mathbf{s}}(t) = \mathbf{y}(t)$. In this work to

demonstrate the proposed method the exactly determined convolutive BSS problem i.e. $N = M = 2$.

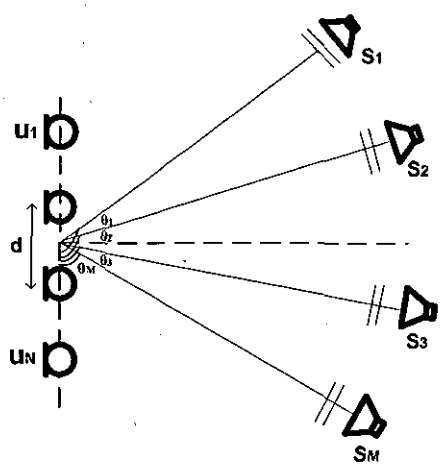

Figure 6.3. Microphone and source layout

The delay element between source l and sensor k i.e. $h_{kl}(\omega)$ is calculated by using the angle of arrival information obtained from tracking.

$$h_{kl}(\omega) = e^{j(k-l)d\cos(\theta_l)\omega/c} \quad k = 1, ..., N \quad l = 1, ..., M \tag{6.5.3}$$

where d is the distance between the sensors and c is the speed of sound in air. Then $\mathbf{H}(\omega)$ is formulated as:

$$\mathbf{H}(\omega) = \begin{bmatrix} h_{11}(\omega) \cdots h_{1M}(\omega) \\ \vdots \quad \vdots \\ h_{N1}(\omega) \cdots h_{NM}(\omega) \end{bmatrix} \tag{6.5.4}$$

Ideally, $h_{kl}(\omega)$ should be the sum of all echo paths, but these cannot all be found, therefore an approximation is used by neglecting the room reverberations.

The unmixing matrix $\mathbf{W}(\omega)$ for each frequency bin can be approxi-

mated from beamforming methods. In recent years, many beamforming methods have been proposed, such as the linearly constrained minimum variance (LCMV) method and the minimum variance distortionless response (MVDR) method [135]. A post filtering approach has also been utilized to improve these methods [136]. However, the LCMV method and the MVDR method need estimates of statistical information of the input or noise signals, which are not accessible in the context of BSS for moving sources e.g. for signals with length equal to $0.4sec$, FFT block length $T = 2048$ and sampling frequency $f_a = 8KHz$, only one sample would be available at each frequency bin $round(0.4f_a/T) = 1$ therefore it is not possible to calculate the covariance matrix required in MVDR method and LCMV method. Furthermore, a diffuse noise field assumption is used in [136], which is not valid in the context of BSS. The post filter method has been used in [29] to perform speech enhancement for both stationary and moving cases, however, the model used in that speech enhancement is a single-input-single-output (SISO) model, which is different from the multi-input-multi-output (MIMO) model in the context of BSS, thus this post filtering approach is also not suitable for the solution of BSS. To the best of my knowledge, in the context of BSS, only the beamforming approach that is directly obtained from the inverse of the mixing matrix model has been successfully used in [137]. To compensate for noninvertibility of the mixing matrix, a regularization term is included in [18], in which the beam pattern obtained from the geometric information is incorporated in the solution of BSS. Similar to that in [18], the unmixing matrix in this method is calculated as:

$$\mathbf{W}(\omega) = (\mathbf{H}(\omega)^H \mathbf{H}(\omega) + \beta I)^{-1} \mathbf{H}(\omega)^H \tag{6.5.5}$$

where $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), ..., \mathbf{w}_N(\omega)]$, $\mathbf{H}(\omega) = [\mathbf{h}_1(\omega), ..., \mathbf{h}_M(\omega)]$, β is a small positive constant such as 0.01 in the simulations, and I represents the identity matrix.

Finally, by placing $\mathbf{W}(\omega)$ in (6.5.2) the sources are estimated. Since the scaling is not a major issue [66] and there is no permutation problem, therefore the estimated sources can be aligned for reconstruction in the time domain.

6.5.2 FastICA based separation

If the sources are stationary for at least two seconds, the sources are extracted with the help of the estimated $\mathbf{H}(\omega)$ from the above section and the whitening matrix for the mixtures, as an initialization of the FastICA algorithm [96]. This approach thereby improves the convergence of the algorithm and also increases the separation performance whilst mitigating the permutation problem. Crucially, in the frequency domain convolutive BSS (FDCBSS) approach, since the algorithm essentially fixes the permutation at each frequency bin, there will be no problem while aligning the estimated sources for reconstruction in the time domain.

Each column of $\mathbf{H}(\omega)$ (6.5.1) is used to initialize the FastICA (presented in Chapter-4) for each frequency bin.

$$\mathbf{w}_i(\omega) = \mathbf{Q}(\omega)\mathbf{h}_i(\omega) \tag{6.5.6}$$

Summary Table: Implementation steps for the proposed method

- *Calibrate the video cameras and calculate calibration parameters.*
- *Detect the face region in the synchronized frames of both cameras.*
- *Find the positions of the lips of each speaker from the face regions in synchronized video frames and calculate the observation of each speakers* $\mathbf{Z}_{i,k}$ *at each state* k *in 3-D real world coordinates.*
- *Implement the 3-D visual tracker and find the estimated position of each speaker* $\mathbf{X}_{i,k}$ *at each state* k.
- *Calculate angle of arrivals* $\boldsymbol{\theta}_{i,k}$ *to the sensor array and check the sources are stationary or moving, important parameters are threshold,* T_v *and* T_k.
- *Incorporate the visual information in (6.5.3) and separate the sources accordingly either by beamforming or by the FastICA, important parameters are* N, M, d, c, $\boldsymbol{\theta}$, b, Q, T.

6.6 Experiments and results

6.6.1 Data collection

The simulations are performed on real recorded audio-visual signals generated from a room geometry as illustrated in Figure 6.4. Data are collected in a 4.6 x 3.5 x 2.5 m^3 intelligent office. Two out of eight calibrated colour video cameras ($C1$ and $C2$ shown in Figure 6.4) are utilized to collect the video data. Video cameras are fully synchronized with an external hardware trigger module and frames are captured at $f_v = 25$Hz with an image size of 640x480 pixels, frames were down-scaled if it was necessary, and reducing the resolution by half was a good tradeoff between accuracy and resolution. Both video cameras have overlapping fields of view. The duration between consecutive states is $T_v = 1/25sec$. Audio recordings are taken at $f_a = 8$kHz and are synchronized manually with video recordings. The distance between the audio-sensors is $d = 4cm$. The other important variables are selected as: number of sensors and speakers $N = M = 2$, number of particles $N_p = 600$, $B = 200$, the number of images in the first and second experiment are $k = 525$ & 600, which respectively indicate 21 & $24secs$ of data, $T_k = 5sec$, $threshold = 0.04m$, FFT length $T = 2048$ and filter length $Q = 1024$, height of the cameras in the intelligent office is $2.35m$, and the room impulse duration is 130ms. In the proposed algorithm the non-linearity for FastICA is $G(y) = \log(b+y)$, with $b = 0.1$. In the first experiment on tracking, speaker 2 is stationary and speaker 1 is moving and in the second experiment both speakers are moving around a table in a tele-conference scenario.

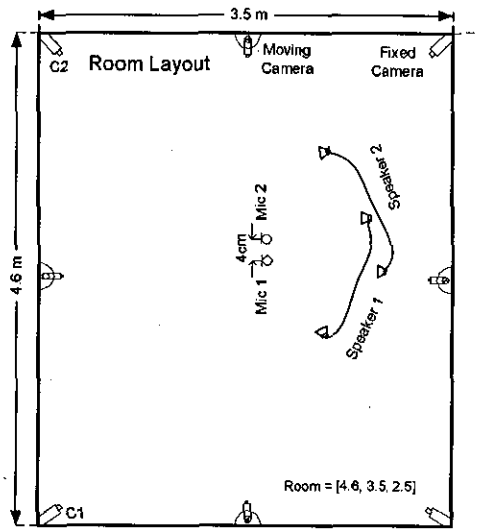

Figure 6.4. A two-speaker two-microphone layout for recording within a reverberant (room) environment. Room impulse response length is 130 ms.

6.6.2 Results and discussion

3-D tracking results

In this section the results obtained from tracking are discussed. Two experiments are performed to evaluate the 3-D visual tracker. The faces of the speakers are detected by using the Viola-Jones face detector [78] which efficiently detected the faces in the frames shown in Figure 6.5. Since in the dense environment as shown in Figure 6.5 it is very hard to detect the lips directly, therefore the center of the detected face region as the position of the lips in each sequence is approximated. More sophisticated and computationally efficient schemes could also be proposed for detecting the face through a sequence of images but the approach adopted in this chapter is sufficient to verify the multimodal CBSS method, the target of this work.

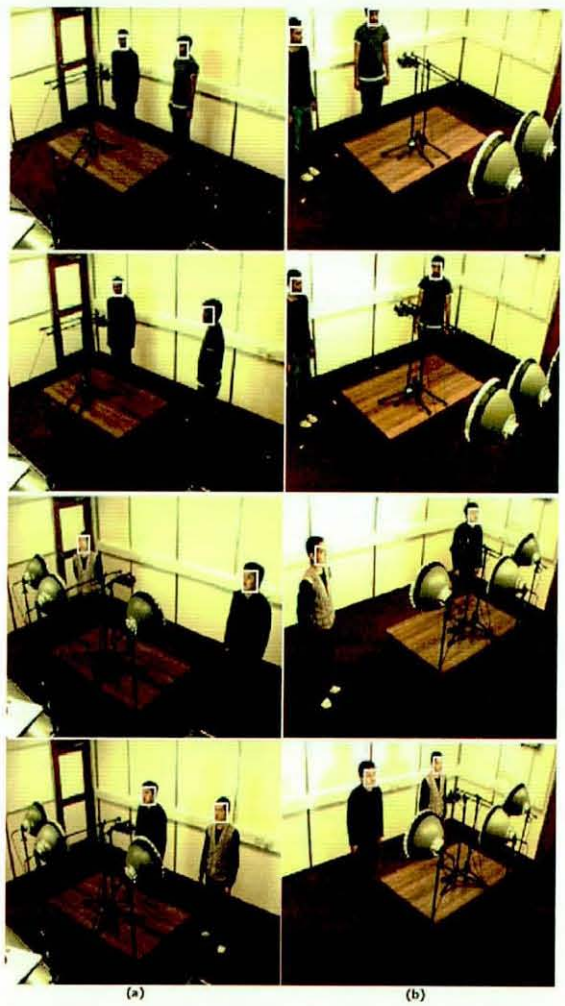

Figure 6.5. 3-D Tracking results 1: frames of synchronized recordings, (a) frames of first camera and (b) frames of second camera; the Viola-Jones face detector [78] efficiently detected the faces in the frames.

The approximate 2-D position of the lips of the speaker in both synchronized camera frames at each state is converted to 3-D world coordinates by using the calibration parameters [75] and the optimization method [79]. With this measurement the particle filter is updated. The number of particles in both experiments for MCMC-PF was the

same $N_p = 600$, $B = 200$, for SIR-PF was $N_p = 1000$ and results were obtained using single runs.

In the first experiment speaker 2 is stationary and speaker 1 is moving around the table so the tracking results of the speaker 1 are discussed in detail in this experiment. The sampling importance resampling particle filer (SIR-PF) is also suitable for this case as used in our work [124]. In the second experiment both speakers are simultaneously moving and their motion is more complicated as they cross over. MCMC-PF is suitable for multispeaker tracking because it improves the sampling efficiency with approximately the same computational cost of the Generic-PF. In both experiments SIR-PF and MCMC-PF are used. The gait of the speakers is not smooth and the speakers are also stationary for a while at some points during walking around the table which provides a good test for the evaluation of the 3-D tracker as well as for source septation methods, and this is also clear in the 3-D tracking results shown in Figures 6.6, 6.7, 6.8 & 6.9.

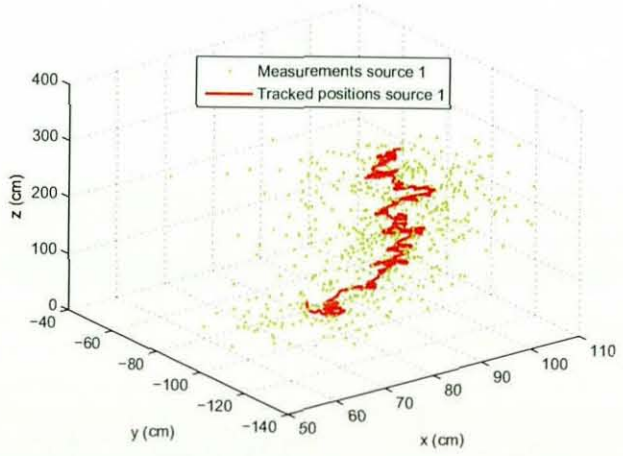

Figure 6.6. 3-D Tracking results 1: SIR-PF based 3-D tracking of speaker 1 while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment.

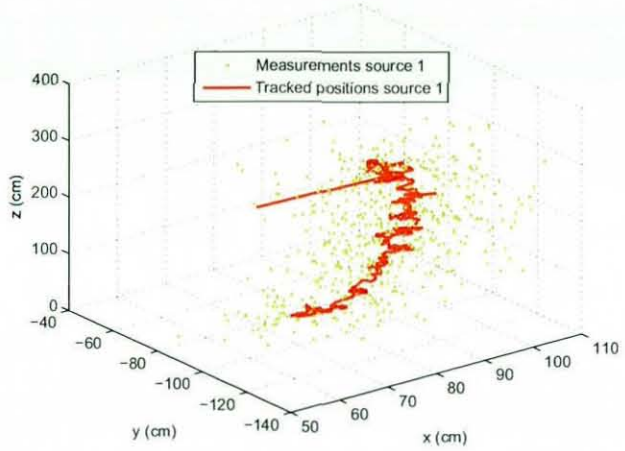

Figure 6.7. 3-D Tracking results 1: MCMC-PF based 3-D tracking of speaker 1 while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment.

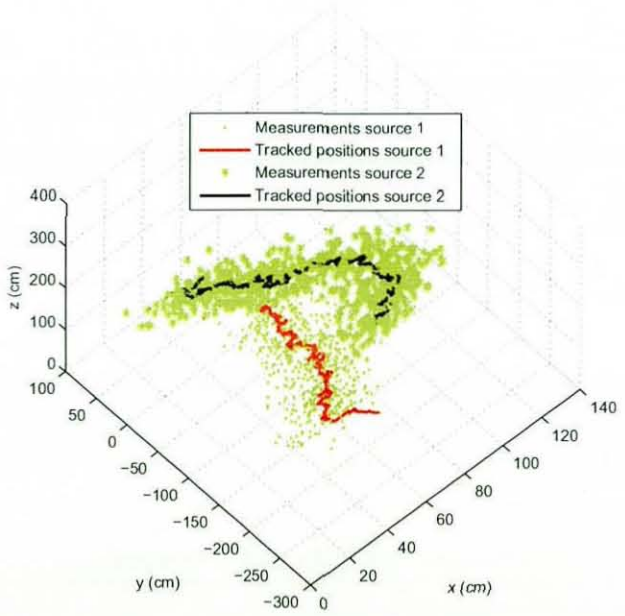

Figure 6.8. 3-D Tracking results 2: SIR-PF based 3-D tracking of the speakers while walking around the table in the intelligent office.

In order to view the tracking results in more detail, the tracking results are plotted in the xy and z axes separately. Figures 6.10, 6.10, 6.12 & 6.13 clearly show that tracking result has removed much of the measurement uncertainty and later in this section the error in detection for the particle filter will be quantified. The benefit of the true 3-

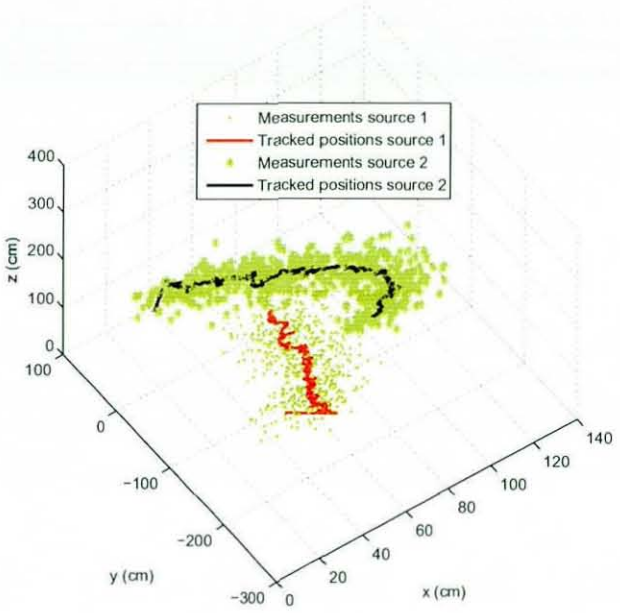

Figure 6.9. 3-D Tracking results 2: MCMC-PF based 3-D tracking of the speakers while walking around the table in the intelligent office.

D tracker is clearly shown in Figure 6.13. In particular, although the speakers would approximately coallese in the 2-D image plane, they are clearly separable in the 3-D real world coordinates due in part to the height of the cameras. In 2-D tracking in the image plane this problem cannot be avoided. The error bars for particle filters at different states are also plotted in these results. It is highlighted that the error bars would appear as 3D surfaces in the pseudo-3D plots and would make the plots cluttered if they were displayed. However, the behavior of the error ellipses on the 2D plots gives a clear indication as to how 3D error bar surfaces would appear on the 3D plots.

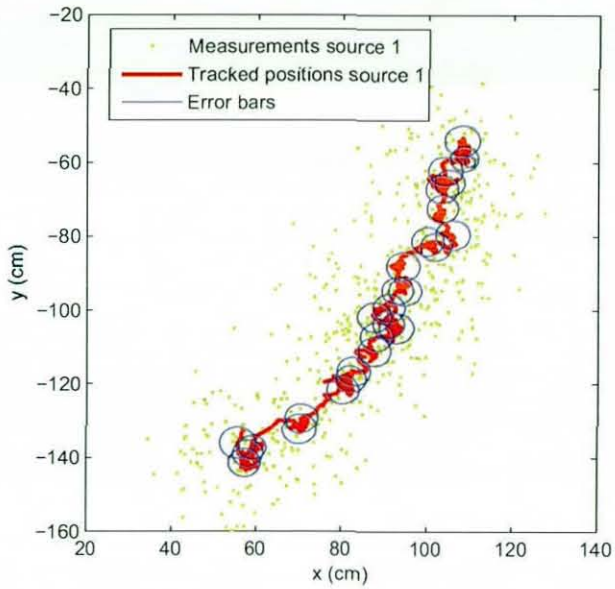

Figure 6.10. 3-D Tracking results 1: SIR-PF based tracking of the speaker 1 in the x and y-axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result provides more in depth view in the x and y-axis.

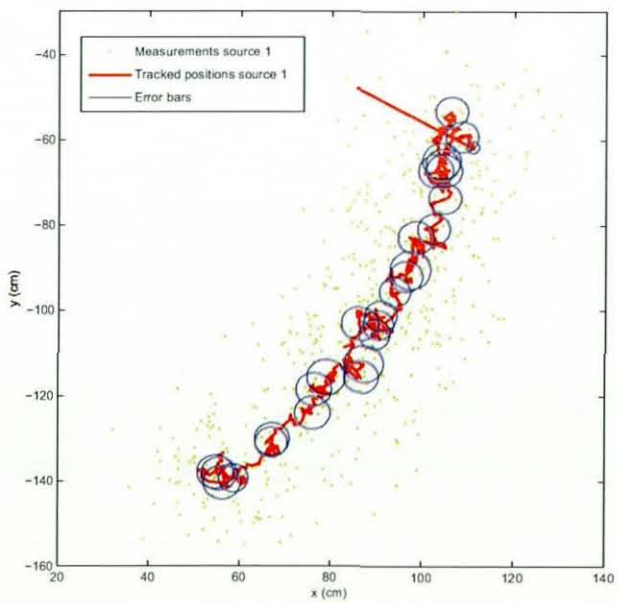

Figure 6.11. 3-D Tracking results 1: MCMC-PF based tracking of the speaker 1 in the x and y-axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result provides more in depth view in the x and y-axis.

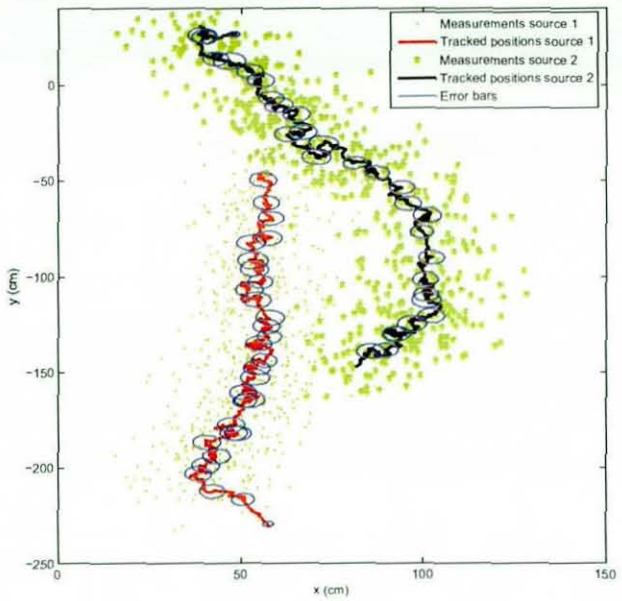

Figure 6.12. 3-D Tracking results 2: SIR-PF based tracking of the speakers in the x and y-axis, while walking around the table in the intelligent office. The result provides more in depth view in the x and y-axis.

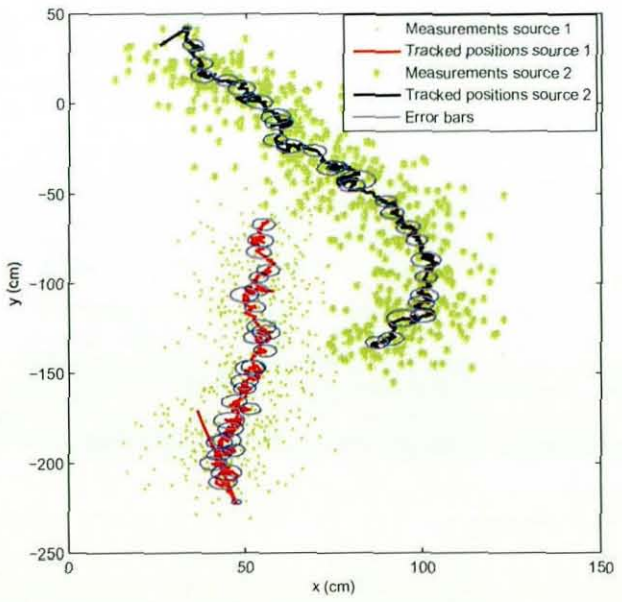

Figure 6.13. 3-D Tracking results 2: MCMC-PF based tracking of the speakers in the x and y-axis, while walking around the table in the intelligent office. The result provides more in depth view in the x and y-axis.

Actually, the height of the speakers is fixed and during walking only the movement in the heads will produce minor change which is clear in Figures 6.14, 6.15, 6.16 & 6.17. Since the speakers and microphones are approximately at the same level therefore it is assumed that effective movement is in the xy plane.

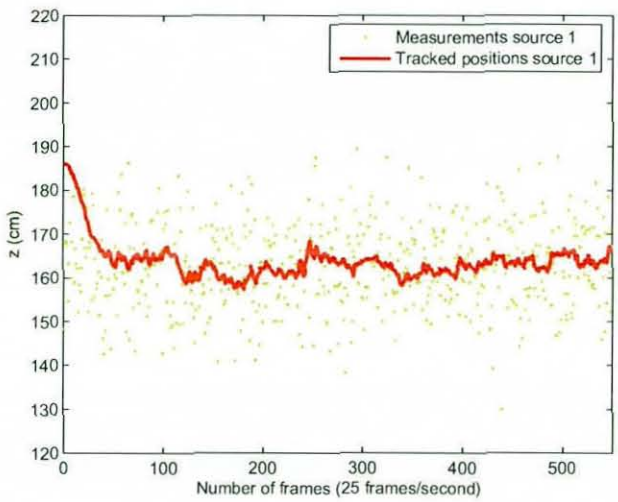

Figure 6.14. 3-D Tracking results 1: SIR-PF based tracking of the speaker 1 in the z-axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result confirms that there is very small change in the z-axis with respect to the x and y-axis.

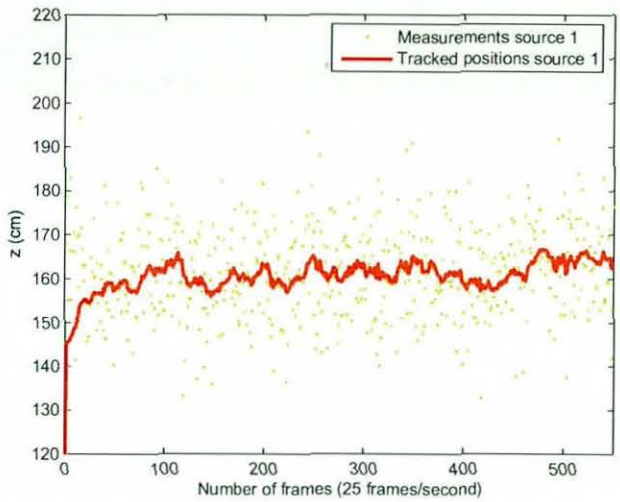

Figure 6.15. 3-D Tracking results 1: MCMC-PF based tracking of the speaker 1 in the z-axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result confirms that there is very small change in the z-axis with respect to the x and y-axis.

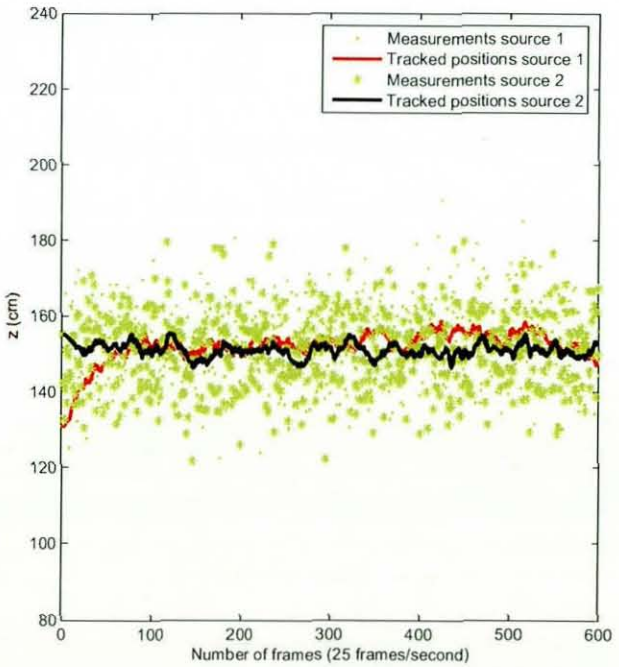

Figure 6.16. 3-D Tracking results 2: SIR-PF based tracking of the speakers in the z-axis, while walking around the table in the intelligent office. The result confirms that there is very small change in the z-axis with respect to the x and y-axis.

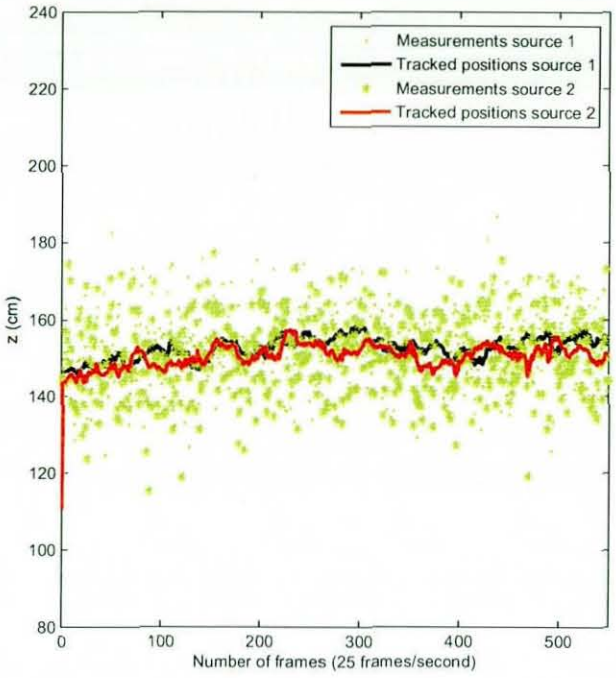

Figure 6.17. 3-D Tracking results 2: MCMC-PF based tracking of the speakers in the z-axis, while walking around the table in the intelligent office. The result confirms that there is very small change in the z-axis with respect to the x and y-axis.

In order to evaluate the performance of the tracker as in [29], the Euclidean distance to the frame-based ground truth is generated at each state. To calculate the ground truth, this time consuming manual task is performed by annotating the mouth position of each speaker in each camera frame. Figures 6.18, 6.19 6.20 & 6.21 provide the Euclidean error at each state for both experiments. In the first experiment the mean error is $0.05m$ and standard deviation is $0.03m$. In the second experiment the mean error is $0.055m$ and standard deviation is $0.032m$ which confirm the good performance of the tracker.

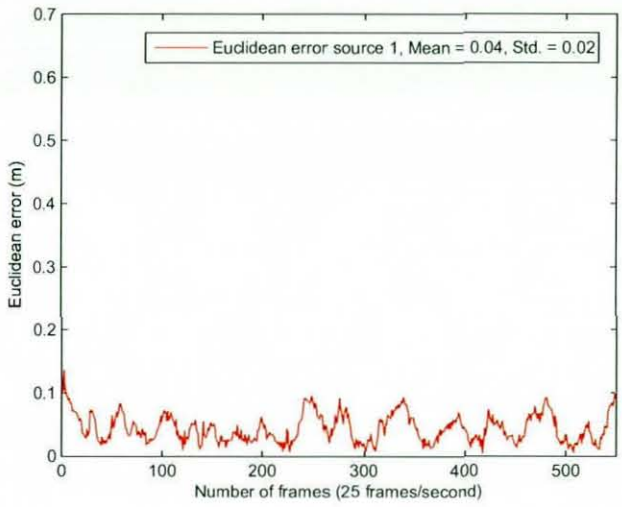

Figure 6.18. 3-D Tracking results 1: SIR-PF based tracking of the speaker 1. Speaker 2 is physically stationary. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of speaker 1.

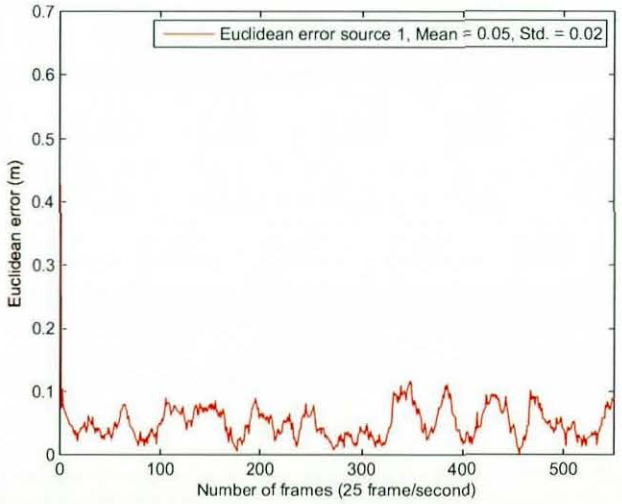

Figure 6.19. 3-D Tracking results 1: MCMC-PF based tracking of the speaker 1. Speaker 2 is physically stationary. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of speaker 1.

Angle of arrival results

The calculated position of the center of the microphones in experiment 1 is $[-0.08, -0.22, 1.62]^T m$, the position of speaker 2 is

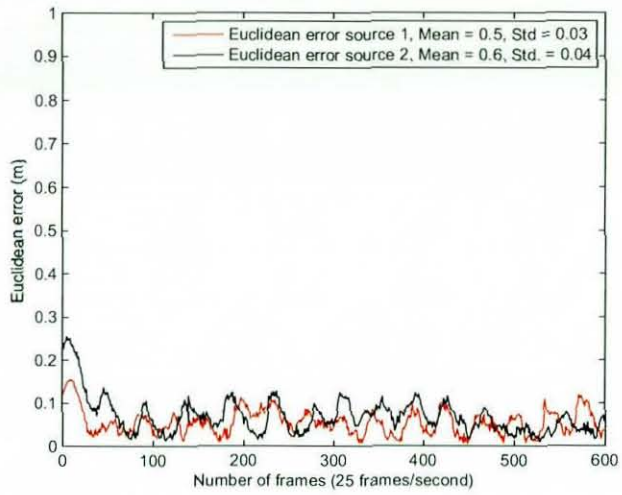

Figure 6.20. 3-D Tracking results 2: SIR-PF based tracking of the speakers. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of the speakers.

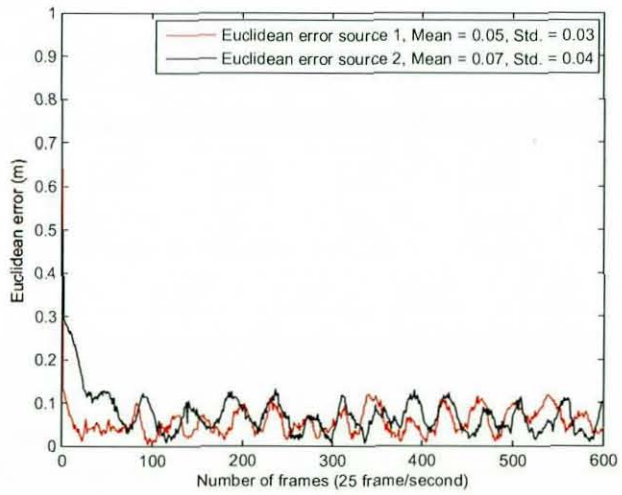

Figure 6.21. 3-D Tracking results 2: MCMC-PF based tracking of the speakers. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of the speakers.

$[0.94, 0.59, 1.63]^T m$ (the reference point in the room is under the table, close to the microphones) and the tracked position of speaker 1 in states $k = 1 : 525$ is shown in Figure 6.11. The angle of arrival of speaker 2 is 128 *degree* and the angles of arrivals of the speaker 1 are shown in Figure 6.22. The calculated position of the center of the microphones

in experiment 2 is $[-0.50, -0.94, 1.60]^T m$. The angle of arrivals of both speakers are shown in Figure 6.23. In the results of both experiments it is found that the effective movement of the speakers were in the x-axis and y-axis therefore the effective change in the angle of arrival was only in the xy plane.

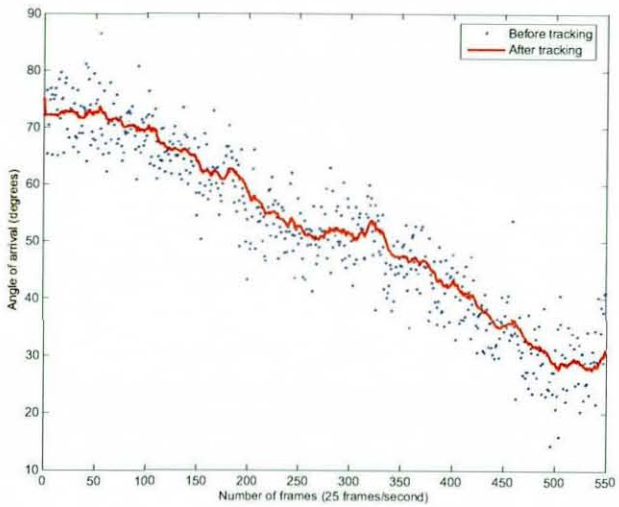

Figure 6.22. Angle of arrival results 1: Angle of arrival of speaker 1 relative to the sensor array. Speaker 2 is physically stationary in this experiment. The estimated angle before tracking and corrected angle by MCMC-PF are shown. The change in angle is not smooth because of the gait of the speaker.

Now a successful tracker is available to provide the required geometric information to perform multimodal blind source separation. Therefore simulations on BSS are discussed next.

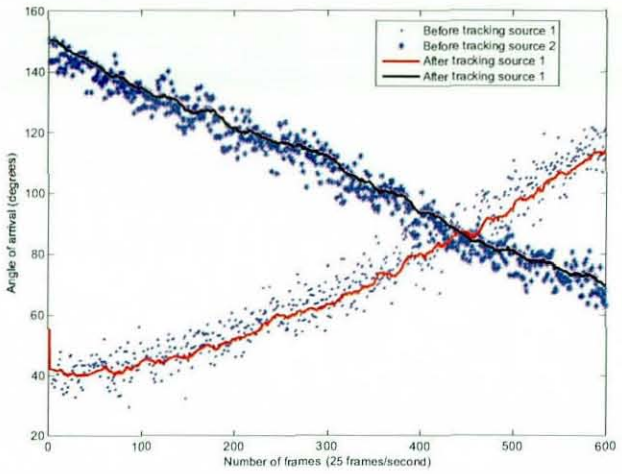

Figure 6.23. Angle of arrival results: Angle of arrival of the speakers to the sensor array. The estimated angle before tracking and corrected angle by MCMC-PF are shown.

BSS results

The objective evaluation of BSS is limited by the requirement of the mixing filter therefore for such testing the audio signals are convolved with real room impulse responses recorded in certain positions of the room. The separation of the real recorded signals in the intelligent office is evaluated subjectively by listening tests and mean opinion scores (MOSs) are provided at the end. In the context of objective evaluation, termed moving source test (MST), it is assumed that the moving sources remain static over a particular time interval less than $0.5sec.$ The justification is that over this interval no frequency domain CBSS algorithm could be used as there would be insufficient number of samples to achieve convergence, but the proposed beamforming is successful as it is independent of data length.

Five simulations for comparison of the proposed algorithm are presented

- FastICA [96](Matlab code available online) based BSS with ran-

dom initialization and length of the signals is $5sec$.

- FastICA based BSS with intelligent initialization and length of the signals is $5sec$.
- FastICA based BSS with intelligent initialization but length of the signals is $0.4sec$ (the MST case).
- Beamforming based BSS and the length of the signals is $0.4sec$ (the MST case).
- Beamforming based BSS when both sources are physically close to each other.

Initially, in the first simulation the recorded mixtures of length of 5sec are separated by the original FastICA algorithm. The performance indices and evaluation of permutation by the original FastICA algorithm [96] with random initialization are shown in Figure 6.24. It is highlighted that thirty-five iterations are required for the performance level achieved in Figure 6.24(a) with no solution for permutation as shown in Figure 6.24(b). The permutation problem in frequency domain BSS degrades the SIR to approximately zero for the recorded mixtures.

In the second simulation recorded mixtures of length of 5sec are again separated. In this simulation the angle of arrival of both speakers obtained from the 3-D tracker is passed one-by-one to (6.5.3) and FastICA is intelligently initialized (as discussed in Section 6.5.2). The resulting performance indices are shown in Figure 6.25(a) which show good performance i.e. close to zero across the majority of the frequency bins. This is due to visual information used in the initialization, and the

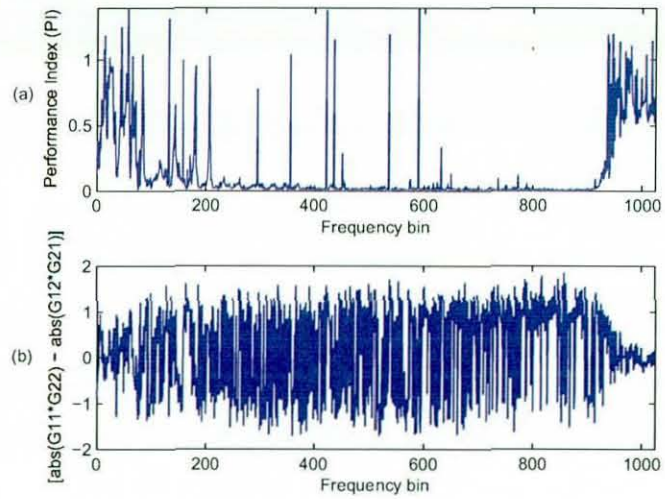

Figure 6.24. BSS Results: performance index at each frequency bin for the original Bingham and Hyvärinen algorithm on the top [96] and evaluation of permutation at the bottom, on the recorded signals of known room impulse response with fixed iteration count = 35, length of the signals is 5 seconds. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

algorithm also converges in six iterations. The visual modality therefore renders this BSS algorithm semiblind and thereby much improves the resulting performance and the rate of convergence. Permutation is evaluated on the basis of the criterion mentioned above. In Figure 6.25(b) the results confirm that the proposed algorithm automatically mitigates the permutation at each frequency bin. Since there is no permutation problem the sources are therefore finally aligned in the time domain. In Figure 6.25(a) at higher frequency bins there is less energy in the mixtures therefore performance in those bins is deteriorated. The SIR is also calculated as in [66] and results are shown in Table 6.1.

In the third simulation the length of the mixtures is reduced to $0.4sec$ i.e. the MST case, and the performance is shown in Figure 6.26. It is obvious in the results that the performance is poor because Fas-

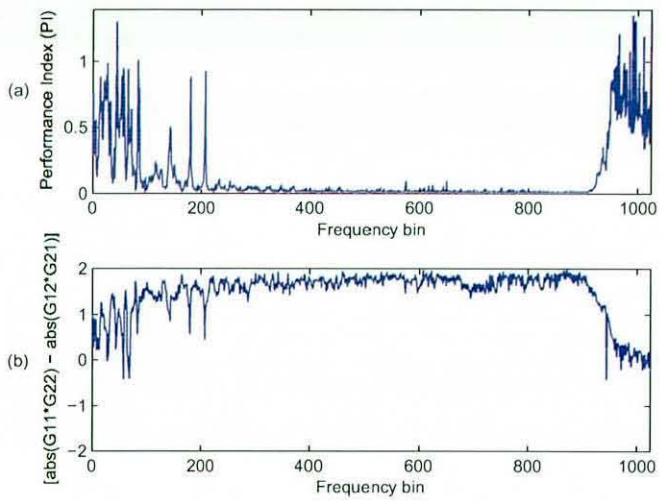

Figure 6.25. BSS Results: performance index at each frequency bin for the proposed intelligently initialized FastICA algorithm at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response with fixed iteration count = 6, length of the signals is 5 seconds. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Table 6.1. BSS Results: comparison of SIR-Improvement between algorithms and the proposed method for different sets of mixtures.

Algorithms	SIR-Improvement (dB)
Parra et al. Method [19]	6.8
Wenwu et al. Method [74]	10.0
IIFastICA	12.9

tICA is based on fourth order statistics and is limited by the data length requirement. For signals with length equal to $0.4sec$, given the block length of the FFT, only one sample would be available at each frequency bin $round(0.4f_a/T) = 1$ and therefore batch-wise BSS algorithms cannot separate the sources of short data length due to insufficient samples to converge, which is a common problem when the sources are moving.

In the fourth simulation the angles of arrival of both speakers obtained from the 3-D tracker are passed to (6.5.3) and the sources were separated by using beamforming (discussed in section 6.5.1) and the

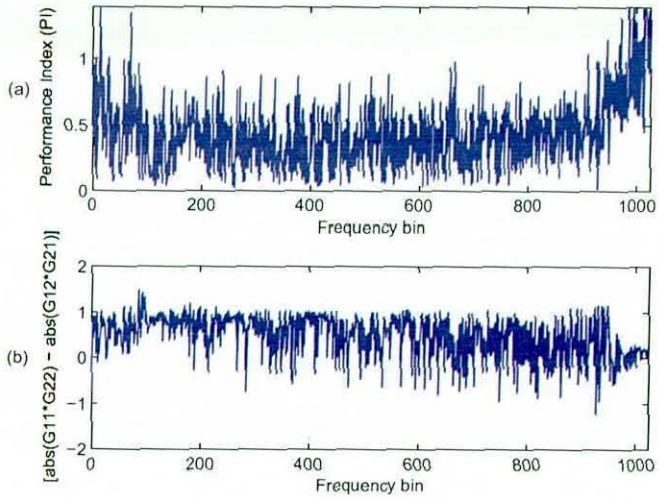

Figure 6.26. BSS Results: performance index at each frequency bin for the proposed intelligently initialized FastICA algorithm at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, length of the signals is 0.4 seconds. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

results are shown in Figure 6.27. The resulting performance indices are shown in Figure 6.27(a) and confirm good performance and Figure 6.27(b) also shows that the beamforming mitigates the permutation. Since there is no permutation problem therefore the sources can be aligned in the time domain. For comparison the data length of the mixtures used in this simulation is $0.4sec$ and SIR in this case is 9.5dB. It is know that the ideal condition for beamforming is when there is no reverberation in the room (instantaneous case), but is not possible in a real environment, however beamformer still works in a moderate reverberant environment as in this case (room impulse response length is 130 ms).

In the last simulation when both speakers are physically close to each other, i.e. at state $k = 393$ (where both speakers are close and sta-

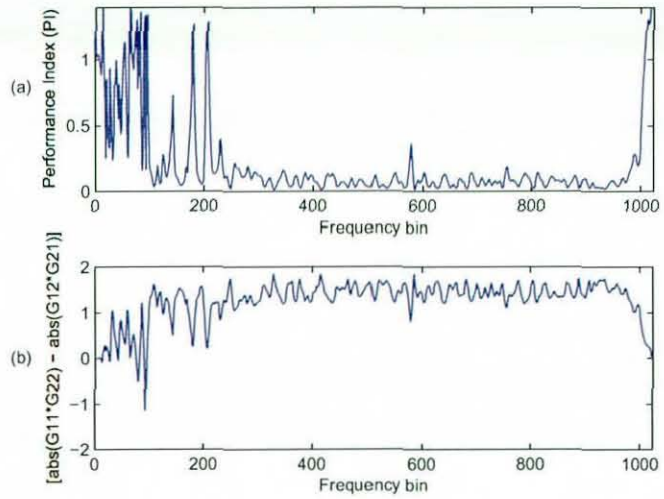

Figure 6.27. BSS *Results*: performance index at each frequency bin for 3-D tracking based angle of arrival information used in beamforming at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, beamforming based separation is independent of length of the signals. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

tionary for $0.4sec$) the position of the speaker 1 is $[0.54, -1.10, 1.59]^T m$ and the position of speaker 2 is $[1.00, -0.91, 1.58]^T m$, the angles of arrivals of both speakers i.e. 81 & $91degrees$ respectively, obtained from the above estimated positions from the 3-D tracker are passed to (6.5.3) and the sources are separated by using beamforming and the results are shown in Figure 6.28. In this case, the performance reduces because of the limitations of the beamformer, i.e. it is unable to discriminate spatially one speaker from another due to the width of its mainlobe being greater than the separation of the speakers, which is particularly clear at lower frequencies. For comparison the data length of the mixtures used in this simulation is also $0.4sec$ and SIR in this case is $8.2dB$. In conclusion, beamforming provides the solution for source separation of moving sources at an acceptable level because beamforming is inde-

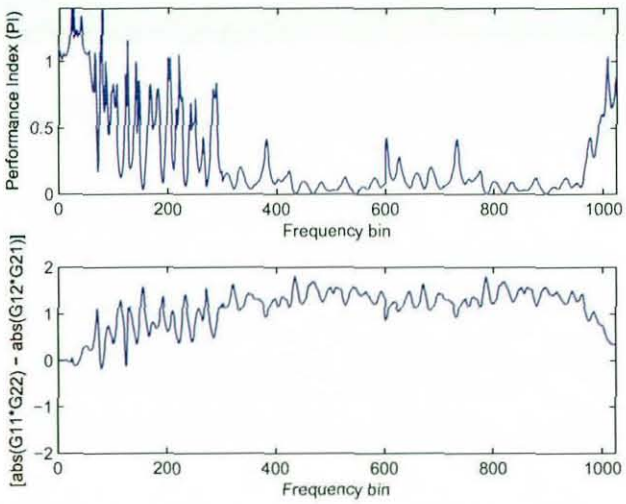

Figure 6.28. BSS Results: performance index at each frequency bin for 3-D tracking based angle of arrival information used in beamforming at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, beamforming based separation is independent of length of the signals. Speakers are physically close to each other therefore performance is reduced. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

pendent of the data length requirement unlike second or fourth order statistics based batch-wise BSS algorithms.

Finally, separation of real room recordings were evaluated subjectively by listening tests, six people participated in the listening tests and mean opinion score is provided in Table 6.2.

Table 6.2. Subjective evaluation: MOS for separation of real room recordings, by IIFastICA when sources are stationary, and by beamforming when sources are moving.

Algorithms	Mean opinion score
IIFastICA	4.7
Beamforming	3.8

6.7 Summary

In this chapter a new multimodal BSS solution was proposed to solve the moving source separation problem. A full 3-D tracker based on MCMC-PF was implemented. Video information was utilized in the 3-D tracker which provided velocity and direction information of sources. Based on the velocity of the source, a criterion for source separation was setup: a beamforming algorithm was used when sources are moving and a BSS algorithm was performed when sources are stationary. The direction information was then utilized to facilitate beamforming and source separation. As shown by the simulation results, the proposed method has a good performance for both stationary and moving sources, which was not previously possible. This work has provided an important step forward towards the solution of the real cocktail party problem.

Chapter 7

CONCLUSIONS AND FURTHER RESEARCH

7.1 Conclusions

This study has provided a substantial step towards the solution of the *cocktail party problem*, by presenting novel multimodal methods, leading to a complete solution to the problem of BSS for stationary sources and a foundation step to the solution of BSS for moving sources. The contributions can be summarized as follows:

1. A novel geometrically constrained multimodal method for CBSS of stationary sources based on second-order statistics.

2. A novel multimodal method based on higher-order statistics for CBSS of stationary and step-wise moving sources.

3. A new approach to overdetermined frequency domain blind source separation.

4. A 3-D visual tracker for tracking of multiple speakers in a room.

5. A novel multimodal method for CBSS of stationary and moving sources by combining beamforming and ICA.

In the first contribution the visual modality of speech was exploited as a geometric constraint within a second-order statistics based CBSS which exploits the non-stationary of speech (audio modality). Audio-visual information was integrated through a penalty function-based formulation to improve the CBSS algorithm. Geometrical positions of speakers were localized by colour video cameras. The face region of each speaker in each image frame was extracted on the basis of a skin model and a face model, these 2-D image coordinates were transformed to 3-D real world coordinates. On the basis of this geometric information the distances between sensors and speakers, and angles of arrivals of each speaker to each sensor were calculated. This visual information was passed to the geometrical model for estimation of the mixing filter matrix, which is integrated into a BSS algorithm in the form of a constraint and the overall SIR-Improvement was 9.8dB and MOS was 4.0.

The second contribution provided a multimodal method for BSS of stationary and step-wise moving sources for higher-order statistics based independent component analysis (ICA) of complex valued frequency domain signals. The mixing filter matrix was estimated in a similar manner to the above contribution. This geometric information based mixing filter matrix with whitening matrix of observed mixtures data was incorporated into the initialization of the complex FastICA algorithm for each frequency bin, which not only solved the inherent permutation problem in the frequency domain CBSS (with complex valued signals) but also improved the rate of convergence for static sources. For the real room recordings it was practically verified that when the sources moved in small steps then the unmixing matrix of the

previous block with the whitening matrix of the current stage mixtures provided the intelligent initialization for FastICA to separate step-wise moving sources. The overall SIR-Improvement was 14.5dB and MOS was 4.7 for stationary sources and respectively 12.8dB and 4.1 for step-wise moving sources.

The third contribution tackled the problem of ill-conditioning of the mixing matrix, which is related to the positions of the sources and inter-element microphone spacings. To achieve this an overdetermined frequency domain blind source separation (BSS) of speech signals was developed which exploited all combinations of observations and hence varying inter microphone spacings. The observations were divided into subgroups and IIFastICA was used in each subgroup. The idea was to evaluate the BSS performance of each group at every frequency bin on the basis of approximately measuring the independence of separated signals, and choosing the output of the group with the best performance at each frequency bin as the overall system output. This provided the particular advantage that the best group was frequency dependent and therefore the best inter-element spacing was likely to be chosen. The separated signals of the overall system were then obtained by transforming their frequency domain representations into the time domain. Averaged PI measurement achieved for RT = 130 and 300ms was 0.080 and 0.101 respectively, confirming good separation.

In the fourth contribution a 3-D visual tracker based on Markov Chain Monte Carlo particle filters (MCMC-PF) to simultaneously track multiple speakers in a room was implemented. It was also evaluated that audio localization was not effective due to the complexity in the case of the multiple concurrent speakers BSS problem and therefore

video localization was applied, and two calibrated cameras with overlapping fields of view (FOV) were used for sensor configuration. Since the generic particle filter is not feasible for real time tracking of multiple speakers therefore the MCMC-PF was implemented, which resulted in high sampling efficiency. The output of the tracker was position and velocity information which helped in BSS of stationary and moving sources. Two speakers in the intelligent office were tracked and the Euclidean error mean was 0.055m and standard deviation was 0.032m.

The last major contribution presented was a multimodal solution to BSS of moving sources. To calculate the time varying mixing filters in the case of moving sources, the visual modality was utilized in this method. The positions and velocities of the sources were obtained from the 3-D visual tracker. The complete BSS solution was formed by integrating a frequency domain blind source separation algorithm and beamforming: on the basis of velocity obtained from the 3-D visual tracker, if the sources were identified as stationary for a certain minimum period, a frequency domain BSS algorithm was implemented. Once the sources are moving, a beamforming algorithm which requires no prior statistical knowledge was used to perform real time speech enhancement and provide separation of the sources. The overall SIR-Improvement was 12.9dB and 8.2dB, and, MOS was 4.7 and 3.8 for frequency domain BSS and beamforming algorithms respectively.

7.2 Future research

The mixing filter matrix calculated in (3.2.3) includes only direct paths between sources and sensors, in reality, in a convolutive environment the actual mixing matrix should include the reverberation terms related to

the reflection of sounds by the obstacles and walls. Therefore a modal which can reflect some reflections in the mixing filter matrix can be a future work.

In the tracking part for BSS of moving sources a more robust approach would entail considering a speaker localization method which includes different head postures and facial directions of speakers. More sophisticated and computationally efficient schemes could also be proposed for detecting the face through a sequence of images.

In BSS of moving sources, the proposed beamforming method is only valid in a low reverberant environment and further research is required in this aspect.

The existing BSS technique to solve the *cocktail party problem* is a statistical approach and in general, is not valid for BSS of all moving sources therefore some more cognitive approach is required to solve the problem, thereby better mimicking a human, and mirroring Colin Cherry's Challenge.

BIBLIOGRAPHY

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

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

[3] S. Haykin and Z. Chen, *New Directions in Statiatical Signal Processing: From Systems to Brain: The Machine Cocktail Party Problem.* The MIT Press, Cambridge, Massachusetts London, 2007.

[4] A. J. Aubrey, *Exploiting The Bimodality Of Speech In The Context Of The Cocktail Party Problem.* PhD thesis, Cardiff University, UK, 2008.

[5] J. F. Cardoso, "Blind signal separation:statistical principles," *Proc. of IEEE*, vol. 86, pp. 2009–2025, 1998.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* New York: Wiley, 2001.

[7] T. W. Lee, *Independent Component Analysis-Theory and Application.* Kluwer academic publishers, 1998.

[8] J. Herault, C. Jutten, and B. Ans, "Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetique en apprentissage non supervis," *In Proc. GRETSI, Nice, France*, pp. 1017–1022, 1985.

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

[10] W. Wang, S. Sanei, and J. A. Chambers, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," *Proc. ICA, Nara, Japan*, 2003.

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

[12] S. Ding, J. Huang, D. Wei, and A. Cichocki, "A near real-time approach for convolutive blind source separation," *IEEE Trans. Circut and Syestem-1*, vol. 53, no. 1, pp. 114–128, 2006.

[13] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP2002*, pp. 881–884, 2002.

[14] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," in *International Conference on Independent Component Analysis and Signal Separation*, pp. 365–371, Jan 1999.

[15] C. Jutten and J. Herault, "Blind separation of sources, part i: an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[16] H. Nguyen and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, pp. 209–229, 1995.

[17] L. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," in *Proc. NNSP2001*, pp. 273–282, Sep. 2001.

[18] L. Parra and C. V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 352–362, Sep. 2002.

[19] L. Parra and C. Spence, "Convolutive blind source separation of nonstationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[20] S. Douglas, "Blind separation of acoustic signals (in microphone arrays: Techniques and applications)," *Springer, Berlin*, 2001.

[21] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[22] P. Smaragdis, "Information theoretic approaches to source separation," Master's thesis, MIT Media Lab, 1997.

[23] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind separation of convolved mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, 2005.

[24] S. Makino, "Blind source separation of convolutive mixtures of speach,(in adaptive signal processing: Applications to real-world problems)," *Springer, Berlin*, 2003.

[25] C. Jutten, "Blind separation of sources: An algorithm for sepration of convolitive mixtures," *In Proc. Internation Signal Processing Workshop, Elsevier*, pp. 275–278, 1992.

[26] A. Mansour, C. Jutten, and P. Loubaton, "Subspace method for blind separation of sources and for a convolutive mixture model," *in Signal Processing VIII, Theories and Applications. Elsevier*, pp. 2081–2084, 1996.

[27] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal Acoustical Society of America.*, pp. 212–215, 1954.

[28] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature 264*, pp. 746–748, 1976.

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

[30] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audiovisual sensor array," *Proc. IEEE Int. Conf. Multimedia (ICME)*, pp. 1382–1385, 2005.

[31] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," *in Proc. IEEE Conf. Multimedia Interfaces (ICMI)*, pp. 183–190, 2005.

[32] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Inf. Fusion*, vol. 3, no. 2, pp. 209–223, 2001.

[33] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. IEEE*, vol. 92, no. 3, pp. 485–494, 2004.

[34] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *in Proc. IEEE ICASSP*, 2001.

[35] J. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audiovisual fusion and segregation," *in Proc. Neural Inf. Process. Syst. (NIPS).*, pp. 772–778, 2000.

[36] D. Gatica-Perez, G. Lathoud, I. McCowan, and J. M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," *in Proc. IEEE Int. Conf. Comput. Vision, Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, 2003.

[37] J. Hershey and J. Movellan, "Real-time speaker tracking using particle filter sensor fusion," *Audio vision: Using audiovisual synchrony to locate sounds, in Proc. Neural Inf. Process. Syst. (NIPS)*, 1999.

[38] B. Kapralos, M. Jenkin, and E. Milios, "Audiovisual localization of multiple speakers in a video teleconferencing setting," *Int. J. Imaging Syst. Technol.*, vol. 13, pp. 95–105, 2003.

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

[40] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," *in Proc. Int. Conf. Comput. Vision ICCV*, pp. 741–746, 2001.

[41] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," *Proc. IEEE ICASSP*, pp. 425–428, 2005.

[42] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Process*, vol. 12, no. 5, pp. 520–529, 2004.

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

[44] S. Haykin, *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. John Wiley, New York, 2000.

[45] U. A. Lindgren and H. Broman, "Source separation using a criterion based on second-order statistics," *IIEEE Trans. on Signal Processing*, vol. 46, no. 7, pp. 1837–1850, 1998.

[46] H. Broman, U. Lindgren, H. Sahlin, and P. Stoica, "Source separation: A TITO system identification approach," *Signal Processing, Elsevier*, vol. 73, no. 1, pp. 169–183, 1999.

[47] H. Sahlin and H. Broman, "MIMO signal separation for FIR channels: A criterion and performance analysis," *IEEE Trans. Sig. Proc.*, vol. 48, no. 3, pp. 642–649, 2000.

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

[49] A. G. Lindgren, T. P. von Hoff, and A. N. Kaelin, "Stability and performance of adaptive algorithms for multichannel blind source separation and deconvolution," *In Proc. of EUSIPCO*, vol. 2, pp. 861–864, 2000.

[50] H. Sahlin and H. Broman, "Separation of real-world signals," *Signal Processing, Elsevier*, vol. 64, pp. 103–113, 1998.

[51] D. C. B. Chan, P. J. W. Rayner, and S. J. Godsill, "Multi-channel signal separation," *In Proc. of ICASSP*, pp. 649–652, 1996.

[52] C. Simon, C. Vignat, P. Loubaton, C. Jutten, and G. dUrso, "On the convolutive mixture source separation by the decorrelation approach," *In Proc. of ICASSP*, vol. 4, pp. 2109–2112, 1998.

[53] J. Liang and Z. Ding, "Blind MIMO system identification based on cumulant subspace decomposition," *IEEE Trans. Sig. Proc.*, vol. 51, no. 6, pp. 1457–1468, 2003.

[54] R. Liu, Y. Inouye, and H. Luo, "A system theoretic foundation for blind signal separation of MIMO-FIR convolutive mixtures a review," *In Proc. of ICA*, 2000.

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

[56] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved nonstationary signals," *Neurocomp.*, vol. 22, no. 1-3, pp. 157–171, 1998.

[57] H. C. Wu and J. C. Principe, "Simultaneous diagonalization in the frequency domain (SDIF) for source separation," *In Proc. of ICA*, pp. 245–250, 1999.

[58] A. Souloumiac, "Blind source detection and separation using second order non-stationarity," *In Proc. of ICASSP*, pp. 1912–1915, 1995.

[59] A. Ahmed, P. J. W. Rayner, and S. J. Godsill, "Considering non-stationarity for blind signal separation," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 99)*, 1999.

[60] S. Haykin, *Adaptive Filters.* John Wiley, 1994.

[61] S. Roberts and R. Everson, *Independent Component Analysis.* Cambridge, England: Cambridge University Press, 2001.

[62] J. C. A. Van-Der-Lubbe, *Information Theory.* Cambridge, England: Cambridge University Press, 1997.

[63] A. J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[64] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, pp. 101–104, April 1997.

[65] T. W. Lee, A. J. Bell, and R. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems 9*, pp. 758–764, MIT Press, 1997.

[66] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks, "A geometrically constrained multimodal approach for convolutive blind source separation," *Proc. IEEE ICASSP*, pp. 969–972, 2007.

[67] M. Z. Ikram and D. R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proc. ICASSP*, 2000.

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

[69] 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 Trans. Speech and Audio Processing*, vol. 12, pp. 530–538, Sep. 2004.

[70] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measuremet in Blind Audio Source Separation," *IEEE Trans. Speech and Audio Processing*, vol. 14, pp. 1462–1469, Jul 2006.

[71] M. G. Jafari, *Novel Sequential Algorithms for Blind Source Separation of Instantaneous Mixtures*. PhD thesis, King's College London, 2002.

[72] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filter for Tracking Applications*. Boston—London: Artech House Publishers, 2004.

[73] W. Wang and J. A. Chambers, "Blind source separation of convolutive mixtures and its application to speech enhancement: a overview," tech. rep., King's College London, Sep. 2002.

[74] W. Wang, S. Sanei, and J. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1654–1669, 2005.

[75] 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. RA-3, no. 4, pp. 323–344, 1987.

[76] J. Y. Lee and S. I. Yoo, "An elliptical boundary modal for skin color detection," *Proc. Imaging Science, Systems, and Technology*, 2002.

[77] R. Fisher, S. Perkins, A. Walker, and E. Wolfart, "Roberts cross edge detector," *Image Processing Learning Resources*, 2003.

[78] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *in Proc. IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, pp. 511–518, 2001.

[79] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2001.

[80] C. C. Took, S. Sanei, J. Chambers, R. Rickard, and S. Dunne, "Fractional delay estimation for blind source separation and localisation of temporomandibular joint sounds," *IEEE Transaction on Biomedical Engineering*, vol. 53, pp. 2123 – 2126, March 2008.

[81] C. C. Took, K. Nazarpour, S. Sanei, and J. Chambers, "Blind separation of temporomandibular joint sounds by incorporating fractional delay estimation," *Proc. of IEE IMA 2006, Cirincester, UK.*

[82] R. Mukai, H. Sawada, S.Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," *Proc. IEEE ICASSP, Hong Kong*, 2003.

[83] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind source separation of moving speakers in real reverberant environment," *Proc. IEEE ICASSP*, pp. 1133–1136, 2000.

[84] R. E. Prieto and P. Jinachitra, "Blind source separation for time-variant mixing systems using piecewise linear approximations," *Proc. IEEE ICASSP*, pp. 301–304, 2005.

[85] T. Tsalaile, S. M. Naqvi, K. Nazarpour, S. Sanei, and J. A. Chambers, "Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound," *Proc. IEEE ICASSP, Las Vegas, USA*, 2008.

[86] J. Herault and C. Jutten, "Space or time adaptive signal processing by neural network models," *in Neural Networks for computing: AIP Conference Proceedings 151, J. S. Denker, Ed., American Institute of Physics, New York*, 1986.

[87] P. Comon, C. Jutten, and J. Herault, "Blind separation of sources, part ii: problems statement," *Signal Processing*, vol. 24, pp. 11–20, 1991.

[88] E. Sorouchyari, "Blind separation of sources, part iii: stability analysis," *IEEE Trans. on Audio, Speech and Language processing*, vol. 24, pp. 21–29, 1991.

[89] A. Cichocki and L. Moszczynski, "A new learning algorithm for

blind separation of sources," *Electronic Letters*, vol. 28, no. 21, pp. 1986–1987, 1992.

[90] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[91] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits and Systems*, vol. 43, no. 11, pp. 894–906, 1996.

[92] T. W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski, "A unifying information-theoretic framework for independent component analysis," *Computers and Mathematics with Applications*, vol. 31, no. 11, pp. 1–12, 1996.

[93] A. Hyvärinen, "Fast and roubst fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, 1999.

[94] A. D. Back and A. C. Tsoi, "Blind deconvolution of signals using a complex recurrent network," *Proc. IEEE Workshop, Neural Networks for Signal Processing 4*, pp. 565–574, 1994.

[95] E. Moreau and O. Macchi, "Complex self adaptive algorithms for source separation based on higher order contrasts," *Proc. VII European Signal Processing Conference (EUSIPCO'94), Edinburgh, Scotland*, vol. II, pp. 1157–1160.

[96] E. Bingham and A. Hyvärinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Networks*, vol. 10, no. 1, pp. 1–8, 2000.

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

[98] C. L. Nikias and A. P. Petropulu, "Higher order spectra analysis. a nonlinear signal processing framework," *Prentice-Hall*, 1993.

[99] C. W. Therrien, "Discrete random signals and statistical signal processing," *Prentice-Hall*, 1992.

[100] S.Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures," *EURASIP J. Appl. Signal Process.*, no. 11, pp. 1157–1166, 2003.

[101] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: John Wiley and Sons, 1969.

[102] S. Haykin, *Unsupervised adaptive filtering.* Wiley, 2000.

[103] A. Westner and V. M. Bove, "Blind separation of real world audio signals using overdetermined mixtures," *Proc. ICA*, 1999.

[104] M. Joho, H. Mathis, and R. H. Lambert, "Overdetermined blind source separation: using more sensors than source signals in a noisy mixture," *Proc. ICA2000*, pp. 81–86, 2000.

[105] R. M. H. Sawada, S. Araki and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," *Proc. NNSP2002*, pp. 465–474, Sep. 2002.

[106] H. S. S. Winter and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation,"

EURASIP Journal on Applied Signal Processing, vol. 2006, pp. 1–11, 2006.

[107] A. Koutras, E. Dermatas, and G. Kokkinakis, "Improving simultaneous speech recognition in real room environments using overdetermined blind source separation," *Proc. Eurospeech 2001*, pp. 1009–1012, Sep. 2001.

[108] E. Bingham and A. Hyärinen, "A fast fixed-point algorithm for independent component analysis of complex valued signals," *International Journal of Neural Systems*, vol. 10, no. 1, pp. 1–8, 2000.

[109] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

[110] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "A multimodal approach for frequency domain blind source separation for moving sources in a room," *Proc. IAPR CIP2008, Santorini, Greece*, 2008.

[111] K. E. Hild-II, D. Erdogmus, and J. C. Principe, "Blind source extraction of time-varying, instantaneous mixtures using an on-line algorithm," *Proc. IEEE ICASSP, Orlando, Florida, USA*, 2002.

[112] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. on Speech and Audio processing*, vol. 10, no. 3, pp. 173–185, 2002.

[113] S. M. Naqvi, Y. Zhang, T. Tsalaile, S. Sanei, and J. A. Chambers, "A multimodal approach for frequency domain independent component

analysis with geometrically-based initialization," *Proc. EUSIPCO, Lausanne, Switzerland*, 2008.

[114] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP MAGAZINE*, pp. 4–21, 1988.

[115] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," *in Proc. Eur. Conf. Comput. Vision (ECCV)*, 2002.

[116] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," *in Proc. IEEE ICASSP*, pp. 881–884, 2004.

[117] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," *in Proc. IEEE Workshop Applications Signal Process. Audio Acoust. (WASPAA)*, 2001.

[118] M. Isard and J. MacCormick, "BRAMBLE: A Bayesian multi-blob tracker," *in Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2001.

[119] G. Lathoud and M. Magimai-Doss, "Asector-based, frequency-domain approach to detection and localization of multiple speakers," *in Proc. IEEE ICASSP*, pp. 265–268, 2005.

[120] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone array measurements," *in Proc. IEEE ICASSP*, pp. 371–374, 1997.

[121] B. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," *in Proc. IEEE ICASSP*, 2004.

[122] C. Faller and J. Merimaa, "Source localization in complex listening situtations: Selection of binaural cues based on interaural coherence," *The Journal Of The Acoustical Society Of America*, vol. 116(5), pp. 3075–3089, 2004.

[123] M. Fallon, S. Godsill, and A. Black, "Joint acoustic source localization and orientation estimation using sequential monte carlo," *in Proc. Digital Audio Effects (DAFx-06)*, 2006.

[124] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "Multimodal blind source separation for moving sources," *Proc. IEEE ICASSP, Taipei, Taiwan*, 2009.

[125] M. Isard and A. Blake, "CONDENSATION: Conditional density propagation for visual tracking," *Int. J. Comput. Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[126] K. Astom, "Introduction to stochastic control theory," *Acdemic Press*, 1970.

[127] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall, 1996.

[128] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* New York: Springer-Verlag, 2001.

[129] Z. Khan, T.Balch, and F.Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," *in Proc. Eur. Conf. Comput. Vision (ECCV), Prague, Czech Republic*, pp. 279–290, 2004.

[130] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IIEEE*

Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1805–1818, 2005.

[131] J. S. Liu, *Monte Carlo Strategies in Scientific Computing.* New York: Springer-Verlag, 2001.

[132] S. K. Pang, S. J. Godsill, J. Li, F. Septier, and S. Hill, *Inference and Learning in Dynamic Models: Sequential Inference for Dynamically Evolving Groups of Objects.* Cambridge University Press, under revision and to appear 2009.

[133] C. Berzuini, N. G. Best, W. R. Gilks, and C. Larizza, "Dynamic Conditional Independence Models and Markov Chain Monte Carlo Methods," *Journal of the American Statistical Association*, pp. 440:1403–1412, 1997.

[134] S. K. Pang, J. Li, and S. J. Godsill, "Models and Algorithms for Detection and Tracking of Coordinated Groups," *IEEE Aerospace Conference*, pp. 1–17, 2008.

[135] S. Haykin, *Adaptive Filter Theory.* Prentice-Hall, 2001.

[136] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[137] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ica and beamforming," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 666–678, 2006.