Robust Variational Bayesian Clustering for Underdetermined Speech Separation

by

Zeinab Youssef Zohny

A Doctoral Thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), at Loughborough University.

October 2016

Advanced Signal Processing Group,
Wolfson School of Mechanical, Manufacturing and Electrical Engineering,
Loughborough University, Loughborough
Leicestershire, UK, LE11 3TU.

I dedicate this thesis to my family.

Abstract

The main focus of this thesis is the enhancement of the statistical framework employed for underdetermined T-F masking blind separation of speech. While humans are capable of extracting a speech signal of interest in the presence of other interference and noise; actual speech recognition systems and hearing aids cannot match this psychoacoustic ability. They perform well in noise and reverberant free environments but suffer in realistic environments. Time-frequency masking algorithms based on computational auditory scene analysis attempt to separate multiple sound sources from only two reverberant stereo mixtures. They essentially rely on the sparsity that binaural cues exhibit in the time-frequency domain to generate masks which extract individual sources from their corresponding spectrogram points to solve the problem of underdetermined convolutive speech separation. Statistically, this can be interpreted as a classical clustering problem.

Due to analytical simplicity, a finite mixture of Gaussian distributions is commonly used in T-F masking algorithms for modelling interaural cues. Such a model is however sensitive to outliers, therefore, a robust probabilistic model based on the Student's t-distribution is first proposed to improve the robustness of the statistical framework. This heavy tailed distribution, as compared to the Gaussian distribution, can potentially better capture outlier values and thereby lead to more accurate probabilistic masks for source separation. This non-Gaussian approach is applied to the state-of the-art MESSL algorithm and comparative studies are undertaken to confirm the improved separation quality.

A Bayesian clustering framework that can better model uncertainties in reverberant environments is then exploited to replace the conventional expectation-maximization (EM) algorithm within a maximum likelihood

estimation (MLE) framework. A variational Bayesian (VB) approach is then applied to the MESSL algorithm to cluster interaural phase differences thereby avoiding the drawbacks of MLE; specifically, the probable presence of singularities and experimental results confirm an improvement in the separation performance.

Finally, the joint modelling of the interaural phase and level differences and the integration of their non-Gaussian modelling within a variational Bayesian framework, is proposed. This approach combines the advantages of the robust estimation provided by the Student's t-distribution and the robust clustering inherent in the Bayesian approach. In other words, this general framework avoids the difficulties associated with MLE and makes use of the heavy tailed Student's t-distribution to improve the estimation of the soft probabilistic masks at various reverberation times particularly for sources in close proximity. Through an extensive set of simulation studies which compares the proposed approach with other T-F masking algorithms under different scenarios, a significant improvement in terms of objective and subjective performance measures is achieved.

Contents

Statement of Originality

The contributions of this thesis are mainly on the enhancement of the statistical framework used for underdetermined blind speech separation. The novelty of the contributions is supported by one journal paper under review, one electronic letter and two conference papers.

In Chapter 3, a novel approach to the probabilistic modelling of the spatial interaural cues used in time-frequency (T-F) masking based speech separation is proposed. Based on the heavy tailed Student's t-distribution, this non-Gaussian modelling is less sensitive to outliers. This approach is important to improve the robustness of T-F algorithms based on clustering spectrogram points in reverberant environments without the need for any reverberation detection method. The results have been published in:

1. Z. Zohny, S. M. Naqvi and J. A. Chambers, "Enhancing the MESSL algorithm with robust clustering based on the Student's t-distribution," Electronics Letters, vol. 50, no. 7, pp. 552-554, 2014

2. Z. Zohny and J. A. Chambers,"Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL," *Proc. Dig. Sig. Process.*, pp. 59-62, 2014.

In Chapter 4, a variational Bayesian framework is proposed for clustering spectrogram points depending only on their interaural phase difference cues. This elegant approach overcomes the drawbacks of the popular expectation maximization (EM) clustering algorithm, particularly the probable presence of singularities and improves the separation performance. The results have been published in:

3. Z. Zohny, S. M. Naqvi and J. A. Chambers,"Variational EM for clustering interaural phase cues in MESSL for blind source separation of speech," *Proc. of 40th Int. Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, 2015.

In Chapter 5, a general probabilistic approach for T-F masking speech separation is presented. This approach is based on integrating the non-Gaussian modelling into a variational Bayesian framework for the joint clustering of interaural cues. The proposed framework combines the advantages of the robust estimation provided by the Student's t-distribution and the robust clustering inherent in the Bayesian approach when modelling uncertainties and hence improves the estimation of the soft probabilistic masks at various reverberation times particularly for sources in close proximity. Comparative studies of the proposed approach with the state-of-the-art algorithms under different scenarios have confirmed a significant improvement in terms of objective and subjective performance measures.

4. Z. Zohny, S. M. Naqvi, J. A. Chambers and W. Wang, "Robust variational Bayesian clustering for speech separation," submitted to IEEE Transactions on Signal Processing, 2016.

Acknowledgements

Many people have given me great support throughout the last four years. It is not easy to find the right words expressing my sincere gratitude. Foremost, my supervisor Professor Jonathon Chambers, who had accepted generously to guide me through this unconventional and exciting journey. Based in Oman, mother of two and a full time lecturer, pursuing a PhD degree seemed more of a dream. Without his encouragement and belief, I would not have been able to fulfil this challenge. I'm very thankful for his continuous support, patience and motivation. Our regular Skype meetings, discussions and his effort in reviewing the publications and the thesis drafts have been invaluable. The amount of knowledge, devotion and resilience I have gained through this work with his exceptional leadership is tremendous.

I would also like to thank Dr. Syed Mohsen Naqvi for his support, interesting discussions and valuable comments. I owe many thanks to all my colleagues in the advanced signal processing group (ASPG), who were always there whenever I needed advice or support.

Special thanks to my dear friend and colleague Ozak Esu, who was always offering great help to me and my family during our visits.

I owe many thanks to my friends in Oman, Miriam Humer and Irene Sasso. With our completely different backgrounds, engineering, art and architecture, we enjoyed long and inspiring conversations, appreciating the beauty of mathematics and the logic in art. I'm also very thankful to my best friends Ingy EL Hakim, Abir Osman and Omeya Sami, who were continuously encouraging me, especially during the writing phase of my thesis.

I'm very grateful to my mother who was always a model in perseverance and determination and my brother with his great sense of humour, listening to completely irrelevant statistical problems with great interest and love. My mother-in-law deserves also many thanks, with the support she gave to

my family, I had the privilege to travel as much as needed, knowing the boys were completely in good hands.

Last, but most importantly, I wish to express my deepest gratitude to my husband and boys. They supported me in every aspect, tolerated long hours of studying during weekends and holidays, my absence on many occasions or my mind absence and swing of moods, without any complaint. Without the love, support and understanding of all my friends and family, I would have never been able to complete this work.

Zeinab Zohny

September, 2016

List of Acronyms

AmI	Ambient Intelligence
ASR	Automatic Speech Recognition
BRIR	Binaural Room Impulse Response
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CPP	Cocktail Party Problem
DFT	Discrete Fourier Transform
DUET	Degenerate Unmixing Estimation Technique
EM	Expectation Maximization
EVD	Eigen Value Decomposition
FDBSS	Frequency Domain Convolutive Blind Source Separation
GCC	Generalized Cross-Correlation
GMM	Gaussian Mixture Model
HMI	Human-Machine Interface
HOS	Higher Order Statistics
ICA	Independent Component Analysis

ILD Interaural Level Difference

IPD Interaural Phase Difference

ITD Interaural Time Difference

IVA Independent Vector Analysis

LOST Line Orientation Separation Technique

MESSL Model-based EM Source Separation and Localization

MLE Maximum Likelihood Estimation

MOSPALOSEP MOdeling SPAtialization LOcalization and SEParation

MOS Mean Opinion Score

OPTIVIP OPTImization of the Visual Implantable Prosthesis

PCA Principal Component Analysis

PESQ Perceptual Evaluation of Speech Quality

PHAT PHAse Transform

PI Permutation Index

SIR Signal-Interference-Ratio

SMM Student's t-distribution Mixture Model

SOS Second Order Statistics

T-F Time-Frequency

VBSS Variational Bayesian source Separation

VB Variational Bayesian

List of Symbols

Scalar variables are denoted by plain lower-case letters, (i.e., x), vectors by bold-face lower-case letters, (i.e., $\mathbf{x}$), and matrices by upper-case bold-face letters, (i.e., $\mathbf{X}$). Some frequently used notations are as follows:

$\mathrm{E}\{\cdot\}$	Statistical expectation
$(\cdot)^T$	Transpose
$(\cdot)^H$	Hermitian transpose
$\|.\|_2$	Euclidean norm
$(.)^{-1}$	Matrix inverse
$\mathbf{H}$	Mixing matrix
$\mathbf{W}$	Estimated Unmixing matrix
$\mathbf{P}$	Permutation matrix
$\mathbf{\Lambda}$	Diagonal matrix
$\mathbf{Q}$	Orthogonal matrix
N	Number of sources
M	Number of sensors
$diag(\mathbf{d})$	diagonal matrix with vector $\mathbf{d}$ on its main diagonal
$KL(.)$	Kullback-Leibler divergence

$\Gamma\{\cdot\}$	Gamma fuction

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Cocktail Party Problem

Humans manage effortlessly to focus their attention and follow a single speaker while filtering out other interference such as simultaneous conversations, music or noise as shown in Figure 1.1. The cocktail party problem (CPP) introduced by Colin Cherry [1] refers to this psychoacoustic ability. For more than sixty years, various studies in many disciplines have been dedicated to understand the human auditory system and to seek computational solutions imitating its capability [2].

Figure 1.1: Cocktail party problem.

The need for machines capable of sound localization and separation is becoming more crucial especially with the growing number of applications requiring human-machine interfaces (HMIs). The emerging field of ambient intelligence (AmI) [3] aims to create a digital environment sensitive, responsive and supportive of people in their daily lives relying essentially on these interfaces. An interesting example of an HMI is Siri. Siri is an application for Apple's iOS, it acts as personal assistant and navigator. According to Apple it should adapt to the individual's preferences, personalizing results and performing tasks such as recommending nearby restaurants, or getting directions [4].

Unfortunately, a full understanding of the CPP phenomenon is still incomplete and automatic speech recognition (ASR) based systems suffer from many limitations. They work well in noise free and low reverberant environments but their performance degrades particularly in the presence of other interfering speakers. It is also well known that listeners with hearing impairment have difficulty in the presence of background noise since the effectiveness of existing hearing aids diminishes in a typical noisy and reverberant CPP environment. Advances in the field of speech separation and recognition would potentially benefit the design of hearing aids and improve the speech intelligibility of their wearers in actual social conditions.

Different environmental assumptions can be made about the CPP problem. In the instantaneous case, the signals are assumed to arrive instantly at the sensors with only an intensity difference. Whereas, the anechoic case considers the arrival delays between the sources and sensors. Both cases are not realistic and differ from the real scenario where acoustic signals take multiple paths to the sensors. The natural reverberant CPP environment can be ideally represented by the echoic convolutive mixing model shown in Figure 1.2.

Different algorithms aiming at solving the CPP problem have originated

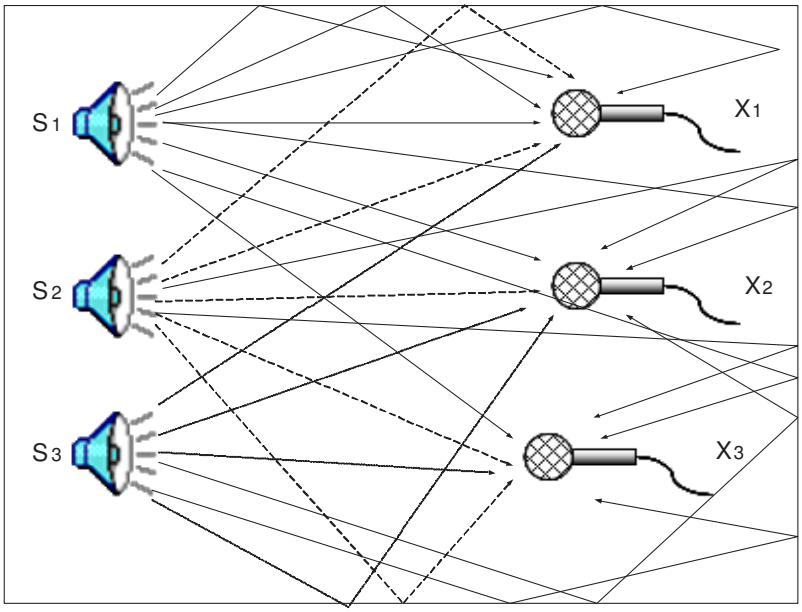


Figure 1.2: An enclosed convolutive mixing environment with three sources and three sensors.

in the fields of blind source separation (BSS) and computational auditory scene analysis (CASA). These methods as explained in detail in the following sections rely basically on statistical properties or assumptions to separate a mixture of signals.

The main objective of this thesis is to enhance the speech separation through the provision of a statistical framework that suits best the realistic conditions of the CPP environment, i.e. is capable of extracting a source of interest among multiple speakers, reverberations and similar to humans using only two microphones.

1.2 Blind Source Separation

Following the work of Jutten and Herault [5] and Comon [6] in the early nineties, extensive research has been dedicated to formulating a mathematical framework for the recovery of signals from their mixtures, given that

neither the original sources nor the mixing process are known by the separation algorithm. There are multiple potential applications for BSS [7]; in array signal processing, for the recognition of sources from unknown arrays and in wireless communications BSS can potentially be used for blind code-division multiple access, blind multi-input multi-output equalization and timing recovery [8], [9]. Source separation can also be applied in speech processing for automatic voice recognition in noisy acoustic environments [10] and in the decomposition of brain imaging such as electroencephalogrophy [11]. Other areas where BSS is applied include financial time series analysis [12] and cosmology for the analysis of cosmic microwave background [13].

Although BSS approaches can be categorized based on the separation assumptions such as the number of sensors relative to the number of sources, the nature of the signals, the domain and the criterion for separation, they all rely on some statistical discriminant to separate sources in a blind manner. In the context of second order statistics (SOS) based algorithms such as that due to Parra-Spence [14], the sources are separated in the frequency domain based only on their uncorrelatedness while making use of other assumptions; basically the non-stationarity of the underlying signals is exploited to achieve the separation.

On the other hand, higher order statistics (HOS) based algorithms such as independent component analysis (ICA) [15], assume that the sources are statistically independent, have non-Gaussian distributions with at most one Gaussian component and the convolutive separation is performed in the frequency domain through optimizing some objective function related to the non-Gaussianity of the components. The absolute value of Kurtosis and negentropy are commonly used as a measure of non-Gaussianity. Different approaches to the objective functions are discussed in detail in [15] and a survey of existing algorithms applied to convolutive audio mixtures can be found in [16]. By performing the separation in the frequency domain [14] [15],

the convolutive problem is decomposed into smaller multiplicative problems at each frequency bin, which improves the computational and time efficiency, compared to the time domain based solutions [17]. Unfortunately, frequency domain SOS and ICA based methods suffer mainly from the permutation problem [18]. Due to the fact that both the original sources and the mixing matrix are unknown, the sources are recovered at each bin but not necessarily in the same order. This ambiguity should be solved since only consistent permutations for all frequencies reconstruct the original signals properly. By introducing a new multivariate cost function and a dependency model capturing the inter-frequency dependencies, independent vector analysis (IVA) simultaneously achieves the separation and solves the permutation problem [19].

BSS has enriched the field of signal processing, however it has certain weaknesses when solving the CPP problem. The independence assumption is unrealistic in a neurobiological context since a mixing environment can have a varying number of speakers or any form of noise such as laughing or coughing, and yet the human capability remains the same regardless of the variations in the auditory scene [2]. Moreover, most of the ICA/BSS techniques are based on linear filtering which involves the pseudo-inversion of the mixing matrices. This is only possible when the number of sources is equal or less than the number of sensors, i.e. for determined or overdetermined cases [20].

Nevertheless, the problem of extracting more sources than sensors (underdetermined or overcomplete separation) is more challenging and other approaches such as time-frequency (T-F) masking are commonly used as explained in the next section.

1.3 T-F masking

Humans are capable of extracting a sound of interest from a mixture reaching their ears. The auditory scene analysis describes this process in two stages [21]. In the first stage, the sound undergoes a decomposition into the T-F domain then the auditory system reassembles the spectrogram points belonging to the same source based on different cues into different streams. Each source in the mixture generates various cues that can be used to potentially group the relevant points, these cues can be related to the spatial location of each source or other intrinsic sound properties such as amplitude and frequency modulations, harmonicity, temporal continuity and trained speech models [22]. The availability of many cues ensures the grouping process in case one of the cues fails to indicate the correct grouping.

The T-F masking approach originated in the field of computational auditory scene analysis (CASA) [22]. Driven by the work of Bregman in 1990 [23], CASA aims at the design of computational methods imitating the human auditory system and thereby capable of directly extracting a sound of interest in a cocktail party environment. The concept of W-disjoint orthogonality introduced by Yilmaz and Rickard in [24] has played a major role in T-F masking approaches. It assumes that the energy at a T-F point mostly belongs to one source. In their proposed method, the degenerate unmixing estimation technique (DUET) [24], the ratio of T-F representations of the mixtures is used to construct a two-dimensional histogram with only one peak for each source. The location of the peak corresponds to the relative interaural level and time differences of each source. Using the histogram, binary T-F masks are obtained to separate the sources from their mixtures.

In contrast to the hard binary masks of DUET, many of the binaural T-F masking algorithms such as the model-based expectation maximization source separation and localization (MESSL) and the system for modeling,

spatialization, localization and separation (MOSPALOSEP) perform separation using soft probabilistic masks [25] [26]. Different interaural cues resulting from the time and level differences between the signals reaching the spatially distinct microphones are used to achieve the separation. These cues are commonly modelled using Gaussian mixture models (GMMs) and the main approach in these algorithms is clustering in the selected interaural feature space via the expectation maximization (EM) algorithm, to generate soft masks extracting the speech of interest from the associated spectrogram points [22]. The choice of the EM framework and the Gaussian mixture modelling is briefly justified in the following section.

1.4 EM for Gaussian mixture models

The EM algorithm is an iterative algorithm for maximum likelihood estimation (MLE) known as one of the most popular methodologies in the field of statistical signal processing for estimating the parameters of a probability distribution function. EM is commonly used when the direct access to data required for parameters estimation is not possible, typically in binning or histogram operations [27]. Hence, it is ideally suitable for clustering problems where the points belonging to each cluster are not known. In other words, EM is able to estimate the parameters when there is a many-to-one mapping from an underlying distribution to the actual distribution fitting the observations [27].

The likelihood optimization through EM is performed in two steps as the name of the algorithm implies, the expectation step followed by the maximization step. In the E step, the expected values of the unknown variables are computed given the current estimate of the parameters and observations then new estimates of the parameters are determined in the M step. The algorithm iterates between these steps until convergence as shown in Figure

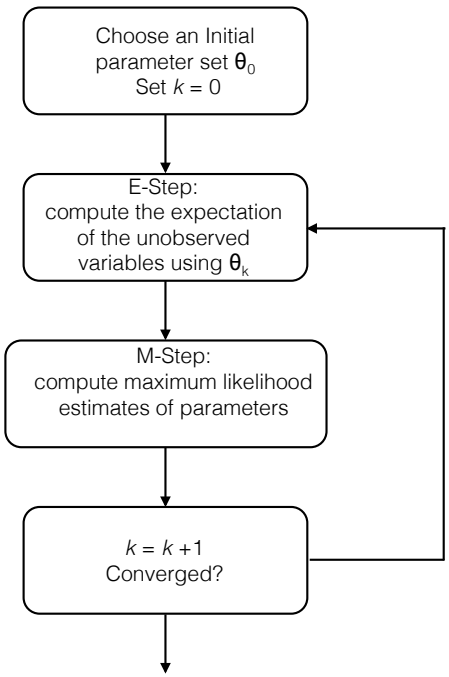


Figure 1.3: An overview of the EM algorithm. After initialization, it alternates between the E and M steps until convergence.

1.3. Since the seminal work of Dempster [28] in 1977, the EM algorithm was largely applied in many areas of engineering and signal processing, such as image modelling and reconstruction, speech recognition, channel estimation and information theory [29]. Its popularity has increased due to its analytical simplicity, guaranteed stability and local convergence [27]. For multivariate data problems attention was directed to the use of the Gaussian distribution for probabilistic modelling because of their computational simplicity since their parameters are easily determined via the EM framework.

This approach suffers from two major limitations related to both the shortcomings of the EM algorithm as a clustering methodology and the Gaussian distribution modelling known by its sensitivity to outliers. As explained in [30] the tails of the Gaussian distribution are considered less

significant than required and hence the estimation of individual component parameters is easily affected by atypical observations. On the other hand, the EM algorithm requires the knowledge of the posterior distribution of the hidden variables given the observations thereby restricting its application to complex problems [29]. The potential presence of singularities is another disadvantage associated to the EM applied to GMMs [31]. Maximizing the log-likelihood function is an ill-posed problem since if one of the Gaussian components collapses on a data point and its variance tends to zero, the log-likelihood will tend to infinity. Additional techniques should be employed when adopting this framework to solve the problem of singularities [31].

The work developed in this thesis aims at improving the statistical framework for underdetermined T-F based speech separation by exploiting the following areas:

- Non-Gaussian probabilistic modelling of interaural cues to provide robustness and improve the speech separation in real reverberant environments.
- Variation Bayesian inference approach as an alternative to the traditional EM algorithm to overcome its shortcomings.
- Multivariate modelling of interaural cues to avoid unnecessary assumptions of independence or additional effort for formulating their correlation.

An overview of the work structure and a short summary of each chapter are provided in the following section.

1.5 Thesis Outline

This thesis is organized as follows:

- Chapter 2 provides a relevant survey of the different methods used to solve the CPP problem particularly the frequency domain based techniques. The first section of Chapter 2 focuses on the determined/overdetermined case and three different convolutive BSS techniques are presented. The different criteria used to evaluate BSS performance quality are introduced followed by some illustrative comparative results using sources from the standard databases and the impulse responses generated by the imaging method. The second section of this chapter gives an overview of the T-F masking concept used widely to tackle the underdetermined case, discusses various ICA and CASA based approaches with a focus on the state-of-the-art MESSL algorithm which uses Gaussian mixture modelling for the interaural cues at each spectrogram point and an EM estimation procedure.

- Chapter 3 exploits non-Gaussian probabilistic modelling as an alternative to the Gaussian modelling employed in MESSL. The Student's t-distribution mixture models (SMMs) are introduced and applied for fitting the mixtures of cues at each spectrogram point. They are used to model independently both the IPD and ILD cues. SMMs enhance the robustness in the presence of reverberations and the performance of MESSL with SMMs is shown to improve considerably.

- Chapter 4 describes the variational Bayesian framework. The proposed framework is then used for clustering spectrogram points based on their IPD cues. This elegant approach overcomes the drawbacks of the EM for GMMs and improves the separation especially when the sources are in close proximity. Simulation studies based on speech

mixtures formed from the TIMIT database confirm the advantage of the proposed approach.

- Chapter 5 presents a novel probabilistic approach for T-F masking based speech separation where non-Gaussian modelling is integrated into the variational Bayesian framework for the joint clustering of IPD and ILD cues. Bayesian SMMs are described and the variational EM update rules for SMMs are presented. The performance of the proposed approach using real impulse responses is evaluated and compared with the state-of-the-art algorithms under different scenarios in terms of objective and subjective performance measures.

- Chapter 6 concludes the thesis and suggests directions for future research.

Chapter 2

A RELATED SURVEY OF CONVOLUTIVE SPEECH SEPARATION METHODS

This chapter gives an overview of the various approaches used to solve the cocktail party problem (CPP) particularly frequency domain based techniques. BSS based techniques can be categorized into SOS methods and HOS methods, but they all assume that the number of sensors is greater or equal to the number of sources. Three major frequency domain convolutive BSS (FDCBSS) techniques are presented and compared using sources from standard databases and the impulse responses generated by the imaging method.

For the underdetermined case, an overview of the T-F masking concept, used widely to tackle this problem, is given. Different T-F masking methods developed in both the ICA and CASA communities are discussed, with a focus on the state-of-the-art algorithm referred to as model-based expectation-maximization source separation and localization (MESSL). Based on the locations of the sound sources, MESSL attempts to separate multiple sources from their binaural mixtures in the presence of reverberation, thereby achieving underdetermined convolutive blind source separation.

2.1 Introduction

Many signal processing applications rely on the recovery of various signals from certain observations while having limited information about the mixing process and the original sources [20]. BSS was initially designed to extract these sources from their instantaneous mixtures. Assuming N sources and M sensors (or microphones), where $M \geq N$, the instantaneous noise free model formulated by Comon in [6] can be expressed as

$$x_i(t) = \sum_{j=1}^{N} h_{ij} s_j(t) \tag{2.1.1}$$

where $x_i(t)$ is the ith element of the observed vector $\mathbf{x}(t) \in \mathbb{R}^M$, $s_j(t)$ denotes the jth element of the source vector $\mathbf{s}(t) \in \mathbb{R}^N$, t denotes the discrete time index and h_{ij} is the ith row, jth column element of the mixing matrix $\mathbf{H} \in \mathbb{R}^{M \times N}$; Equation (2.1.1) can also be written as

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t) \tag{2.1.2}$$

Separation is achieved by estimating the separating or unmixing matrix $\mathbf{W}$ and the estimated sources are expressed as follows:

$$y_j(t) = \sum_{i=1}^{M} w_{ji} x_i(t) \tag{2.1.3}$$

where $j = 1, \ldots, N$, $y_j(t)$ is the jth element of the estimated column vector $\mathbf{y}(t)$ and w_{ji} is the jth row, ith column element of the separating matrix $\mathbf{W}$; or alternatively

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2.1.4}$$

The instantaneous assumption is not generally realistic and differs from the real scenario where signals take multiple paths to the sensors. The CPP

problem is better represented by a convolutive mixing model

$$x_i(t) = \sum_{j=1}^{N} \sum_{p=0}^{P-1} h_{ij}(p) s_j(t-p) \tag{2.1.5}$$

where $i = 1, \ldots, M$, $p = 0, \ldots, P$-1 and h_{ij} is the P-tap impulse response from source j to microphone i and the pth slice of the FIR filter matrix is

$$\mathbf{H}(p) = \begin{bmatrix} h_{11}(p) & \cdots & h_{1N}(p) \\ \vdots & \ddots & \vdots \\ h_{M1}(p) & \cdots & h_{MN}(p) \end{bmatrix} \tag{2.1.6}$$

The estimated sources can be obtained as follows

$$y_j(t) = \sum_{i=1}^{M} \sum_{q=0}^{Q-1} w_{ji}(q) x_i(t-q) \tag{2.1.7}$$

where j=1,...,N, $q = 0$,...,Q-1 and the qth slice of the unmixing filter matrix is

$$\mathbf{W}(q) = \begin{bmatrix} w_{11}(q) & \cdots & w_{1M}(q) \\ \vdots & \ddots & \vdots \\ w_{N1}(q) & \cdots & w_{NM}(q) \end{bmatrix} \tag{2.1.8}$$

Many methods known as multichannel blind deconvolution type algorithms have been proposed to address the convolutive BSS problem [32]. They can be categorized according to the domain of separation criterion into time and frequency domain approaches. The time domain based solutions commonly suffer from slow convergence and significant computational load [17]. Frequency domain approaches, on the other hand, provide a fast alternative; a survey of the FDCBSS algorthims can be found in [33].

Applying a T-point discrete Fourier transform (DFT), the observed vector for each frequency bin can be expressed as

$$\mathbf{x}(\omega, t_k) = \mathbf{H}(\omega)\mathbf{s}(\omega, t_k) \tag{2.1.9}$$

where $\mathbf{s}(\omega, t_k) = [s_1(\omega, t_k), .., s_N(\omega, t_k)]^H$ and $\mathbf{x}(\omega, t_k) = [x_1(\omega, t_k), .., x_M(\omega, t_k)]^H$ are the time-frequency representations of the source and the observed vectors, respectively, t_k denotes the discrete time block index, ω is the frequency index, K is total the number of time blocks and $(.)^H$ denotes Hermitian transpose.

The frequency domain representation of the multiple impulse responses is a set of M by N matrices at all frequencies, each expressed as

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

The deconvolution problem is thus transformed into an instantaneous BSS problem at each frequency ω. The sources are separated in every frequency bin by estimating the N by M unmixing matrix $\mathbf{W}(\omega)$ such that

$$\mathbf{y}(\omega, t_k) = \mathbf{W}(\omega)\mathbf{x}(\omega, t_k) \tag{2.1.11}$$

and

$$\mathbf{W}(\omega) = \begin{bmatrix} w_{11}(\omega) & \cdots & w_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ w_{N1}(\omega) & \cdots & w_{NM}(\omega) \end{bmatrix} \tag{2.1.12}$$

where $\mathbf{y}(\omega, t_k) = [y_1(\omega, t_k), .., y_N(\omega, t_k)]^H$ is the time-frequency representation of the estimated sources and $\mathbf{W}(\omega)$ is the frequency domain representation of the unmixing matrix. The estimated time domain sources can be obtained by applying an inverse Fourier transform to the outputs in the fre-

quency domain.

FDCBSS methods decompose the convolutive problem into smaller multiplicative problems at each frequency bin to improve the computational efficiency substantially compared to the time domain methods [16]. Unfortunately, they suffer from permutation and scaling ambiguity [18]. The reason is mainly due to the fact that both the sources and the mixing matrix are unknown which means that the order of the recovered sources at each frequency bin cannot be determined. Additionally, it is not possible to determine the energy of the original sources; any scalar multiplier α_i for one of the sources s_i could always be cancelled by dividing the corresponding column of $\mathbf{H}$ by the same scalar as follows

$$\mathbf{x} = \sum_i (\frac{1}{\alpha_i} h_i)(s_i \alpha_i) \tag{2.1.13}$$

In other words, the separating matrix $\mathbf{W}$ can be expressed as

$$\mathbf{W} = \mathbf{P}\mathbf{\Lambda}\mathbf{H}^{-1} \tag{2.1.14}$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{\Lambda}$ is a diagonal matrix conveying the scaling ambiguity. If the permutation problem is not solved, the estimated time domain signals might combine contributions from different sources into a single channel. The scaling ambiguity would result in an overall filtering of the estimated sources [16].

FDCBSS techniques separate mixtures of speech signals through the optimization of some statistical discriminant based on SOS or HOS assumptions made about the sources, given that the amplitude and permutation ambiguities discussed above are mitigated. Since they rely essentially on the pseudo-inversion of the mixing matrices. This is only possible for the determined or overdetermined cases [20].

For underdetermined or overcomplete separation, T-F masking is com-

monly used. The T-F approach has its origin in the field of computational auditory scene analysis (CASA) but has been developed in both the CASA and ICA communities [22]. The T-F representation can be achieved by transforming the mixtures into a Fourier or Wavelet basis or a windowed auditory filterbank [34]. The separation relies on the powerful assumption of sparseness that acoustic sources exhibit in a given basis. Based on this assumption, the probability of overlapping of two signals in the T-F domain is considered very low [16]. In other words, most of the energy at each T-F point belongs to a single source. Accordingly, a mask can be applied to preserve the energy in the T-F points belonging to the speech source of interest and attenuate the energy of the interfering sources in the rest of the spectrograms points. The basic idea behind using T-F masking for sound separation has been used for decades [35]. For instance, the classical Wiener filter [36] can be viewed as a T-F mask where each T-F unit of the mask represents the ratio of the target energy to the mixture energy within the unit.

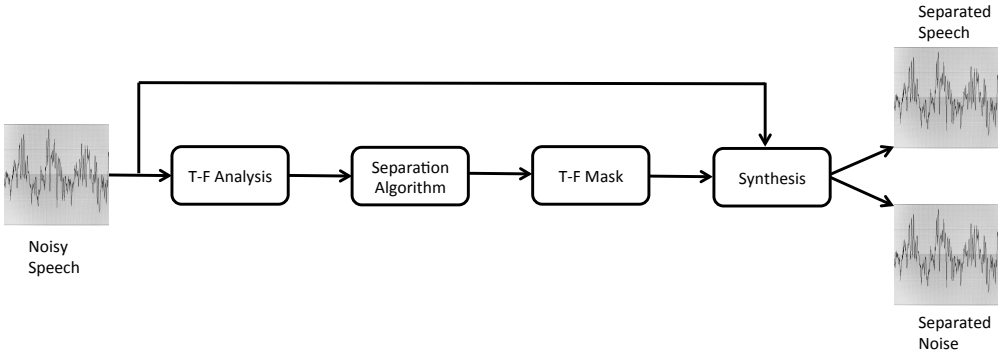


Figure 2.1: Block diagram depicting the main steps in a T-F masking system for speech separation.

As shown in Figure 2.1, a noisy speech signal first undergoes T-F analysis, a separation algorithm is then applied and the outcome of the separation is a T-F mask used in a synthesis step to convert the estimated sources back to the waveform representation. Naturally, binary masks were initially considered, where a value 1 indicates that the energy in the corresponding spectrogram point should be preserved and a value 0 indicates that the energy should be removed [37]. In order to avoid artifacts caused by applying these binary or hard masks such as musical noise, recent studies proposed the use of smooth or soft masks [35].

Section 2.2 focuses on the non-sparse methods employed for overdetermined/determined BSS, wherein three major techniques are presented to provide an insight on the different statistical approaches and recent development in the literature. Section 2.3 describes the various performance measures used to assess the performance of these techniques. In Section 2.4, experimental results comparing their separation performance are undertaken. Section 2.5 reviews the major approaches for underdetermined BSS and finally the chapter is summarized in Section 2.6.

2.2 Overdetermined/Determined BSS

The following section describes three FDCBSS techniques. The first one is based on second order statistics while the others exploit higher order statistics. These techniques work effectively when the number of sensors is sufficient ($M \geq N$).

2.2.1 SOS based BSS: Parra-Spence algorithm

In SOS separation algorithms the sources are separated based only on their uncorrelatedness rather than using the stronger condition of independence. However, SOS are not generally sufficient to achieve the separation and

this is why instead of making assumptions on HOS, SOS make use of other assumptions such as the non-stationarity or smoothness of the sources. The main advantage of SOS is that it requires less data for its estimation and is less sensitive to noise and outliers [38].

Parra and Spence tackled the problem by exploiting the cross-correlations at multiple times which provides a sufficient set of constraints for estimating the unknown channels. The algorithm searches for the set of separating matrices $\mathbf{W}(\omega)$ that diagonalizes simultaneously the cross-correlation matrices of the estimated sources for K different time lags. The algorithm transforms the problem to the frequency domain as explained in the previous section and solves a separation problem for every frequency bin.

Since in the noise free case, $\mathbf{y}(\omega, t_k) = \mathbf{W}(\omega)\mathbf{x}(\omega, t_k)$, then

$$\mathbf{R}_y(\omega, t_k) = \mathbf{W}(\omega)\mathbf{R}_x(\omega, t_k)\mathbf{W}^H(\omega) \tag{2.2.1}$$

$$= \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{\Lambda}_s(\omega, t_k)\mathbf{H}^H(\omega)\mathbf{W}^H(\omega) \tag{2.2.2}$$

where $\mathbf{\Lambda}_s(\omega, t_k)$ is the diagonal covariance matrix of the sources and $\mathbf{R}_x(\omega, t_k)$ is the covariance matrix of $\mathbf{x}(\omega, t_k)$.
The cross-power-spectrum of the recorded mixtures can be estimated using a sample average as follows:

$$\hat{\mathbf{R}}_x(\omega, t_k) = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{x}(\omega, t + lT)\mathbf{x}^{\mathbf{H}}(\omega, t + lT) \tag{2.2.3}$$

where T is the block length of the DFT.

The cost function J_m which is used to diagonalize $\mathbf{R}_y(\omega, t_k)$ is defined as:

$$J_m = \sum_{w=1}^{T}\sum_{k=1}^{K}||E(\omega, t_k)||_F^2 \tag{2.2.4}$$

where $E(\omega, t_k) = \mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t)\mathbf{W}^{\mathbf{H}}(\omega) - \mathbf{\Lambda}_s(\omega, t_k)$ and $||.||_F^2$ is the squared Frobenius norm. The cost function J_m is computed at $t_k = kTL$ where

$k = 1, \ldots, K$ and K is the number of matrices to diagonalize. In order to minimize J_m the method of steepest descent is used

$$\frac{\partial J_m}{\partial \mathbf{W}^*(\omega)} = 2 \sum_{k=1}^{K} E(\omega, t_k) \mathbf{W}(\omega) \hat{\mathbf{R}}_x(\omega, t_k) \tag{2.2.5}$$

and the update equation for the separating matrix at each frequency bin is written as

$$\mathbf{W}_{j+1}(\omega) = \mathbf{W}_j(\omega) - \mu \sum_{k=1}^{K} E(\omega, t_k) \mathbf{W}_j(\omega) \hat{\mathbf{R}}_x(\omega, t_k) \tag{2.2.6}$$

where j is the iteration index and μ is the learning rate.

A serious problem of the Parra-Spence algorithm is permutation. $E(\omega, k)$ is insensitive to permutation of the coordinates. Only consistent permutations for all frequencies will result in the proper reconstruction of the sources. A solution to the permutation problem is proposed in [14] by imposing a smoothness constraint on the separating filters. This can be achieved by constraining the filter length Q to be much less than the size of the DFT. In other words, $\mathbf{W}(\tau) = 0$ for $\tau > Q$ and $Q << T$, but this is not always successful.

2.2.2 HOS: Independent Component Analysis (ICA)

ICA as defined in [39] is a statistical model in which the observed data are generated as a linear combination of the original sources considered as latent variables. These variables are assumed non-Gaussian and independent and hence they are termed independent components. The objective is to estimate the underlying hidden variables and the mixing matrix.

The ICA generative model is expressed as follows

$$\mathbf{x} = \mathbf{H}\mathbf{s} \tag{2.2.7}$$

where $\mathbf{x}$ is the observed vector, $\mathbf{s}$ is the vector of statistically independent latent variables and $\mathbf{H}$ is the unknown mixing matrix.

The core of the theory of ICA is based on the realization that the above model is identifiable under the following assumptions:

- The components s_i are mutually statistically independent, which implies that the joint probability density functions of the sources can be factorized as the product of the marginal distribution of the individual components: $P(s_1, ..., s_N) = \Pi_{i=1}^{N} P(s_i)$
- The components s_i have non-Gaussian distributions with at most one Gaussian component.
- The mixing matrix $\mathbf{H}$ is square ($N = M$) and invertible.

Based on these assumptions, each component is determined up to a multiplying scale factor, i.e. the scales and signs of the components are not determined. In addition, the order of the components cannot be determined. Most ICA algorithms perform the separation in two steps [15]: A preliminary spatial whitening followed by the actual ICA estimation. The first step is also referred to as principal component analysis (PCA). Both steps are discussed in the following subsections.

Principal Component Analysis PCA

The objective of PCA is to transform the observed vector $\mathbf{x}$ into a vector $\mathbf{z}$ with spatially uncorrelated components such that

$$E(\mathbf{z}\mathbf{z}^T) = I \tag{2.2.8}$$

The whitening is usually performed after data centering in which the mean is subtracted from the observed data vector. PCA might be done using eigenvalue decomposition (EVD) of the covariance matrix $E(\mathbf{x}\mathbf{x}^\mathbf{T}) = \mathbf{E}\mathbf{D}\mathbf{E}^\mathbf{T}$

where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $E(\mathbf{x}\mathbf{x}^\mathbf{T})$ and $\mathbf{D}$ is the diagonal matrix of its eigenvalues $\mathbf{D} = diag(d_1, \ldots, d_n)$. $E(\mathbf{x}\mathbf{x}^\mathbf{T})$ is computed as a time average using samples of the observed vector $\mathbf{x}(1), \ldots, \mathbf{x}(\mathrm{T})$. The whitened vector $\mathbf{z}$ is obtained as follows

$$\mathbf{z} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^\mathbf{T}\mathbf{x} = \tilde{\mathbf{H}}\mathbf{s} \tag{2.2.9}$$

where $\tilde{\mathbf{H}}$ denotes the new mixing matrix after whitening. PCA reduces the number of unknown parameters. Instead of n^2 elements of the mixing matrix $\mathbf{H}$, only $n(n-1)/2$ elements of the new orthogonal matrix $\tilde{\mathbf{H}}$ need to be estimated.

FastICA

The output vector $\mathbf{z}$ of PCA has uncorrelated components but whitening does not result in a unique decomposition of the data since any orthogonal transform $\mathbf{Q}\mathbf{z}$ is also white, $\mathbf{Q}$ being any orthogonal matrix. For non-Gaussian variables, whitening does not imply independence, there is more information in the data that needs to be exploited. This is performed through optimizing some objective/contrast function related to the non-Gaussianity of the components. Typically, the absolute value of kurtosis or fourth-order cumulant is used, it is equal to zero for a Gaussian variable and greater than zero for most non-Gaussian random variables. However, kurtosis estimation lacks robustness due to its sensitivity to outlier values [40]. Negentropy is another measure of non-Gaussianity, that is closely related to the entropy which is the basic concept of information theory. It is always non-negative and equal to zero only for a Gaussian variable. The estimation of negentropy is complex and is often approximated using higher order moments which suffer as well from non-robustness [15]. It can also be approximated using the maximum entropy principle to provide more robust estimation [41]. Other

measures are related to the minimization of the mutual information or maximum likelihood estimation [42], [43]. Different approaches to the objective functions are discussed in detail in [15].

The fast fixed point algorithm (FastICA) proposed in [39] is presented in this section. In FastICA, HOS are implicitly embedded into the algorithm by arbitrary non-linearities, which have proven to be more robust against atypical values and more computational efficient compared to kurtosis-based ICA methods [39]. The non-linear function chosen for this work is $G(y) = log(a + y)$ where $a = 0.1$. In the one-unit version of ICA, the contrast function at each frequency bin is expressed as follows

$$J_G(\mathbf{w}) = E\{G(|\mathbf{w}^H\mathbf{z}|^2)\} \tag{2.2.10}$$

where $\mathbf{z}$ is the whitened vector and $\mathbf{w}$ is a column vector of the separating matrix $\mathbf{W}$.

Optimizing $E\{G(|\mathbf{w}^H\mathbf{z}|^2)\}$ under the constraint $E\{|\mathbf{w}^H\mathbf{z}|^2\} = ||\mathbf{w}||^2 = 1$ can be performed by calculating the gradient and equating to zero, i.e.

$$\nabla E\{G(|\mathbf{w}^H\mathbf{z}|^2)\} - \beta\nabla E\{|\mathbf{w}^H\mathbf{z}|^2\} = 0 \tag{2.2.11}$$

where $\beta \; \epsilon \; \mathbb{R}$. The Newton method is used to solve (2.2.11) and the fixed point algorithm for one unit is expressed as [39]

$$\mathbf{w}^+ = E\{\mathbf{z}(\mathbf{w}^H\mathbf{z})^* g(|\mathbf{w}^H\mathbf{z}|^2)\} - E\{g(|\mathbf{w}^H\mathbf{z}|^2)\} + |\mathbf{w}^H\mathbf{z}|^2 g'(|\mathbf{w}^H\mathbf{z}|^2)\mathbf{w}\} \tag{2.2.12}$$

$$\mathbf{w}_{new} = \frac{\mathbf{w}^+}{||\mathbf{w}^+||} \tag{2.2.13}$$

In order to prevent units from converging to the same maxima, the outputs are decorrelated after every iteration using Gram-Schmidt-like decorrela-

tion [18]. After estimating of $\mathbf{w}_1, .., \mathbf{w}_p$, during the estimation of $\mathbf{w}_{p+1}$, the projections of the previously estimated p vectors are subtracted from $\mathbf{w}_{p+1}$ followed by normalization of $\mathbf{w}_{p+1}$ as explained in (2.2.14) and (2.2.15)

$$\mathbf{w}_{p+1} = \mathbf{w}_{p+1} - \sum_{j=1}^{P} \mathbf{w}_j \mathbf{w}_j^H \mathbf{w}_{p+1} \tag{2.2.14}$$

$$\mathbf{w}_{p+1} = \frac{\mathbf{w}_{p+1}}{||\mathbf{w}_{p+1}||} \tag{2.2.15}$$

This can also be achieved through a symmetric decorrelation as follows

$$\mathbf{W} = \mathbf{W}(\mathbf{W}^H \mathbf{W})^{-1/2} \tag{2.2.16}$$

The main problem of ICA applied in the frequency domain is the permutation of the solutions over different frequency bins. This implies that permutations should be sorted out for the separating matrices at each frequency bin so that the signals in the time domain can be reconstructed properly. The new approach termed IVA proposed in [19] was designed to solve the permutation problem by optimizing a new cost function modelling the inter-frequency dependencies of data. This was achieved by improving the modelling of the source priors as explained in the following section.

2.2.3 HOS: Independent Vector Analysis

Conventionally, in the ICA literature, the source priors are assumed independent at each frequency bin and are modelled using a super-Gaussian distribution such as the Laplacian distribution [32]. By using higher order dependencies and considering the sources as vectors with a multivariate super-Gaussian distribution, IVA simultaneously estimates the unmixing matrix while theoretically avoiding any permutation indeterminacy [19]. The cost function proposed for multivariate source separation is the Kullback-Leibler

(KL) divergence. Statistically, KL divergence is used as a measure of independence between two functions. In this case, the two functions are the exact joint probability density functions of the estimated sources $p(\mathbf{y}_1, ..., \mathbf{y}_N)$ and the product of approximated probability density functions of the individual source vectors $\Pi_1^N q(\mathbf{y_i})$.

$$
\begin{aligned}
C &= KL\Big(p(\mathbf{y}_1, ..., \mathbf{y}_N || \Pi_1^N q(\mathbf{y}_i)\Big) \\
&= \int p(\mathbf{y}_1 ... \mathbf{y}_N) log \frac{p(\mathbf{y}_1, ..., \mathbf{y}_N)}{\Pi_1^N q(\mathbf{y_i})} d\mathbf{y}_1 ... d\mathbf{y}_N \\
&= \int p(\mathbf{x}_1 ... \mathbf{x}_M) log p(\mathbf{x}_1 ... \mathbf{x}_M) d\mathbf{x}_1 ... d\mathbf{x}_M \\
&\quad - \sum_{k=1}^{K} log|det \mathbf{W}^{(k)}| - \sum_{i=1}^{N} \int p(\mathbf{y_i}) log q(\mathbf{y_i}) d\mathbf{y_i} \\
&= const. - \sum_{k=1}^{K} log|det \mathbf{W}^{(k)}| - \sum_{i=1}^{N} E log q(\mathbf{y_i})
\end{aligned}
\tag{2.2.17}
$$

where $\mathbf{y}_i = [y_i^{(1)} \dots y_i^{(K)}]^T$, $\mathbf{x}_i = [x_i^{(1)} \dots x_i^{(K)}]^T$ and K is the number of frequency bins.

The quantity $\int p(\mathbf{x}_1 ... \mathbf{x}_M) log p(\mathbf{x}_1 ... \mathbf{x}_M) d\mathbf{x}_1 ... d\mathbf{x}_M$ is the entropy of the observations and it is constant. In the above equations, each source is multivariate and the KL divergence is minimum when the sources are as independent as possible while the dependency between the components of each individual source vector is still preserved.

The gradient descent method is used to minimize this contrast function; the derivation of the Newton method with fixed point iteration similar to the FastICA can be found in [19]. By differentiating the cost function relative to the coefficients of the separating matrices w_{ij} the gradients can be written

as follows [32]

$$\triangle w_{ij}^{(k)} = -\frac{\partial C}{\partial w_{ij}^{(k)}} = w_{ij}^{-H(k)} - E\phi^{(k)}(y_i^{(1)} \dots y_i^{(K)}) x_j^{*(k)} \tag{2.2.18}$$

where $(\mathbf{W}^{(k)-1})^H \equiv \{w_{ij}^{-H(k)}\}$. By multiplying both sides of the gradient equation by $\mathbf{W}^{(k)H}\mathbf{W}^{(k)}$ the natural gradient algorithm [44] is obtained.

$$\triangle w_{ij}^{(k)} = \sum_{l=1}^{L} \Big(I_{il} - E\phi^{(k)}(\mathbf{y}_i^{(1)} \dots \mathbf{y}_i^{(K)}) \mathbf{y}_l^{*(k)} \Big) w_{lj}^{(k)} \tag{2.2.19}$$

where I_{il} is only unity when $i = l$ and 0 otherwise. The update rule is

$$w_{ij}^{(k)new} = w_{ij}^{(k)old} + \eta \triangle w_{ij}^{(k)} \tag{2.2.20}$$

The non-linear score function $\phi^{(k)}$ is given as

$$\phi^{(k)}\big(\mathbf{y}_i^{(1)} ... \mathbf{y}_i^{(K)}\big) = -\frac{\partial log q\big(\mathbf{y}_i^{(1)} \dots \mathbf{y}_i^{(K)}\big)}{\partial \mathbf{y}_i^{(k)}} \tag{2.2.21}$$

The major difference between ICA and IVA is the form of the score function. The source prior of a vector with frequency independent Laplacian distribution is expressed as

$$P(\mathbf{s}_i) = \Pi_{k=1}^{K} p\Big(s_i^{(k)}\Big) = \alpha \Pi_{k=1}^{K} exp\Big(- \frac{|s_i^{(k)} - \mu_i^{(k)}|}{\sigma_i^{(k)}} \Big) \tag{2.2.22}$$

where α is the normalization term, $\mu_i^{(k)}$ is the mean and $(\sigma_i^{(k)})^2$ is the variance of the ith source at the kth frequency bin. Assuming zero mean and unit variance the score function can be written as

$$\phi^{(k)}\Big(\mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(K)}\Big) = \frac{\partial \sum_{k=1}^{K} |y_i^{(k)}|}{\partial y_i^{(k)}} = \frac{y_i^{(k)}}{|y_i^{(k)}|} \tag{2.2.23}$$

It can be seen from (2.2.23) that $\phi^{(k)}$ depends only on a single variable

and hence is not a multivariate function. On the other hand, the source prior proposed in [32] representing the inter-frequency dependencies is a dependent multivariate super-Gaussian distribution [45].

Assuming a K dimensional random variable defined by

$$\mathbf{s}_i = \sqrt{v}.\mathbf{z}_i + \mu_i \tag{2.2.24}$$

where μ_i is K dimensional deterministic variable, $\mathbf{z}_i$ is a K dimensional random variable and v is a scalar random variable. $\mathbf{z}_i$ has a Gaussian distribution with zero mean and Σ_i as covariance matrix

$$p(\mathbf{z}_i) = \alpha_z exp\Big(- \frac{\mathbf{z}_i^H \Sigma_i^{-1} \mathbf{z}_i}{2}\Big) \tag{2.2.25}$$

where α_z is a normalization term, whereas v has a gamma distribution defined by

$$p(v) = \alpha_v v^{\frac{K-1}{2}} exp\Big(- \frac{v}{2}\Big) \tag{2.2.26}$$

where α_v is a normalization term.

Therefore, the conditional distribution $p(\mathbf{s}_i/v)$ is Gaussian with mean μ_i and covariance $v\Sigma_i$. The marginal distribution of $\mathbf{s}_i$ can be written as

$$\int_0^\infty p(\mathbf{s}_i|v)p(v)d(v) \tag{2.2.27}$$

$$= \hat{\alpha} \int_0^\infty \sqrt{v} exp\Big(- \frac{1}{2}\Big(\frac{(\mathbf{s}_i - \mu_i)^H \Sigma_i^{-1}(s_i - \mu_i)}{v} + v\Big) d(v) \tag{2.2.28}$$

$$= \alpha exp\Big(- \sqrt{(\mathbf{s}_i - \mu_i)^H \Sigma_i^{-1}(s_i - \mu_i)}\Big) \tag{2.2.29}$$

Since the frequency domain separation is achieved after conversion using the Fourier transform, components from different frequency bins are uncorre-

lated and have zero mean. Equation (2.2.29) can be written in this form

$$P(\mathbf{s}_i) = \alpha exp\left(- \sqrt{\left| \frac{s_i^{(k)}}{\sigma_i^{(k)}} \right|^2} \right) \tag{2.2.30}$$

where $\sigma_i^{(k)}$ is related to the standard deviation of the ith source at the kth frequency bin. Assuming unit variance and scaling the frequency components after separation, the multivariate score function can be written as

$$\phi^{(k)}\left(y_i^{(1)} \dots y_i^{(K)}\right) = \frac{\partial \sqrt{\sum_{k=1}^{K} |y_i^{(k)}|^2}}{\partial y_i^{(k)}} = \frac{y_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |y_i^{(k)}|^2}} \tag{2.2.31}$$

Equation (2.2.31) is one form of the multivariate score function used to represent inter-frequency dependency of the sources. However, this score function is not unique it varies according to the type of dependency. Finding suitable score functions adapted to the nature of the sources is a subject of on-going research [32]. IVA solves the permutation problem inherently without the need of any post-processing step and the scaling problem is solved using the minimal distortion principal [46]. Once the algorithm is finished the separating matrix is scaled as follows

$$\mathbf{W}^{(k)} \leftarrow diag\big(\mathbf{W}^{-1(k)}\big)\mathbf{W}^{(k)} \tag{2.2.32}$$

Finally, an inverse Fourier transform is applied to reconstruct the sources in the time domain. IVA can be viewed as a generalization of the ICA algorithm where the multivariate treatment of the observations recovers the independence of the vectors while preserving the natural frequency dependencies within the components of the same source vector.

In order to evaluate and compare the performance of the FDCBSS algorithms, various performance measures commonly used in BSS are presented

in the following section.

2.3 Performance measures

Different performance measures can be used for the evaluation of blind audio speech source separation, they can be categorized into objective and subjective measures. Objective measures/indices measure the quality of the estimated mixing matrix or the estimated sources. Since these measures require the knowledge of the original system parameters which essentially are not available in practice, subjective measures can be used such as the Mean Opinion Score (MOS) tests for voice specified by the ITU-T recommendation P.800 [47]. The following objective measures are used for comparing the FDCBSS techniques:

2.3.1 Signal-Interference-Ratio SIR

The SIR proposed in [48] is expressed as:

$$SIR = 10log_{10}\frac{||\mathbf{s}_{target}||_2^2}{||\mathbf{e}_{intf}||_2^2} \tag{2.3.1}$$

where

$$\mathbf{s}_{target} = \langle \hat{\mathbf{s}}_i, \mathbf{s}_i \rangle \mathbf{s}_i / ||\mathbf{s}_i||_2^2 \tag{2.3.2}$$

$$\mathbf{e}_{intf} = \sum_{i \neq j} \langle \hat{\mathbf{s}}_i, \mathbf{s}_j \rangle \mathbf{s}_i / ||\mathbf{s}_j||_2^2 \tag{2.3.3}$$

and s_{target} is the source of interest, e_{intf} represents the interference resulting from other sources and the inner product $\langle \hat{\mathbf{s}}_i, \mathbf{s}_j \rangle = \sum_{t=1}^{T} \hat{s}_i(t) s_j(t)$.
The SIR can be estimated in the frequency domain and is expresssed as:

$$SIR = 10log_{10}\frac{\Sigma_i \Sigma_\omega |\mathbf{H}_{ii}(\omega)|^2 \langle |s_i(\omega)|^2 \rangle}{\Sigma_{i,j,i \neq j} \Sigma_\omega |\mathbf{H}_{ij}(\omega)|^2 \langle |s_j(\omega|^2 \rangle} \tag{2.3.4}$$

where $\mathbf{H}_{ii}$ and $\mathbf{H}_{ij}$ represent the diagonal elements the off diagonal elements of the mixing matrix, respectively and $s_i(\omega)$ is the frequency domain representation of the target source. A high SIR is achieved when the sources are mutually orthogonal.

2.3.2 Performance index PI

The PI is a function of the global matrix $\mathbf{G} = \mathbf{WH}$ and it is expressed as [49]:

$$PI(G) = \Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{k=1}^{m}\frac{abs(\mathbf{G}_{ik})}{max_k abs(\mathbf{G}_{ik})}-1\Big)\Big]+\Big[\frac{1}{m}\sum_{k=1}^{m}\Big(\sum_{i=1}^{n}\frac{abs(\mathbf{G}_{ik})}{max_i abs(\mathbf{G}_{ik})}-1\Big)\Big] \tag{2.3.5}$$

where $\mathbf{G}_{ik}$ denotes the elements of the global matrix G. This criterion allows the performance evaluation at every frequency bin, as it approaches zero the better the separation. The upper bound of $PI(G)$ depends on the normalization factor.

2.3.3 Permutation Evaluation

For the case of N=M=2, $[abs(\mathbf{G}_{11}\mathbf{G}_{22}) - abs(\mathbf{G}_{12}\mathbf{G}_{21})]$ can be used to indicate whether the outputs are permuted or not. It is greater than zero for a permutation free separation [50].

2.4 Experimental results and discussions

The aim of this section is to present the results obtained when applying the three FDCBSS methods on speech signals. Simulations are performed using two real recorded speech signals. The dimensions of the rooms are $5m \times 5m \times 5m$. The sources are assumed to be positioned at [1 2 1.5] and [3.5, 2, 1.5]. The microphones are positioned at [2.47, 2.5, 1.5] and [2.53, 2.5,

1.5] relative to the reference of the room, which is the corner. The length of the Fast Fourier transform $T = 1024$ and the sampling frequency of the speech signals is 8820Hz. The performance of the algorithms is evaluated using the SIR, PI as well as the permutation index. Figure 2.2 shows the results obtained when applying the Parra-Spence algorithm.

The SIR computed at the input is equal to -0.03 dB and the output =10.01

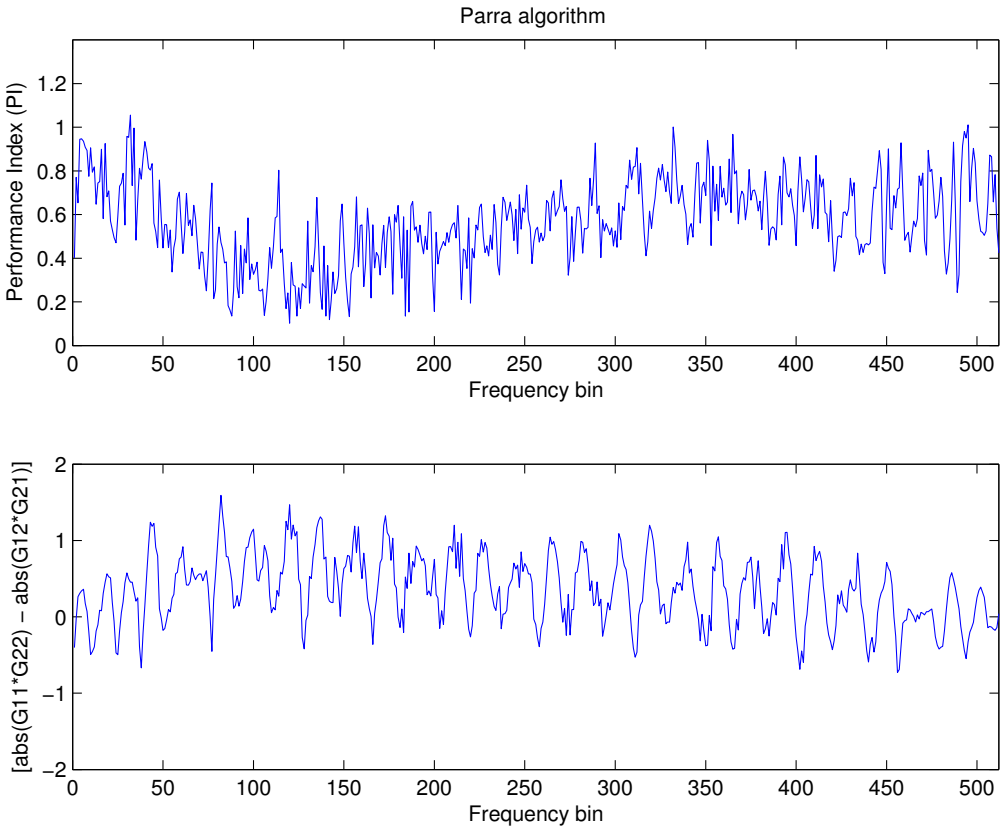


Figure 2.2: Performance index and permutation evaluation at each frequency bin for the Parra-Spence algorithm.

dB with an improvement of 10.04dB. The permutation is solved using a constraint on the filter length in the time domain which provides a solution but not in all frequency bins as shown in Figure 2.2 where $[abs(\mathbf{G}_{11}\mathbf{G}_{22}) - abs(\mathbf{G}_{12}\mathbf{G}_{21})]$ is less than zero. Permutation results in lower SIR ratio which would be dramatically improved with a complete solution of the permutation

problem. Figure 2.3 shows the results obtained when applying the FastICA algorithm. FastICA does not provide a solution for permutation this is why a post-processing step should be performed in order to reconstruct the sources properly. The SIR computed at the input is equal to -0.03 dB and the output =20.58 dB with an improvement of 20.61 dB.

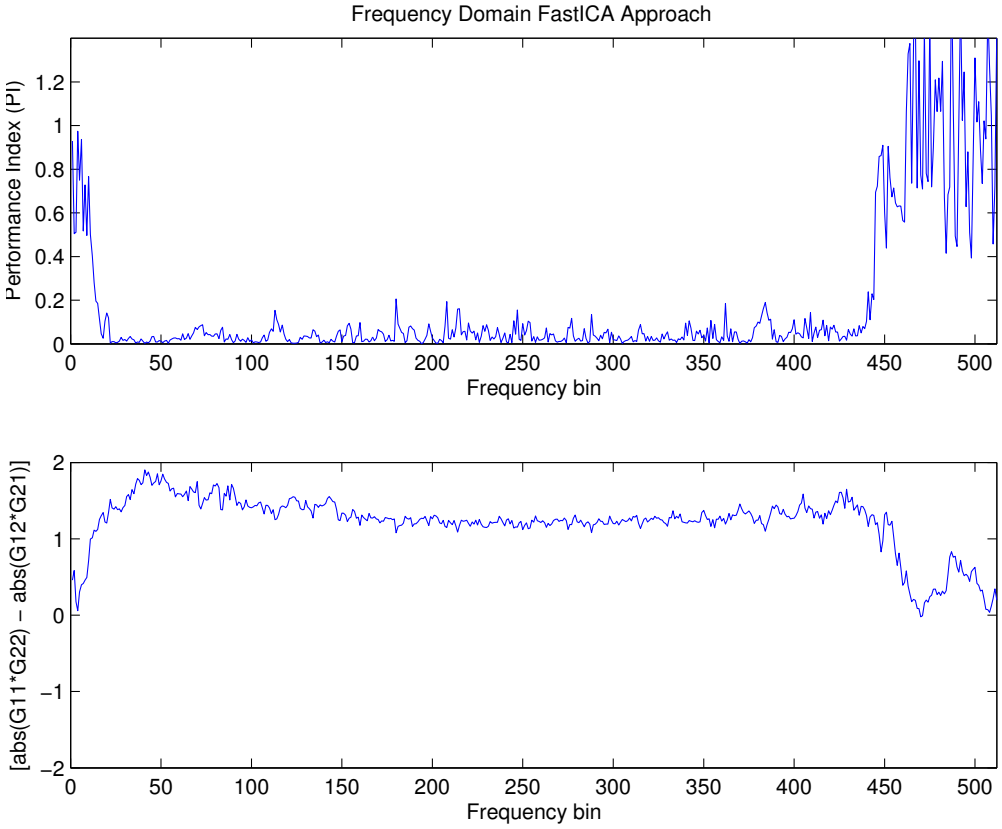


Figure 2.3: Performance index and permutation evaluation at each frequency bin for the FastICA algorithm.

Figure 2.4 shows the performance indices of the IVA algorithm, permutation is solved within the algorithm with no need for post-processing step or constraints on the filter length. The SIR computed at the input is equal to -0.03 dB and the output =23.80 dB with an improvement of 23.83 dB. For the three algorithms the separation performance is poor for low frequencies

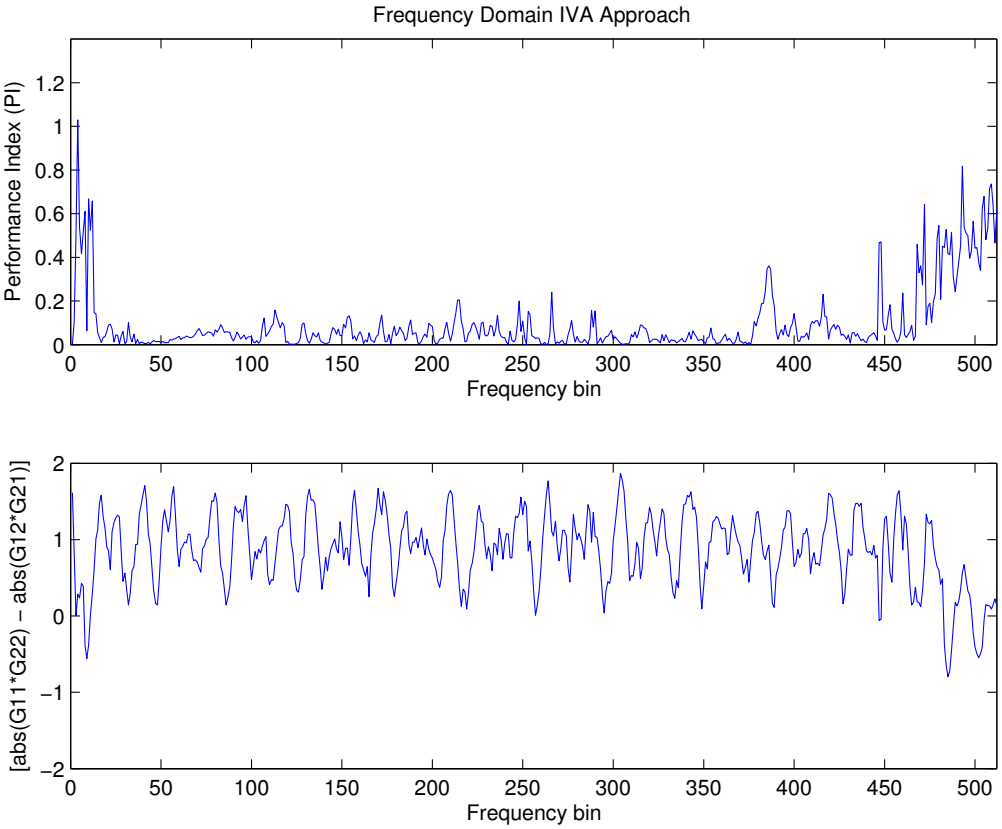


Figure 2.4: Performance index and Permutation evaluation at each frequency bin for the IVA algorithm.

due to the inter-microphone spacing (6cm) which is much smaller than the wavelength at the low frequencies (spatial aliasing). Low performance can also be depicted for the three algorithms at high frequencies and this is due to the low energy of speech signals at high frequencies. It can also be seen that involving HOS improves the quality of the separation which can be seen by comparing PI and SIR of (FastICA /IVA) and Parra-Spence algorithms.

The FDCBSS methods discussed in this section are based on linear filtering, which involves the pseudo-inversion of the mixing matrices while relying on some assumptions made about the nature of the sources such as non-stationarity (SOS) or independence (HOS). However, the sparseness of the speech sources in a given basis is a more powerful assumption [20] as it

imposes a simpler structure on the mixing process that can be useful to perform the separation even in the challenging case of less sensors than sources as explained in the next section.

2.5 Underdetermined BSS

The human auditory system is able to solve the CPP problem effortlessly using only two ears. The T-F approach which originated in the field of computational auditory scene analysis (CASA) aims at designing a machine imitating the human capability of extracting a speech of interest in the presence of other simultaneous sources using only two microphones. Extensive effort has been dedicated in the fields of CASA and ICA to develop methods solving underdetermined BSS [22], they all rely on the sparsity of the acoustic sources in the frequency domain. The concept of W-disjoint orthogonality introduced by Yilmaz and Rickard in [24], on which T-F masking approaches are based, is explained below.

W-disjoint Orthogonality

For two continuous signals s_j and s_k and window function $W(t)$, the W-disjoint orthogonality assumption can be expressed as follows [24]

$$\mathcal{F}^W[s_j](\tau_f, \omega)\mathcal{F}^W[s_k](\tau_f, w) = 0, \forall \tau_f, \omega \tag{2.5.1}$$

where

$$\mathcal{F}^W[s_j](\tau_f, \omega) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t - \tau_f) s_j(t) e^{-i\omega t} dt \tag{2.5.2}$$

where $\mathcal{F}^W[.]$ is the windowed Fourier transform. Since the energy at each spectrogram point of an active source is rarely zero, (2.5.1) is only approximately satisfied. However, speech is known to be sparse in the T-F domain, i.e. most of the energy of each source is captured in a small percentage of its T-F coefficients. The advantage of this sparsity is that the probability

of simultaneous overlapping of two active speech sources is low and hence speech signals are said to be approximately disjoint or orthogonal.

ICA and T-F masking

The T-F masking ICA based method combines the sparseness assumption with the ICA approach, it was firstly proposed in [51] to separate three sources using only two sensors. The separation is performed in two steps; in the first step a binary mask is generated to extract one source then ICA is applied to the mixtures after removal of the first source from the observed data. This approach was modified in [52] by replacing the binary mask with a continuous soft mask based on the direction of arrival to minimize music noise resulting from the excessive zero padding associated with binary masking. The second stage in these proposed methods is followed by a post-processing step to solve the ICA permutation problem and properly group the frequency components coming from the same source. This is performed through the maximization of correlation coefficients related to the activity of the same source such as the correlation coefficients of the amplitude envelopes [53]. Unfortunately, these correlation coefficients are not an accurate criterion for deciding whether different frequency components belong to the same source. A more general approach was presented in [53], it also consists of two stages. In the first stage, mixture samples are clustered in the frequency domain using the line orientation separation technique (LOST) [20], [54], [55]. LOST achieves underdetermined separation by identifying lines in a scatter plot. The orientation of each line is determined through an iterative procedure similar to the expectation maximization (EM) algorithm [54], which generates posterior probabilities representing the membership of each observation to a corresponding class. However, the class order is not the same for each frequency. Permutation ambiguities are solved in the second stage through k-means clustering of the posterior probabilities obtained in the first stage.

The proposed permutation alignment gives better performance compared to the method based on amplitude envelopes.

CASA and T-F masking

T-F CASA based methods, on the other hand, rely on the auditory masking principle consisting, as previously mentioned, of two main stages [21]; decomposition of the sound waveform into the T-F domain followed by grouping of the T-F points belonging to the same source based on different cues into separate streams. Various cues can be used to group the relevant points, three spatial interaural cues resulting from the time, phase and level differences between the signals reaching both ears/microphones are mainly used for localization and separation in CASA systems. The interaural time difference (ITD) is caused by the difference in the arrival times of the signals at the left and right ears. The interaural phase difference (IPD) is related to ITD but is more convenient for narrowband signals when it is difficult to differentiate between the delay resulting from more than one cycle and the corresponding delay of less than a cycle. This ambiguity represents a form of spatial aliasing [25]. Whereas, the interaural level difference (ILD) results from the attenuation of the signal reaching the far ear, this is due to the fact that the head obstructs the wavelengths of sounds comparable to its size. This effect is named shadowing and typically occurs for frequencies above 3 to 4 kHz.

Based on W-disjoint orthogonality, each T-F point is dominated by one source and the problem of underdetermined speech separation can be interpreted as a clustering data problem relying on one or more of the aforementioned cues.

In DUET [24], the ratio of T-F representations of the mixtures is used to construct a two-dimensional histogram with only one peak for each source. The location of the peak corresponds to the interaural cues of each source.

Using the histogram, binary T-F masks are obtained to separate the sources from their mixtures. DUET assumes that interaural cues are the same for all frequencies and requires that a separation of less than $\pi c/f_m$ between microphones to avoid spatial aliasing, where f_m is the maximum frequency of the speech sources and c is the speed of sound in air.

Other T-F masking algorithms such as MESSL and MOSPALOSEP [26], [56] generate soft probabilistic masks to avoid musical noise [22]. In MOSPALOSEP, ITDs are mapped into phase differences and modelled using GMMs, whereas in MESSL, GMMs are used for modelling both IPDs and ILDs [25]. In both, clustering is achieved via the EM algorithm.

Several other methods have also been developed in the literature, for example, introducing additional cues for mask estimation such as the use of joint monaural and binaural cues [35], video aided mixing vector and binaural cues [57], joint mixing vector and binaural cues [58], estimating the mask with a spatial covariance model [59] and extending the binaural case to multichannel scenario with directional statistics [60].

The auditory masking principle is known to be more general than ICA based underdetermined speech separation as it is independent of the source distributions (diffuse or sparse) [22]. The main blocks of a probabilistic T-F masking CASA algorithm are depicted in Figure 2.5. The noisy signals $L(t_s)$ and $R(t_s)$ arriving at the left and right microphones respectively, denote the stereo mixtures at discrete time indices t_s. These signals undergo a T-F analysis using the short-time Fourier transform resulting in $L(\omega, t)$ and $R(\omega, t)$. The ratio of these spectrograms at each time frame t and frequency ω is the interaural spectrogram characterized by IPD/ITD and ILD cues [25]. Localization systems rely on one or both of the interaural cues. Following localization, most algorithms perform separation through clustering using finite mixture of distributions.

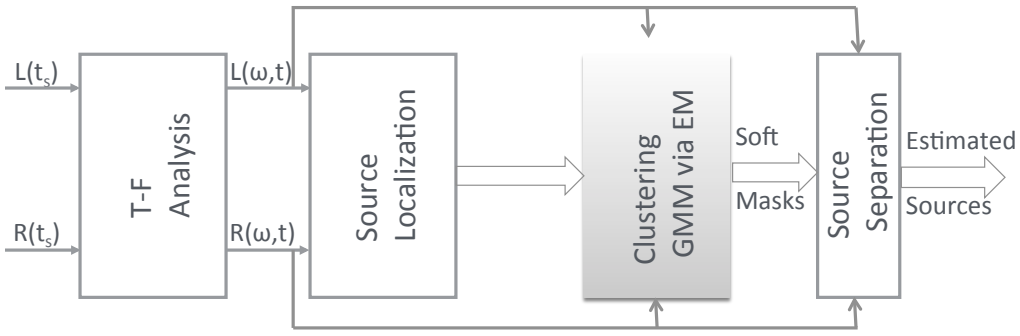


Figure 2.5: Overview of the main processing steps of a typical probabilistic T-F masking algorithm.

MESSL

The state-of-the-art MESSL compared to other underdetermined speech separation methods has offered many advantages. It does not depend on the distributions of the sources in the mixture, does not require prior knowledge of the room, microphones spatial set up and can achieve better separation quality in terms of objective and subjective measures under reverberant conditions [61]. The MESSL framework is flexible and allows the addition of other cues such as source models [62]. Localization of the sound sources in MESSL is considered a key feature enabling the separation, it outperforms other comparable algorithms, with a 40% lower mean absolute error [61]. In anechoic conditions, the separation performance achieved by MESSL compares to that of humans. However, in the presence of reverberations, it reaches approximately 20-25% of human performance [61]. For small azimuthal separation between the target and interferers, the separation performance of MESSL as well as other CASA algorithms relying on spatial cues degrades significantly [25]. Localization in MESSL is explained in the next section. The probabilistic modelling of the interaural cues and the

clustering framework employed for the generation of masks will be discussed thoroughly in the following chapters.

2.5.1 Localization in MESSL

Humans localize sounds in azimuth, elevation and distance; for each dimension different cues can be used. Azimuthal localization depends on the interaural level and time differences. Humans rely on IPDs for azimuthal localization for frequencies below 1.5 kHz and ILD cues for frequencies higher than 4 kHz [63]. Elevation is associated with the pinna cues and distance depends on the direct to reverberation ratio, the level and high frequency content of the sound [61]. In MESSL, localization is essentially azimuthal and is based on the interaural phase difference. The delay associated with each speech source position is estimated using the phase transform (PHAT) [64]. PHAT belongs to the generalized cross-correlation (GCC) framework [65] in which estimates of the delays are computed by maximizing a weighted cross-correlation function between both channels over all possible values of delays.

Assuming a single sound source $x(t)$ reaching two spatially distinct microphones, the observations at the left and right microphones are given by

$$l(t) = a_l x(t - \tau_l) * n_l(t) \tag{2.5.3}$$

$$r(t) = a_r x(t - \tau_r) * n_r(t) \tag{2.5.4}$$

For simplicity, a_l, a_r, τ_l and τ_r are assumed to be frequency independent. The left channel can be expressed in terms of the right channel as follows

$$l(t) = a_{lr} r(t - \tau_{lr}) * n_{lr}(t) = a_{lr} r(t) * \delta(t - \tau_{lr}) * n_{lr}(t) \tag{2.5.5}$$

where the delays, gains and noises have been combined into a_{lr}, τ_{lr} and $n_{lr}(t)$. The relative delay between two signals can be estimated through their cross-correlation defined by

$$r_{lr}(\tau) \equiv \sum_{t=0}^{N-1} l(t)r(t-\tau) \tag{2.5.6}$$

where $\tau = -N+1, \ldots, N-1$,

$$r_{lr}(\tau) = \sum_{t=0}^{N-1} l(t)r\big(-(\tau-t)\big) = l(t) * r(-t) \tag{2.5.7}$$

Using (2.5.5), (2.5.7) can be written as

$$r_{lr}(\tau) = a_{lr}r(t) * \delta(t-\tau_{lr}) * n_{lr}(t) * r(-t) \tag{2.5.8}$$

$$= a_{lr}r_{rr}(\tau) * \delta(t-\tau_{lr}) * n_{lr}(t) \tag{2.5.9}$$

where $r_{rr}(\tau)$ is the auto-correlation of $r(t)$. Taking the Fourier transform of both sides gives

$$S_{lr}(\omega) = L(\omega)R^*(\omega) = a_{lr}S_{rr}(\omega)e^{-j\omega\tau_{lr}}N_{lr}(\omega) \tag{2.5.10}$$

The cross-correlation is a copy of the autocorrelation but peaked at τ_{lr} instead of zero. Thus, the time-delay between the two signals can be estimated by maximizing the cross-correlation.

By computing the cross-correlation in the frequency domain, its inverse $r_{lr}(\tau)$ can be expressed as

$$r_{lr}(\tau) \propto \sum_{k=-N/2}^{N/2-1} L(\omega)R^*(\omega)e^{j\omega\tau} \quad \omega = \frac{2\pi k}{N} \tag{2.5.11}$$

where $\omega = \frac{2\pi k}{N}$. The Generalized Cross-Correlation introduces a weighting

function into (2.5.11) such that the cross-correlation is expressed as [65]

$$g(\tau) = \sum_{\omega} \psi(\omega) L(\omega) R^*(\omega) e^{j\omega\tau} \tag{2.5.12}$$

In PHAT, $\psi(\omega)$ is chosen to cancel the magnitude of the left and right channels [64]

$$\psi(\omega) = \frac{1}{|L(\omega)||R^*(\omega)|} \tag{2.5.13}$$

such that the cost function is given by

$$g(\tau) = \sum_{\omega} \frac{L(\omega) R^*(\omega) e^{j\omega\tau}}{|L(\omega)||R^*(\omega)|} \tag{2.5.14}$$

The proposed localization in MESSL has many advantages over other conventional generalized cross-correlation methods. It has proven to estimate the true delay in reverberant environments, can localize multiple sources and does not depend on statistical assumptions related to the Gaussianity or stationarity of the sound sources [61]. This localization method will be used in the subsequent chapters to initialize the speech separation.

2.6 Summary

This chapter gives an overview of the different FDCBSS approaches used to solve the cocktail party problem. Three non-sparse main techniques were thoroughly studied and discussed; Parra-Spence algorithm, ICA and IVA. The Parra-Spence algorithm relies on SOS while exploiting the non-stationarity of the speech signals. ICA and IVA are based on HOS which improves the separation quality but requires more data and are computationally more complex. HOS is represented by non-linear contrast functions which are optimized through learning algorithms such as the gradient descent algorithm or fixed-point methods to estimate the separating matrix. SOS and ICA methods suffer from the permutation problem which should be solved since only consistent permutations for all frequencies reconstruct the original signals properly. The permutation problem can be solved by imposing a smoothness constraint that translates into smoothing the separating filter or other post-processing steps. However, by introducing a new multivariate cost function and a dependency model capturing the inter-frequency dependencies IVA achieves the separation and solves the permutation problem. By comparing the three techniques, it can be concluded that IVA offers a complete solution for FDCBSS in which good separation is achieved while avoiding the permutation problem. This chapter also introduces the T-F masking concept used to separate a mixture of speech sources in the case of less sensors than sources, typically two sensors. Various T-F masking ICA and CASA based algorithms are discussed with a focus on the state-of-the-art MESSL algorithm, which uses Gaussian mixture modelling for the interaural cues at each spectrogram point and an EM estimation procedure. Localization in MESSL is explained in detail as it will be used throughout this thesis to initialize the separation.

Chapter 3

MODELLING INTERAURAL CUES WITH STUDENT'S T-DISTRIBUTION FOR ROBUST CLUSTERING IN MESSL

In this chapter, a novel approach to the probabilistic modelling of the interaural cues commonly used in time-frequency (T-F) based speech separation systems is presented. The Student's t-distribution known by its heavy tails can potentially better capture outlier values in the univariate parametric modelling of the T-F points and thereby lead to more accurate probabilistic masks for source separation. Gaussian mixture models (GMMs) used in MESSL for modelling the interaural phase difference (IPD) and the interaural level difference (ILD) cues are replaced by the Student's t-distribution mixture models (SMMs) to better represent the uncertainties introduced by noise, reverberations as well as the statistical non-stationarity of speech signals. Simulation studies based on speech mixtures formed from the TIMIT database confirm the advantage of the proposed approach particularly when

the speech sources are in close proximity.

3.1 Introduction

Using finite mixture models in the statistical analysis of data has enormously increased since 1995 [66]. They have proven to provide a useful and flexible mathematical tool for probabilistic modelling of a large variety of random phenomena and have been successfully applied to various fields including engineering, medicine, biology, astronomy, economics as well as other areas requiring data mining, statistical analysis and machine learning [67]. The reason behind their importance and ever increasing applications is their potential ability to model unknown distributional shapes or as stated in [68] they can provide models for "unobserved population heterogeneity". This can be simply illustrated in Figure 3.1. The left hand side represents the case

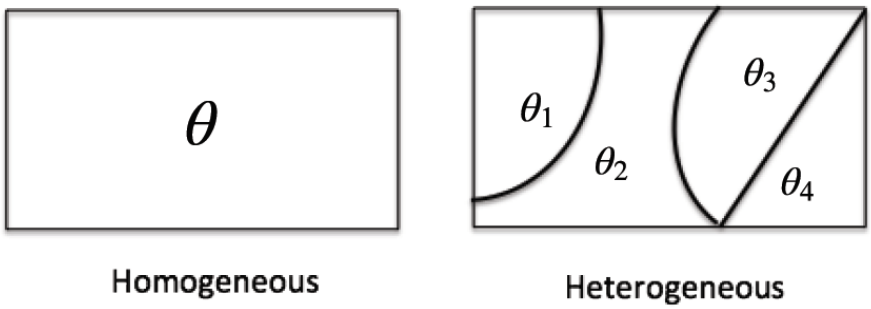


Figure 3.1: The homogeneous population on the left side depicts the case where the observed data can be modelled by one parametric density function. On the other hand, the heterogeneous case represents the variation of the parameter over various subpopulations.

where data are modelled by one-parametric density function $f(y, \theta)$ where θ is the parameter of the population and y is a univariate value belonging to the sample space Y. However, this model cannot define the parameter variation in the case where the population includes different subpopulations having the

same form of the density function but different parameters as shown on the right hand side of Figure 3.1. Since it is not known to which subpopulation or cluster each observation belongs, let $\mathbf{z} = \{z_k\}$ denote the latent vector defining the cluster membership, then $f(y, z_k)$ can be expressed as $f(y|z_k)p(z_k)$ and the marginal density of $f(y)$ is expressed as $\sum_{k=1}^{K} f(y|z_k)p(z_k)$ equivalent to $\sum_{k=1}^{K} f(y;\theta_k)\psi_k$ where K is the number of clusters and ψ_k is the probability of belonging to cluster k defined by the parameter θ_k.
In most natural phenomena, it is practically assumed that $f(y;\theta_k)$ is Gaussian and it is also common to observe elliptical clusters for multivariate continuous data [30]. Within the class of elliptical distributions, the Gaussian density was primarily considered for fitting the mixture of distributions due its analytical convenience.

A Gaussian mixture model for a multivariate $\mathbf{y}$ can be written as [67]

$$p(\mathbf{y}) = \sum_{k=1}^{K} \psi_k \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k, \Sigma_k) \tag{3.1.1}$$

where $\boldsymbol{\mu}_k$ is k-th component mean vector, Σ_k is the corresponding covariance matrix and ψ_k is the mixing coefficient. Assuming $\mathbf{z}$ is a K dimensional binary random vector in which a particular element z_k is equal to 1 and all other elements are equal to 0 such that $\sum_k z_k = 1$. Since there are K possible states of the vector $\mathbf{z}$ depending on which element is nonzero, the marginal distribution $p(\mathbf{z})$ can be written in terms of the mixing coefficients as follows

$$p(\mathbf{z}) = \prod_{k=1}^{K} \psi_k^{z_k} \tag{3.1.2}$$

where $p(z_k = 1) = \psi_k$ and ψ_k must satisfy the following conditions

$$0 \leq \psi_k \leq 1 \tag{3.1.3}$$

$$\sum_{k=1}^{K} \psi_k = 1 \tag{3.1.4}$$

The conditional distribution of $\mathbf{y}$ given $\mathbf{z}$ is Gaussian

$$p(\mathbf{y}|z_k = 1) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k, \Sigma_k) \tag{3.1.5}$$

which can also be expressed as

$$p(\mathbf{y}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k)^{z_k} \tag{3.1.6}$$

The marginal distribution of $\mathbf{y}$ is determined by summing the joint distribution $p(\mathbf{y}, \mathbf{z})$ over all possible states of $\mathbf{z}$

$$p(\mathbf{y}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{y}|\mathbf{z}) = \sum_{k=1}^{K} \psi_k \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k) \tag{3.1.7}$$

The conditional probability $p(z_k = 1|\mathbf{y})$ can also be expressed using Bayes' theorem. Let ν_k denote $p(z_k = 1|\mathbf{y})$ then

$$\nu_k = \frac{p(z_k = 1)p(\mathbf{y}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{y}|z_j = 1)} \tag{3.1.8}$$

$$= \frac{\psi_k \mathcal{N}(\mathbf{y}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \psi_j \mathcal{N}(\mathbf{y}|\mu_j, \Sigma_j)} \tag{3.1.9}$$

where ψ_k is viewed as the prior probability of $z_k = 1$ and ν_k is the corresponding posteriori probability. This posteriori probability ν_k can also be considered as the responsibility that component k takes for explaining the given observation $\mathbf{y}$.

Two sets of data are shown in Figure 3.2, the set on the right hand side is said to be complete i.e. the values of the latent variables z_k are known [67], whereas the other case represents a classical missing data problem and the data set is said to be incomplete.

In a typical CPP environment, different sound sources are simultane-

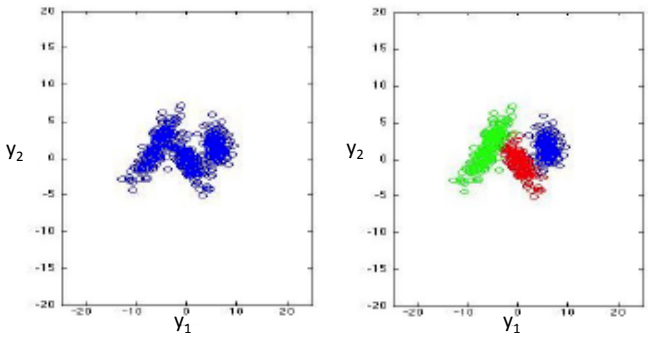


Figure 3.2: Data on the left are obtained from mixing three bivariate Gaussian distributions with different means and precisions. On the right hand side, data are clustered, and each point n in the sample space is associated to a vector $\mathbf{z}_n$ indicating the cluster membership, different clusters have different colours.

ously active and for the target source to be extracted and recognized by the listener, the auditory mixture is partitioned in the T-F domain and the correct fragments are then assigned to their corresponding sources. This process of grouping and segregation represents the auditory scene analysis as explained in [23]. It can be statistically interpreted as a clustering problem, T-F points are simply grouped according to various cues which differ from one algorithm to another. In MOSPALOSEP [56], the interaural time differences (ITDs) are mapped into phase differences used for the separation. Whereas in MESSL [25], as mentioned in Chapter 2, both IPD and ILD cues are used for clustering. Gaussian mixture models define the distribution of the cues as they can be easily fitted iteratively by maximum likelihood estimation (MLE) via the EM algorithm [67].

In this chapter, the use of GMMs in the T-F masking based speech separation is demonstrated through the state-of-art MESSL algorithm. Then, a wider class of distributions is proposed to improve the robustness. In Section 3.2, the probabilistic modelling of the spatial interaural cues employed in MESSL is introduced. In Section 3.3, the EM framework employed to

estimate the model parameters is thoroughly described. Section 3.4 introduces the Student's t-distribution and SMMs are used as an alternative to GMMs for modelling both IPD and ILD cues. Different simulation studies showing the improvement in speech separation are undertaken in Section 3.5. Finally, in Section 3.6, the chapter is summarized.

3.2 Spatial interaural cues

The human auditory system uses efficiently different cues to localize and separate sound waves. MESSL relies essentially on the azimuthal cues which are related to the level and time differences between the signals arriving at the left and right ears as shown in Figure 3.3.

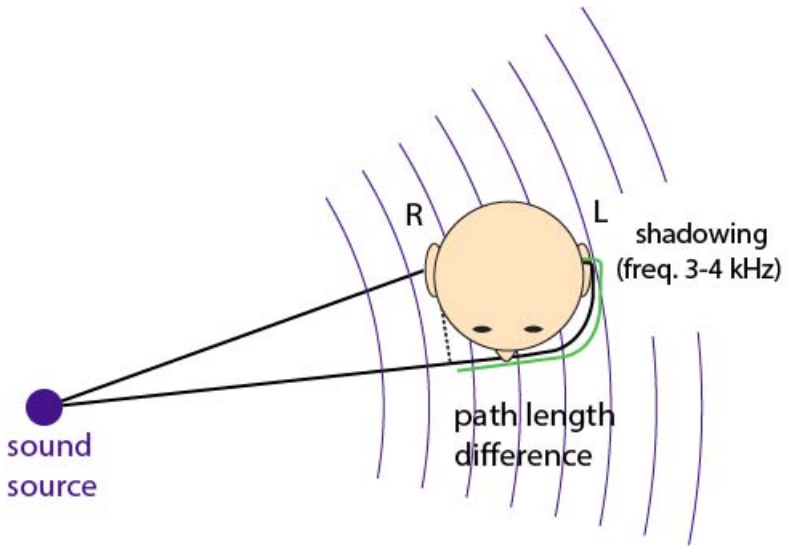


Figure 3.3: Causes of interaural differences.

The interaural time difference (ITD) is known to be ambiguous and only determines a cone of possible source locations namely the "cone of confusion". Humans use pinna cues to solve this confusion [61]. The IPD can be used instead to localize a sound source and can be mapped uniquely to a

specific delay in the absence of spatial aliasing. The ILD results from the attenuation of the signal reaching the far ear, this is due to the fact that the head obstructs the wavelengths of sounds comparable to its size. This effect is named shadowing and typically occurs for frequencies above 3 to 4 kHz [61].

Following [25], $L(\omega, t)$ and $R(\omega, t)$ are assumed to be the spectrograms of a sound source arriving at two spatially distinct microphones such that the interaural spectrogram is expressed as

$$\frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{3.2.1}$$

In order to avoid spatial aliasing [25], the phase residual $\hat{\phi}(\omega, t; \tau)$ expressed as

$$\hat{\phi}(\omega, t; \tau) = arg\Big(e^{j\phi(\omega,t)} e^{-j\omega\tau(\omega)}\Big) \tag{3.2.2}$$

was proposed as an alternative to $\phi(\omega, t)$. The phase residual is the difference between the IPD resulting from a delay of τ samples and the actual IPD and is constrained to the interval $(-\pi, \pi)$.

The phase residual can be modelled with a circular probability distribution particularly the Von Mises distribution [69] but it can also be approximated by a single Gaussian as follows

$$p(\phi(\omega, t)|\tau(\omega), \sigma(\omega)) \approx \mathcal{N}(\hat{\phi}(\omega, t)|\xi(\omega), \sigma^2(\omega)) \tag{3.2.3}$$

where $\xi(\omega)$ and $\sigma^2(\omega)$ denote the frequency dependent mean and variance, respectively.

The interaural level difference can also be modelled by a single Gaussian [70] with frequency-dependent mean $\mu(\omega)$ and variance $\eta^2(\omega)$

$$p(\alpha(\omega, t)|\mu(\omega), \eta^2(\omega)) = \mathcal{N}(\alpha(\omega, t)|\mu(\omega), \eta^2(\omega)) \tag{3.2.4}$$

The ILD and the IPD models are combined together assuming that they are conditionally independent given their parameters as follows

$$p(\phi(\omega,t)\alpha(\omega,t)|\Theta) = \mathcal{N}\Big(\hat{\phi}(\omega,t)|\xi(\omega),\sigma^2(\omega)\Big).\mathcal{N}\Big(\alpha(\omega,t)|\mu(\omega),\eta^2(\omega)\Big) \tag{3.2.5}$$

where Θ represents all of the model parameters. Equation (3.2.5) can be used to determine the likelihood of an observation at any T-F point given the set of parameters Θ. Spectrogram points are assumed to be independent such that the total likelihood can be computed for any set of points as the product of their individual likelihoods.

The parameters of the model described by equation (3.2.5) cannot be estimated directly from the observations, since different points of the spectrogram belong to different sources at different delays $\tau(\omega)$ expressed as [25]

$$\tau(\omega) = \tau + \omega^{-1}\xi(\omega) \tag{3.2.6}$$

where τ is modelled as a discrete random variable used for localization while the parameter $\xi(\omega)$ is varying randomly with frequency in the interval $(-\pi,\pi)$. The number of sources is assumed to be known a priori. On the other hand, the source i dominating each spectrogram point as well as the delay τ are combined into one latent variable $z_{i\tau}(\omega,t)$. This parameter is equal to one with a corresponding probability $\psi_{i\tau}$, if the spectrogram point belongs to source i and delay τ and is zero otherwise. In other words, $z_{i\tau}(\omega,t) \in \{0,1\}$ and $\sum_{i,\tau} z_{i\tau}(\omega,t) = 1$.

The maximum likelihood parameters can be estimated with an expectation maximization (EM) algorithm as explained in the following section. Details of the derivation of EM for mixtures of Gaussians can be found in Appendix A.

3.3 EM for GMMs

Let $\Theta \equiv \{\xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\}$ denote the set of parameters. The total likelihood for a given observation can be expressed as

$$\mathcal{L}(\Theta) = \sum_{\omega,t} \log p\big(\phi(\omega,t), \alpha(\omega,t)|\Theta\big) \tag{3.3.1}$$

$$= \sum_{\omega,t} \log \sum_{i,\tau} \Big[p\big(\phi(\omega,t), \alpha(\omega,t)|z_{i\tau}(\omega,t), \Theta\big).p\big(z_{i\tau}(\omega,t)|\Theta\big)\Big] \tag{3.3.2}$$

$$= \sum_{\omega,t} \log \sum_{i,\tau} \Big[\Big(\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\Big).\mathcal{N}\Big(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)\Big).\psi_{i\tau}\Big] \tag{3.3.3}$$

This is the likelihood of a Gaussian mixture model, with one Gaussian per (i,τ) and $\psi_{i\tau}$ as the mixing coefficients. Using the EM algorithm, the parameters can be estimated by maximizing the objective function $Q(\Theta|\Theta_s)$ with respect to Θ.

$Q(\Theta|\Theta_s)$ is defined as

$$\sum_{\omega,t}\sum_{i,\tau} \Big[p\big(z_{i\tau}(\omega,t)|\phi(\omega,t), \alpha(\omega,t), \Theta_s\big). \log(z_{i\tau}(\omega,t), \phi(\omega,t), \alpha(\omega,t)|\Theta)\Big] \tag{3.3.4}$$

where Θ_s is the estimate of the parameters Θ after s iterations of the algorithm. Maximum likelihood estimation can be performed in two steps, the E step, in which the expectation of $z_{i\tau}(w,t)$ denoted by $\nu_{i\tau}(\omega,t)$, is determined given the observations and the parameter estimate Θ_s, followed by the M step in which the objective function is maximized with respect to Θ given $\nu_{i\tau}(\omega,t)$

$$E\big(z_{i\tau}(\omega,t)\big) = \nu_{i\tau}(\omega,t) \equiv p\big(z_{i\tau}(\omega,t)|\phi(\omega,t), \alpha(\omega,t), \Theta_s\big) \tag{3.3.5}$$

$$\propto p\big(z_{i\tau}(\omega,t)|\phi(\omega,t), \alpha(\omega,t)|\Theta_s\big) \tag{3.3.6}$$

$$= \psi_{i\tau}.\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\Big).\mathcal{N}\Big(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)\Big) \tag{3.3.7}$$

Since $z_{i\tau}(w,t)$ is a binary random variable, the conditional probability $p\big(z_{i\tau}(\omega,t)|\phi(\omega,t),\alpha(\omega,t),\Theta_s\big)$ is equal to its expectation. This expectation is used in the M step to estimate the parameters as follows

$$\mu_i(\omega) = \Big\langle \alpha(\omega,t) \Big\rangle_{t,\tau} \tag{3.3.8}$$

$$\eta_i^2(\omega) = \Big\langle \big(\alpha(\omega,t) - \mu_i(\omega)\big)^2 \Big\rangle_{t,\tau} \tag{3.3.9}$$

$$\xi_{i,\tau}(\omega) = \Big\langle \hat{\phi}(\omega,t;\tau) \Big\rangle_{t} \tag{3.3.10}$$

$$\sigma_{i\tau}^2(\omega) = \Big\langle \Big(\hat{\phi}(\omega,t;\tau) - \xi_{i,\tau}(\omega)\Big)^2 \Big\rangle_{t} \tag{3.3.11}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega,t} \nu_{i\tau}(\omega,t) \tag{3.3.12}$$

where the operator

$$\langle x \rangle \equiv \frac{\sum_{t,\tau} x \nu_{i\tau}(\omega,t)}{\sum_{t,\tau} \nu_{i\tau}(\omega,t)} \tag{3.3.13}$$

After convergence, the mask extracting each source i from the microphone signals $L(\omega,t)$ or $R(\omega,t)$ can be determined by summing the expectations of the latent indicators $\nu_{i\tau}(\omega,t)$ over the delay τ

$$M_i(\omega,t) \equiv \sum_{\tau} \nu_{i\tau}(\omega,t) \tag{3.3.14}$$

This mask represents the probabilities of each spectrogram point belonging to a specific source i. Preliminary experiments [61] have shown that converting these probabilities to more Wiener filter-like coefficients can enhance the separation performance.

MESSL achieves underdetermined speech separation in a reverberant environment and performs better than other comparable algorithms in terms of objective and subjective separation performance measures [25]. In addition, its probabilistic framework is flexible and allows the addition of other cues such as source models [62] and mixing vectors [57] and [58].

However, its performance degrades significantly for nearby sources as the

spatial cues become more similar particularly in the presence of reverberations. As explained in [71], the reverberation structure consists of early reflections and dense late reverberations. The early reflections generally affect speech positively by amplifying it. On the other hand, the late reverberation reflections are poorly correlated with the speech and act as additive noise (outliers). In reverberant conditions, the distribution of interaural cues is broadened which deteriorates the speech intelligibility and the human ability to use these cues for the separation of multiple speech sources [72].

In order to minimize the impact of reverberations, another source named as a "garbage" source is introduced in MESSL to account for all spectrogram points which are not fitted by other source models independent of their locations [25]. This added source allows a better estimation of the actual source parameters by avoiding poorly fitting points. Non-Gaussian mixture modelling, as explained in the following section, is an elegant approach that improves the robustness against outliers without the need for any reverberation detection and modelling algorithm to differentiate between unreliable reverberant points and direct-path points.

3.4 Student's t-distribution

GMMs are popular in the field of cluster analysis due to their tractability and their elliptical clustering ability relative to the spherical K-means type algorithms which cannot model correlation between variables of the feature space [30]. Unfortunately, they are extremely sensitive to outliers, the tails of the Gaussian distribution are often lighter than required and for a set of data containing observations with a distribution having more significant tails, the use of a Gaussian component might bias the fit of the mixture model [30]. Consequently, the estimates of the component means and variances would be affected by the presence of outliers which are typically not

normally distributed.

The problem of protection against outliers was tackled in [73] and [74]. These approaches assumed that the observed set of data is divided into two subsets, the good observations subset reflecting the actual population or cluster and the outlier subset. The goal is to separate them and use only the good data in the statistical analysis.

On the other hand, using the Student's t-distribution as an alternative to the Gaussian distribution achieves the robustness inherently without the need of an additional outlier detection algorithm, only by giving to the outliers a reduced weight in the estimation of the component parameters [30].

Adding a scaling random variable u, the Student's t-distribution for a vector $\mathbf{y}$ is interpreted as an infinite mixture of Gaussians having equal means but different precisions [67]

$$St(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}\big(\mathbf{y}|\boldsymbol{\mu}, u\boldsymbol{\Lambda}\big)\mathcal{G}\big(u|\nu/2, \nu/2\big)du \tag{3.4.1}$$

$$= \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{d}{2}}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\big[1 + \frac{1}{\nu}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})\big]^{-\frac{d+\nu}{2}} \tag{3.4.2}$$

The Gaussian and the Gamma distributions are given by

$$\mathcal{N}\big(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}\big) = (2\pi)^{-\frac{d}{2}}e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})} \tag{3.4.3}$$

$$\mathcal{G}\big(u|\kappa, \eta\big) = \frac{\eta^\kappa}{\Gamma(\kappa)}u^{\kappa-1}e^{-\eta u} \tag{3.4.4}$$

where $\boldsymbol{\mu}$ and Λ are the Gaussian mean vector and precision matrix, respectively. $\Gamma(.)$ is the Gamma function, κ and η are the Gamma distribution parameters, d is the dimension of the feature space and $(.)^T$ is the transpose operator. Parameter $\nu > 0$ is termed the degree of freedom and is considered as a robustness tuning parameter which can be fixed or adaptively estimated

from the data [30]. As shown in Figure 3.4, smaller values of ν lead to heavier tails while as ν tends to infinity the Student's t-distribution reduces to a Gaussian distribution.

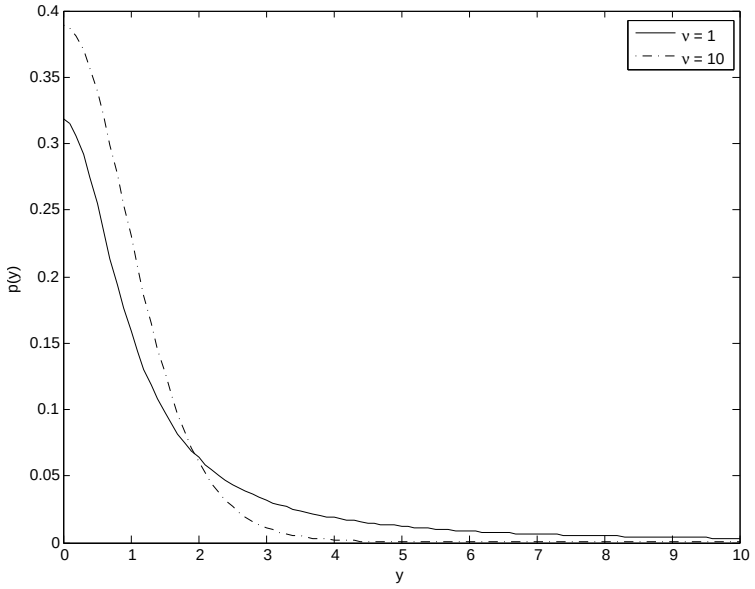


Figure 3.4: Student's t distribution for a univariate y. Changing the value of the degree of freedom parameter ν alters the pdf, smaller values of ν result in heavier tails.

Modelling using the mixtures of Student's t-distribution was proposed as an alternative to GMMs in image registration [75]. Image registration is an essential process in many applications including medical imaging, remote sensing and multisensor robot vision. The robustness to atypical pixel values was achieved even in cases of low signal to noise ratio (SNR) [75]. Motivated by their success in the field of image processing, SMMs are exploited for modelling binaural cues as a primary step in improving the statistical framework for underdetermined T-F based speech separation.

3.5 Student's t-distribution for IPD and ILD

In this section, IPD and ILD cues are fitted independently by the Student's t-distribution such that their joint distribution is given by

$$p(\phi(\omega,t),\alpha(\omega,t)|\Theta) = p(\hat{\phi}(\omega,t)|\Theta_p).p(\alpha(\omega,t)|\Theta_l) \tag{3.5.1}$$

where

$$\begin{aligned} &p(\hat{\phi}(\omega,t)|\Theta_p) = St(\hat{\phi}(\omega,t)|\Theta_p) \\ =&\frac{\Gamma\left(\frac{\nu_p+1}{2}\right)}{\Gamma\left(\frac{\nu_p}{2}\right)}\left(\frac{\lambda_p(\omega)}{\pi\nu_p}\right)^{\frac{1}{2}}\left(1+\frac{\lambda_p(\omega)(\hat{\phi}(\omega,t)-\xi(\omega))^2}{\nu_p}\right)^{-\frac{\nu_p+1}{2}} \end{aligned} \tag{3.5.2}$$

ILD can be modelled similarly to IPD by

$$\begin{aligned} &p(\alpha(\omega,t)|\Theta_l) = St(\alpha(\omega,t)|\Theta_l) \\ =&\frac{\Gamma\left(\frac{\nu_l+1}{2}\right)}{\Gamma\left(\frac{\nu_l}{2}\right)}\left(\frac{\lambda_l(w)}{\pi\nu_l}\right)^{\frac{1}{2}}\left(1+\frac{\lambda_l(\omega)(\alpha(\omega,t)-\eta(\omega))^2}{\nu_l}\right)^{-\frac{\nu_l+1}{2}} \end{aligned} \tag{3.5.3}$$

where $\Theta_p \equiv \{\xi(\omega),\lambda_p\}$ denotes the set of IPD parameters, $\Theta_l \equiv \{\mu(\omega),\lambda_l\}$ denotes the set of ILD parameters, ν_p and ν_l are fixed in advance. As $\nu_p \to \infty$ $St(\hat{\phi}(\omega,t)|\Theta_p)$ reduces to a Gaussian with mean $\xi(\omega)$ and precision $\lambda_p(\omega) = \frac{1}{\sigma^2(\omega)}$. Similarly as $\nu_l \to \infty$, $St(\alpha(\omega,t)|\Theta_l)$ reduces to a Gaussian with mean $\mu(\omega)$ and precision $\lambda_l(\omega) = \frac{1}{\eta^2(\omega)}$.

Let $\Theta \equiv \{\xi_{i\tau}(\omega),\lambda_{pi\tau}(\omega),\mu_i(\omega),\lambda_{li}(\omega),\psi_{i\tau}\}$ denote the collection of the parameters of the models and the degree of freedom is fixed in advance. The likelihood for a given observation can be expressed as

$$\mathcal{L}(\Theta) = \sum_{\omega,t}\log\sum_{i,\tau}[St(\hat{\phi}(\omega,t)|\theta_p).St(\alpha(\omega,t)|\Theta_l)\psi_{i\tau}] \tag{3.5.4}$$

Using the EM algorithm [75] the maximum likelihood estimation can be performed in two steps. In the E step, the expectation of the latent variable

$z_{i\tau}(\omega,t)$, the expectations of the phase scaling $u_{pi\tau}(\omega,t)$ and the level scaling $u_{li}(w,t)$ for each observation are computed given the current observations and the parameter estimates as follows

$$\begin{aligned} E\big(z_{i\tau}(\omega,t)\big) &= \kappa_{i\tau}(\omega,t) \\ &= \frac{\psi_{i\tau}.St(\hat{\phi}(\omega,t)|\Theta_p).St(\alpha(\omega,t)|\Theta_l)}{\sum_{i,\tau}\psi_{i\tau}.St(\hat{\phi}(\omega,t)|\Theta_p).St(\alpha(\omega,t)|\Theta_l)} \end{aligned} \tag{3.5.5}$$

$$u_{pi\tau}(\omega,t) = \frac{1+\nu_p}{\nu_p + (\hat{\phi}(\omega,t) - \xi i\tau(\omega))^2/\sigma^2_{i\tau}(\omega)} \tag{3.5.6}$$

$$u_{li}(\omega,t) = \frac{1+\nu_l}{\nu_l + (\alpha(\omega,t) - \mu_i(\omega))^2/\eta^2_i(\omega)} \tag{3.5.7}$$

These values are used to maximize the log-likelihood in the M step and re-estimate the means, variances and mixing weights below

$$\xi_{i\tau}(\omega) = \frac{\sum_t \kappa_{i\tau}(\omega,t).u_{i\tau}(\omega,t).\hat{\phi}(\omega,t;\tau)}{\sum_t \kappa_{i\tau}(\omega,t).u_{i\tau}(\omega,t)} \tag{3.5.8}$$

$$\lambda^{-1}_{pi\tau}(\omega) = \frac{\sum_t \kappa_{i\tau}(\omega,t).u_{i\tau}(\omega,t).\big(\hat{\phi}(\omega,t;\tau) - \xi_{i\tau}(\omega)\big)^2}{\sum_t \kappa_{i\tau}(w,t)} \tag{3.5.9}$$

$$\mu_i(\omega) = \frac{\sum_{t,\tau} \kappa_{i\tau}(\omega,t).u_{li}(\omega,t).\alpha(\omega,t)}{\sum_{t,\tau} \kappa_{i\tau}(\omega,t).u_{i\tau}(\omega,t)} \tag{3.5.10}$$

$$\lambda^{-1}_{li}(\omega) = \frac{\sum_{t,\tau} \kappa_{i\tau}(\omega,t).u_{li}(\omega,t).\big(\alpha(\omega,t) - \mu_i(\omega)\big)^2}{\sum_{t,\tau} \kappa_{i\tau}(\omega,t)} \tag{3.5.11}$$

$$\psi_{i\tau} = \frac{1}{\Omega T}\sum_{\omega,t} \kappa_{i\tau}(\omega,t) \tag{3.5.12}$$

Using the evaluated parameters, probabilistic masks are estimated by marginalizing over delay

$$M_i(\omega,t) \equiv \sum_{\tau} \kappa_{i\tau}(\omega,t) \tag{3.5.13}$$

The distributions of the spatial cues are corrupted by late reverberations which are not normally distributed [25]. The heavy tail behaviour of the SMM allows an accurate fitting of the mixtures of interaural cues by reducing the weight given to reverberations.

The main difference between the EM algorithm of the SMMs with respect to the GMMs is the estimation of the additional phase and level scaling parameters through Equations (3.5.6) and (3.5.7). This increase in the computation complexity leads to enhancing the robustness of clustering and hence more accurate probabilistic masks. Initialization of the proposed approach follows MESSL [25]. The allowed set of values for the discrete random τ is specified a priori. However, estimates of τ for each source are determined using PHAT [64]. PHAT as explained in Chapter 2 is a localization algorithm based on cross-correlation calculations and $\psi_{i\tau}$ is then assumed to have a Gaussian distribution with its mean located at each cross correlation maximum and a standard deviation of one sample. All the other parameters are left in a symmetric and non-informative state. The first E-step is calculated using these parameters followed by the M-step, these two steps are repeated until convergence. The EM algorithm for SMMs is summarized in Table 3.1

Table 3.1: EM algorithm for SMMs

1. Initialization: Using the estimates of τ from PHAT-histogram, initialize $\psi_{i\tau}$ while leaving the phase and level parameters in a symmetric and non-informative state.

2. E-step: Compute the expectation of the latent variable $\kappa_{i\tau}(\omega, t)$, the expectation of the phase scaling $u_{pi\tau}(\omega, t)$ and that of level scaling $u_{li\tau}(\omega, t)$ using the current parameter values.

3. M-step: Using $\kappa_{i\tau}(\omega, t)$, $u_{pi\tau}(\omega, t)$ and $u_{li\tau}(\omega, t)$ to re-estimate the phase parameter set $\{\xi_{i\tau}(\omega), \lambda^{-1}_{pi\tau}(\omega)\}$, the level parameter set $\{\mu_i(\omega), \lambda^{-1}_{pi\tau}(\omega)\}$ and the mixing coefficients $\psi_{i\tau}$.

4. Cycling between E and M until convergence or for a fixed number of iterations.

3.6 Experimental results

Three main experiments were performed in order to evaluate the performance of MESSL and compare between the proposed non-Gaussian modelling of interaural cues and the conventional modelling employed in MESSL. The first experiment examines various complexities of MESSL and their separation performance under different reverberation times. The second experiment focuses on the complexity based only on the IPD cues and the improvement in the speech separation when their GMMs are replaced by SMMs. The third experiment exploits the effectiveness when both cues are modelled using SMMs.

3.6.1 Data sources

Speech utterances of 2.5 s long were chosen randomly from the whole TIMIT acoustic-phonetic continuous speech corpus [76]. Mixtures were formed from different combinations of male and female signals sampled at 16 KHz. These signals were normalized to have the same energy. All experiments included

two speakers, one target and one interferer.

3.6.2 Room Impulse responses

In the first experiment, the speech utterances were convolved with room impulse responses (RIRs) generated using the image method [77]. In the other experiments, binaural real impulse responses (BRIRs) were used [78]. The BRIRs were measured in a real classroom of dimensions 5 m×9 m×3.5 m and a reverberation time (RT60) of 565 ms. The target was always positioned facing the microphones and the interferer was located at various azimuthal angles as shown in Fig. 3.5.

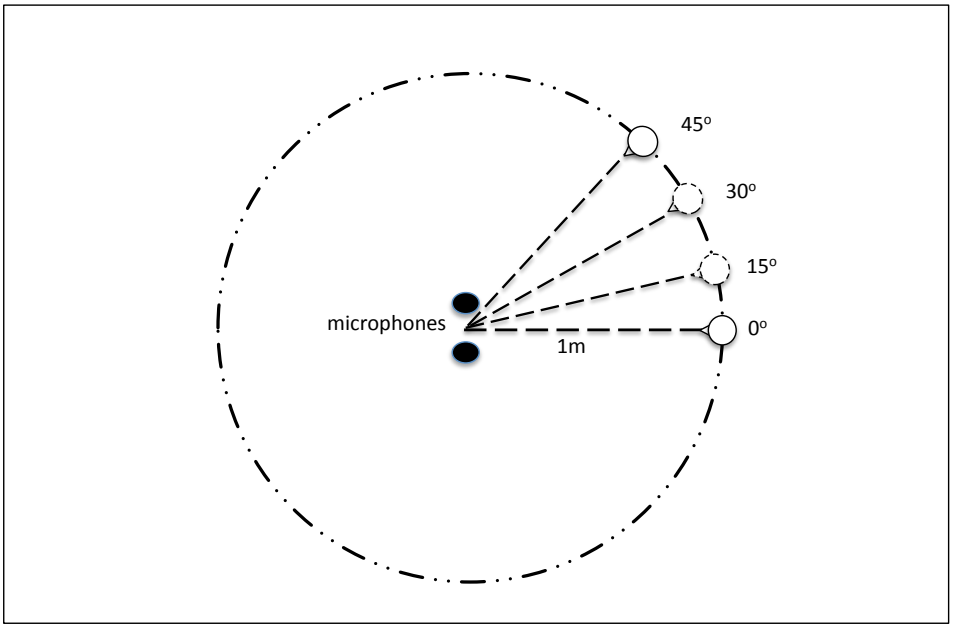


Figure 3.5: The room layout showing approximate positions of the sources and the microphones.

3.6.3 Separation performance evaluation

The performance is evaluated using the signal-to-distortion ratio (SDR) [48]. This metric is the ratio of the energy in the original signal to the energy of interfering signals or other unexplained artifacts. Any energy in the estimated signal that can be explained with a linear combination of delayed versions of the target signal (up to 32ms) counts towards the target energy. Similarly, any energy that can be explained as a linear combination of delayed versions of the interfering signals counts as interference. Any energy that does not belong to any of these projections is considered artifacts such as reverberation from any of the sources.

To describe this mathematically, let $s_i(t)$ denote the set of original anechoic signals, $s_j(t)$ the target signal and $\hat{s}_j(t)$ the estimated target signal. The projection operator $P(x, \{y_i\}, \tau_{max})$ is defined as the projection of the signal x onto versions of the signals $\{y_i\}$ shifted by every integer number of samples up to τ_{max} [61]. The three signals used for the objective evaluation of the speech separation are

$$s_{target}(t) = P(\hat{s}_j, s_j, \tau_{max}) \tag{3.6.1}$$

$$e_{interf}(t) = P(\hat{s}_j, \{s_i\}, \tau_{max}) - P(\hat{s}_j, s_j, \tau_{max}) \tag{3.6.2}$$

$$e_{artif}(t) = \hat{s}_j - P(\hat{s}_j, \{s_i\}, \tau_{max}) \tag{3.6.3}$$

and the three metrics SDR, SIR and SAR are defined as follows

$$\text{SDR} = 10 \log_{10} \frac{||s_{target}||^2}{||e_{interf} + e_{artif}||^2} \tag{3.6.4}$$

$$\text{SIR} = 10 \log_{10} \frac{||s_{target}||^2}{||e_{interf}||^2} \tag{3.6.5}$$

$$\text{SAR} = 10 \log_{10} \frac{||s_{target} + e_{interf}||^2}{||e_{artif}||^2} \tag{3.6.6}$$

where $||.||^2$ indicates the squared vector Euclidean-norm.

3.6.4 MESSL versions

In the first experiment, different MESSL versions are run to separate two sources; the target and the interferer are positioned at 0° and 75° respectively and the reverberation time (RT60) is equal to 300ms.
Different versions of MESSL correspond to various model complexities and the parameters sets are named by their complexity in modelling the interaural cues, 0 refers to the simplest modelling, 1 indicates a more complex modelling and Ω refers to the most complex frequency dependent modelling. For instance, $\Theta_{\Omega 0}$ corresponds to a complex ILD and a simple IPD models, respectively [25].

Table 3.1: SDR for different model complexities, separating 2 speakers in reverberation, RT60=300ms

Name	ILD mean	ILD std	IPD mean	IPD	SDR (dB)
Θ_{00}	0	∞	0	σ_i	4.93
Θ_{11}	μ_i	η_i	$\xi_{i\tau}$	$\sigma_{i\tau}$	6.92
$\Theta_{\Omega\Omega}$	$\mu_i(\omega)$	$\eta_i(\omega)$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$	9.26

Table 3.1 shows the SDR of three model complexities, Θ_{00}, Θ_{11} and $\Theta_{\Omega\Omega}$. Θ_{00} has no ILD contribution and an IPD model with zero mean and standard deviation that varies only by source. Θ_{11} has a frequency-independent mean and a standard deviation varying by source and τ and finally $\Theta_{\Omega\Omega}$ has the full frequency-dependent ILD and IPD model parameters. It can be seen that increasing models complexity improves the SDR; using $\Theta_{\Omega\Omega}$ increases the SDR by 4.3 dB compared to the simplest model Θ_{00}.

Figure 3.6 depicts the performance of the previous three MESSL models

when the interferer is located at 75° while varying the RT60. It can be seen that the performance degrades as the reverberation time increases.

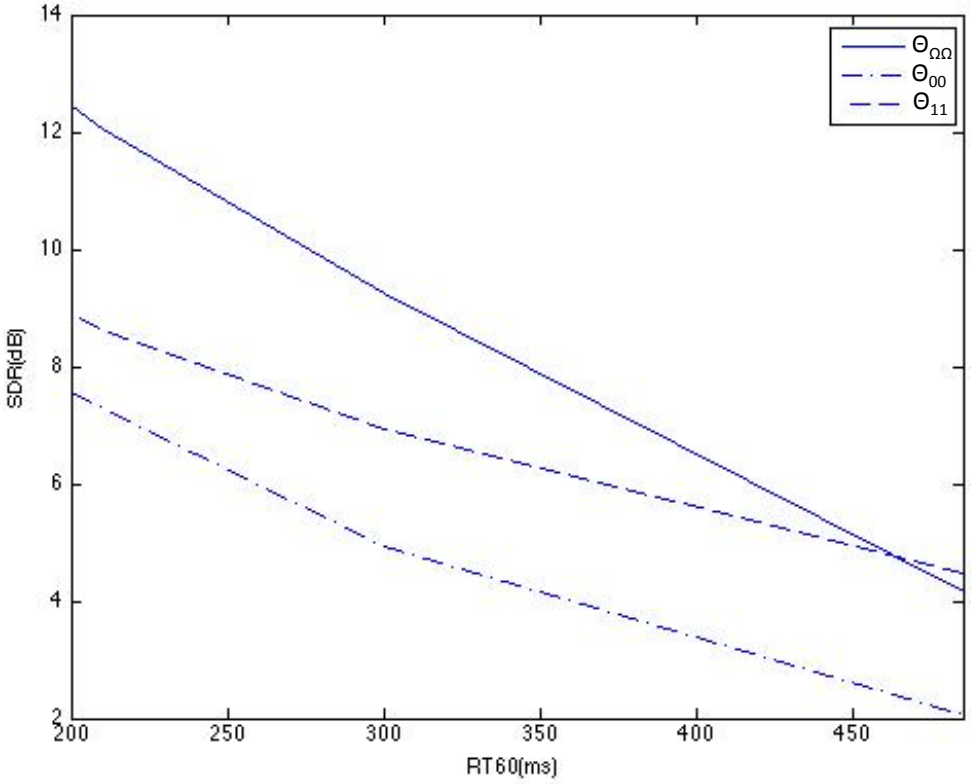


Figure 3.6: SDR of MESSL models at different RT60s.

Finally, $\Theta_{\Omega\Omega}$ was run for different positions of the interferer and Figure 3.7 shows the SDR corresponding to six azimuths between 15° and 90° i.e. [15°, 30°, 45°, 60°, 75°, 90°]. MESSL performs worse as the separation decreases with an average difference of 2.5 dB between the SDRs at 90° and 15°.

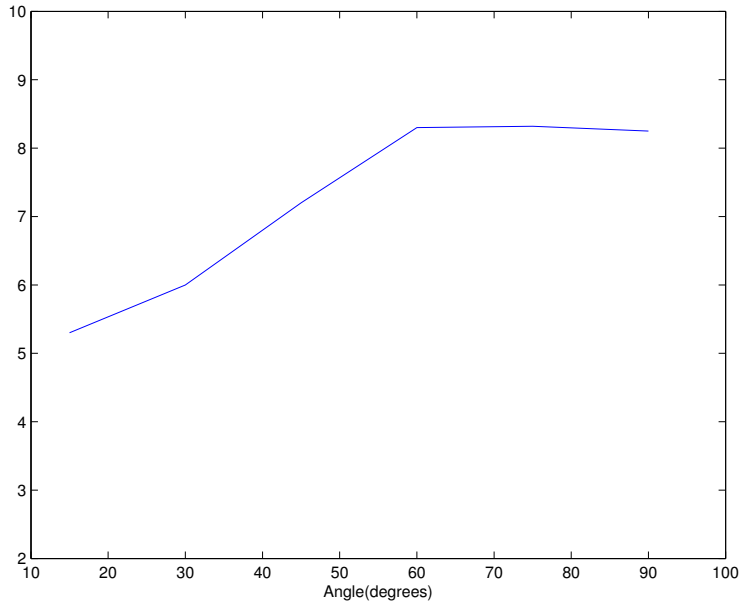


Figure 3.7: SDR of MESSL $\Theta_{\Omega\Omega}$ at different separation angles and RT60=300ms.

3.6.5 SMMs for IPD cues

In this experiment, SMMs are used only to model IPD cues. The MESSL version relying on IPDs and the proposed approach were compared for different values of the degree of freedom ν_p. Ten different mixtures were formed in total from the TIMIT database to test the separation performance particularly at small azimuths where MESSL performance degrades significantly. Two model complexities denoted by Θ_1 and Θ_Ω were tested. In Θ_1, the mean and variance are frequency independent whereas Θ_Ω corresponds to frequency dependent mean and variance. The average results shown in Table 3.2 and Table 3.3, confirm the advantage of the Student's t-distribution.

The average SDR improvement varies with the degree of freedom as well as the azimuthal separation. For Θ_1, $\nu_p = 0.1$ and small separations, an average improvement of 1 dB was obtained while the best individual improvement was 2.3 dB. For Θ_Ω, the average and the best individual improvement

Table 3.2: Separation performance comparison in SDR (dB) for Θ_1

Azimuth angles	$\nu_p = 0.1$	$\nu_p = 1$	$\nu_p = 10$	MESSL
15°	2.96	2.69	1.8	1.61
30°	2.93	2.98	2.43	2.32
45°	3.88	3.99	3.69	3.63

Table 3.3: Separation performance comparison in SDR (dB) for Θ_Ω

Azimuth angles	$\nu_p = 0.1$	$\nu_p = 1$	$\nu_p = 10$	MESSL
15°	3.15	2.9	2.29	2.16
30°	3	3.01	2.57	2.44
45°	3.94	3.99	3.85	3.8

were 0.8 dB and 1.5 dB respectively. For both complexities, the improvement decreases as ν_p increases since the student's t-distribution approaches the Gaussian distribution used in MESSL.

3.6.6 SMMs for both IPDs and ILDs

In this experiment, both interaural cues are modelled using SMMs. Two complexities Θ_{11} and $\Theta_{\Omega\Omega}$ were chosen for comparison. Θ_{11} has a frequency-

Table 3.4: SDR (dB) MESSL $\Theta_{\Omega\Omega}$

Azimuth angles	15°	30°	45°
mix1	-1.51	1.42	3.49
mix2	2.83	1.46	3.17
mix3	0.02	5.80	5.92
mix4	2.27	2.67	4.49
mix5	2.25	1.13	3.45

independent mean and a standard deviation varying by source and τ, whereas $\Theta_{\Omega\Omega}$ has full frequency-dependent ILD and IPD model parameters. Assuming by symmetry that $\nu_p = \nu_l = \nu$, individual SDRs obtained for five mixtures using MESSL and our proposed approach with $\nu = 1$ and $\nu = 10$ are shown in Table 3.4, Table 3.5 and Table 3.6 respectively.

It can be seen that our approach outperforms MESSL for small separation angles. The average results over ten different mixtures shown in Table 3.7 and Table 3.8 confirm the advantage of the SMMs over GMMs.

Table 3.5: SDR (dB) proposed approach $\Theta_{\Omega\Omega}$ $\nu = 1$

Azimuth angles	15°	30°	45°
mix1	0.57	2.14	4
mix2	3.15	2.91	3.56
mix3	4.81	6.38	6.62
mix4	4.13	4.06	5.1
mix5	3.49	3.45	4.17

Table 3.6: SDR (dB) proposed approach $\Theta_{\Omega\Omega}$ $\nu = 10$

Azimuth angles	15°	30°	45°
mix1	-0.22	2.11	3.86
mix2	3.5	2.23	3.5
mix3	2.09	5.93	6.62
mix4	3.16	3.16	4.8
mix5	2.82	2.39	3.76

The average SDR improvement varies with the degree of freedom as well as the azimuthal separation. For $\Theta_{\Omega\Omega}$, the best average improvements were obtained at $\nu = 1$ and they are equal to 1.8 dB, 1.5 dB and 0.7 dB for the azimuthal angles of 15°, 30° and 45° respectively. For Θ_{11}, the average improvements decrease to 1.2 dB, 0.6 dB and 0.3 dB. For both complexities, the improvement decreases as ν increases.

Table 3.7: Separation performance comparison in SDR (dB) for $\Theta_{\Omega\Omega}$

Azimuth angles	$\nu = 0.1$	$\nu = 1$	$\nu = 10$	MESSL
15°	3.03	3.19	2.38	1.39
30°	3.35	3.45	2.69	1.97
45°	4.25	4.36	4.1	3.67
average	3.54	3.67	3.06	2.34

Table 3.8: Separation performance comparison in SDR (dB) for Θ_{11}

Azimuth angles	$\nu = 0.1$	$\nu = 1$	$\nu = 10$	MESSL
15°	2.29	2.26	1.63	1.09
30°	2.47	2.78	2.48	2.24
45°	3.9	4.05	3.84	3.79
average	2.89	3.03	2.65	2.37

Using mixtures of Student's t-distribution for modelling interaul cues has proven to improve the robustness against atypical values and as a by product has improved the separation performance. This distributing was firstly used to model IPDs then was applied to independently model both interaural cues. For $\nu = 1$, an average SDR improvement equivalent to 28% (over three azimuths) was obtained for the frequency independent complexity Θ_{11}, this improvement reached 57% for the frequency dependent complexity $\Theta_{\Omega\Omega}$.

3.7 Summary

In this chapter, non-Gaussian modelling of the spatial interaural cues commonly used in the T-F based speech separation is proposed. The Student's t-distribution whose heavy tails values better reflect outliers, provides a generalization of the Gaussian distribution. This approach is important to improve the robustness of the CASA algorithms based on clustering spectrogram points in noisy and reverberant environments without the need for any reverberation detection method to avoid poorly fitted spectrogram points. The EM algorithm was used to estimate the parameters of the SMMs. These models have proven to be more important when the sources are in close proximity, where accurate representation of the tail behaviour appears to lead to improved separation. Experimental results comparing this proposed approach to the state-of-the-art MESSL confirmed a significant average improvement of the separation performance particularly for the frequency-dependent versions. In the next chapter, variational Bayesian inference is exploited as an alternative to the maximum likelihood estimation to avoid the drawbacks of the EM algorithm and further improve the clustering framework.

Chapter 4

VARIATIONAL EM FOR CLUSTERING INTERAURAL PHASE CUES IN MESSL FOR UNDERDETERMINED SPEECH SEPARATION

Using only two-channel stereo mixtures, MESSL clusters spectrogram points based on their interaural spatial cues. GMMs are assumed for the interaural cues and their corresponding parameters are determined by maximum likelihood estimation via the EM framework. However, the presence of singularities and over-fitting are major drawbacks of MLE. In this chapter, an alternative clustering framework is proposed based on variational Bayesian (VB) inference. This approach overcomes the difficulties associated with the likelihood optimization and improves the separation especially when the sources are in close proximity. The proposed framework is applied for clustering IPD cues fitted using GMMs. Experimental results based on speech mixtures formed from the TIMIT database and convolved with BRIRs confirm the advantage of the VB approach in terms of SDR.

4.1 INTRODUCTION

Bayes' theory was formulated by Thomas Bayes in his work "Essay towards solving a problem in the doctrine of chances" in the 18^{th} century [79]. His work focused on an important problem referred to at that time as the problem of inverse probability, which had emerged with the introduction of insurance as a new concept [67]. The term "inverse" as explained in [80] is concerned with the backwards inference from the data to the parameters. Bayes' framework was rediscovered and generalized by Pierre-Simon Laplace in 1774 [81]; it is still widely applicable for the understanding and describing of probabilistic models in the world of mathematics in general and in the fields of pattern recognition and machine learning in specific. The applications of Bayesian methods have grown significantly in the last decades with the development of algorithms such as variational Bayes and expectation propagation along with the improvements in the computational power of modern computers [67].

For a set of data $\mathbf{Y} = \{y_1, ..., y_N\}$, Bayes' theorem states that

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{p(\mathbf{Y})} \tag{4.1.1}$$

where θ is the model parameter with a prior distribution $p(\theta)$ representing the assumption made about θ before observing the data. The uncertainty in θ can be determined after the observation of the data set in terms of the posterior distribution $p(\theta|\mathbf{Y})$. On the other hand, $p(\mathbf{Y}|\theta)$ represents the likelihood function which describes the probable variation of the data as a function of the parameter θ. It is not a probability distribution and its integral over θ is not equal to one; $p(\mathbf{Y}) = \int p(\mathbf{Y}|\theta)p(\theta)d\theta$ is the normalization constant to ensure that $p(\theta|\mathbf{Y})$ is a proper density function that integrates to one.

Hence, Bayes' theorem can be written as

$$posterior \propto likelihood \times prior \tag{4.1.2}$$

The likelihood function plays a central role in the estimation of the soft masks in the underdetermined T-F speech separation based on the clustering of binaural cues as demonstrated in the previous chapter. The way this function is viewed from the Bayesian approach differs totally from the traditional frequentist perspective on which conventional EM framework is based. In the latter, θ is considered as a fixed parameter which can be determined through an 'estimator' which is commonly the maximization of the likelihood function [67]. Considering the distribution of possible sets of data, the parameter θ is given the value that maximizes $p(\mathbf{Y}|\theta)$.

On the contrary, the Bayesian approach considers only the observed data set and assumes that the parameter is varying with a prior distribution $p(\theta)$. This prior distribution is one of the great advantages of the Bayesian estimation. For instance, in the simple experiment of a tossing a coin three times and getting head each time, maximum likelihood estimation would conclude that the probability of head is one, which implies that any future repetition of the experiment will result in one. By including a prior, the Bayesian approach avoids such an extreme conclusion [67].

For an incomplete data set, the problem of parameter estimation is even more complex. Considering the mixture of Gaussians model, the log likelihood function for a data set Y is given by

$$\ln p(\mathbf{Y}|\psi, \mu, \sigma) = \sum_{n=1}^{N} \ln \Big(\sum_{k=1}^{K} \psi_k \mathcal{N}(y_n|\mu_k, \sigma_k^2 \Big) \tag{4.1.3}$$

where $Y = \{y_1, \ldots, y_N\}$ is a set of independent and identically distributed random variables y_n, $\psi = \{\psi_k\}$ is the set of mixing coefficients μ_k and σ_k^2 denote the component mean and variance, respectively. Maximizing the log

likelihood, as explained in detail in Appendix A.1, is not straightforward due to the summation over K components. Setting the derivatives of the log likelihood to zero will not result in a closed form solution. The EM algorithm provides a powerful iterative solution and has been successfully applied as a clustering framework for classical missing data problems [28]. Unfortunately, the EM framework applied to GMMs suffers from the probable presence of singularities occurring whenever a data point coincides with one of the Gaussian components. These singularities do not occur in the case of homogeneous data modelled by a single Gaussian distribution, since if one data point falls on the Gaussian component, its contribution to the likelihood is multiplicative and the overall likelihood will go to zero rather than infinity. On the other hand, for a GMM with at least two components, the component with the finite variance would assign finite values to all data points, while the other component coinciding on one of the points would contribute to an unbounded added value for the total likelihood as shown in Figure 4.1. A Bayesian approach would avoid these singularities [67].

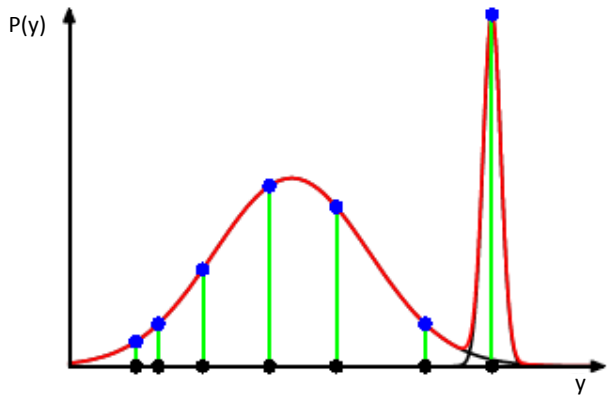


Figure 4.1: Illustration of the probable unbounded property of maximum likelihood estimation for GMM [67].

However, the full Bayesian solution for real applications including image processing, analysis of biomedical signals and source separation, is not analytically tractable and should be approximated [67].

Stochastic approximation such as Markov Chain Monte Carlo (MCMC) is widely used in the field of statistics and digital signal processing [82]. MCMC is non parametric and asymptotically exact. In other words, similar to other sampling methods, the generation of exact results is inefficient in terms of processing time, which limits their applicability to small-scale problems [67]. Gibbs sampling, a simple and standard MCMC method, was applied in the field of underdetermined blind separation of audio signals to separate linear instantaneous mixtures of sources [83]. The separation in the proposed approach is based on the independence of the sources (ICA based approach) combined with the sparsity property that the sound sources exhibit in the frequency domain. The sparsity assumption means that the decomposition of any sound source on a given basis, such as the discrete cosine transform basis, would result in few non zero coefficients. These coefficients were modelled using a Student's t-distribution. Although, this approach showed better results compared the traditional EM framework in terms of the separation quality and robustness to mixing conditions, the MCMC processing time was significantly higher than that of EM. It required approximately 3 hours to separate 3 sources using MCMC compared to a few minutes using the EM framework [83].

An alternative to the MCMC method is the variational approach which falls in the deterministic approximation category [84] [85]. VB is based on mean field theory which originated in the field of statistical physics [86]. Due to its computational efficiency compared to MCMC, the variational approach has gained increasing popularity in the fields of machine learning and pattern recognition [67].

In the ICA based T-F masking approach, the variational framework was proposed as an alternative to the EM algorithm to avoid overfitting and initialization sensitivity as it was shown that depending on initial values the EM update rules might converge to a local maximum resulting only in suboptimal solutions [87]. In [88], a VB implementation for the underdetermined convolutive source separation via frequency bin-wise clustering and a permutation alignment approach developed in [53] is proposed. The VB framework used to cluster a mixture of Gaussians at each frequency bin achieved similar separation quality compared to the original EM based approach and in addition required no knowledge of the initial number of sources. A Bayesian framework was also proposed in [89] to deal with the localization and separation with permutation resolution in a unified framework.

Motivated by the advantages of the Bayesian methods compared to the likelihood optimization and the efficiency of the variational approach as a Bayesian approximation methodology and its successful application in the field of ICA based T-F masking, a VB framework is proposed in this chapter as an alternative to the EM clustering framework employed in MESSL for underdetermined speech separation. Section 4.2 introduces the variational Bayesian inference main concept and assumptions. In Section 4.3, the proposed framework is applied for clustering IPDs in MESSL. In Section 4.4, the Variational EM update rules are thoroughly explained. Experimental results are undertaken in Section 4.5 and the chapter is summarized in Section 4.6.

4.2 Variational inference

Variational inference originated in the 18^{th} century with the work of Euler and Lagrange and others on the calculus of variations introducing the concept of a functional derivative [67]. A good example of a functional is the entropy H[p] which is a mapping that takes the probability distribution as the input and returns an output expressed as

$$H[p] = \int p(y) \ln p(y) dy \tag{4.2.1}$$

Variational formulations can be interpreted as optimization problems where the quantity to be maximized/minimized is a functional and the solutions can be determined by considering all possible functions optimizing this functional. Variational methods approximate the solutions by performing the optimization over a restricted range of functions [90].

Considering the general variational optimization for an inference problem involving a set of N independent and identically distributed data points $\mathbf{Y} = \{y_1, ..y_N\}$ and assuming a Bayesian model, all parameters are given prior distributions and the set of all hidden variables as well as the parameters are combined in one set $\mathbf{Z} = \{z_1, ..., z_N\}$. While the joint distribution $p(\mathbf{Y}, \mathbf{Z})$ is specified, the goal is to approximate the posterior distribution $p(\mathbf{Z}|\mathbf{Y})$ as well as the model evidence $p(\mathbf{Y})$. The log of the marginal probability can be decomposed as follows [67]

$$\ln p(\mathbf{Y}) = \mathcal{L}(q) + KL(q||p) \tag{4.2.2}$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \frac{p(\mathbf{Y}, \mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \tag{4.2.3}$$

$$KL(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{Y})}{q(\mathbf{Z})} d\mathbf{Z} \tag{4.2.4}$$

$KL(q||p)$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{Y})$ which is greater or equal to zero, with the equality if and only if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{Y})$. Hence, $\mathcal{L}(q) \leq \ln p(\mathbf{Y})$, i.e. $\mathcal{L}(q)$ is considered as a lower bound on the log likelihood as shown in Figure 4.2.

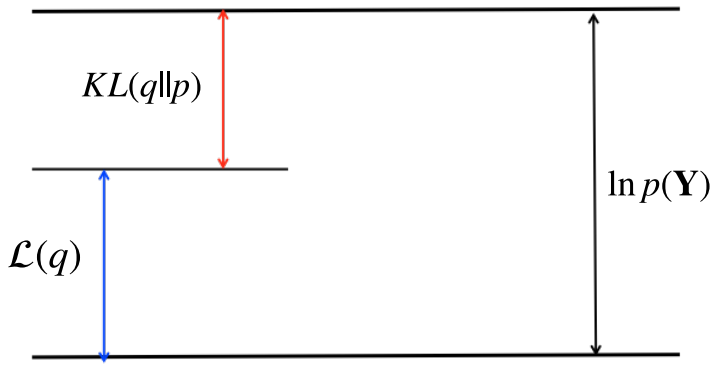


Figure 4.2: Illustration of the decomposition of the log marginal probability $\ln p(\mathbf{Y})$. Since $KL(q||p) \geq 0$, $\mathcal{L}(q)$ is a lower bound on $\ln p(\mathbf{Y})$.

Maximizing the lower bound $\mathcal{L}(q)$ with respect to $q(\mathbf{Z})$ is equivalent to minimizing the KL divergence, the maximum of the lower bound will occur when $q(\mathbf{Z})$ is equal to the posterior distribution $p(\mathbf{Z}|\mathbf{Y})$ [67]. The variational method approximates the Bayesian inference by minimizing the KL divergence over a family of distributions $q(\mathbf{Z})$ satisfying a factorized form as explained in the following section.

4.2.1 Factorized distributions

The hidden variables $\mathbf{Z}$ are assumed to be partitioned into M disjoint groups $\mathbf{Z}_i$, with $i = 1, ..., M$ such that

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i) \tag{4.2.5}$$

This factorization corresponds to the approximation method developed in statistical physics called the mean field theory [86]. This is the only approximation imposed in the variational inference framework as $q_i(\mathbf{Z}_i)$ can have any functional form. For notational simplicity $q_i(\mathbf{Z}_i)$ is denoted by q_i, substituting Equation (4.2.5) into Equation (4.2.3), the lower bound can be expressed as [67]

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \Big\{ \ln p(\mathbf{Y}, \mathbf{Z}) - \sum_i \ln q_i \Big\} d\mathbf{Z} \\
&= \int q_j \Big\{ \int \ln p(\mathbf{Y}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \Big\} d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + const \\
&= \int q_j \ln \tilde{p}(\mathbf{Y}, \mathbf{Z}_j) d\mathbf{Z}_j - \int q_j \ln q_j d\mathbf{Z}_j + const
\end{aligned} \tag{4.2.6}$$

where $\ln \tilde{p}(\mathbf{Y}, \mathbf{Z}_j)$ defines the expectation with the respect to the q distributions over all variables $\mathbf{Z}_i$ for $i \neq j$

$$\ln \tilde{p}(\mathbf{Y}, \mathbf{Z}_j) = E_{i \neq j}\big[\ln p(\mathbf{Y}, \mathbf{Z}) \big] + const \tag{4.2.7}$$

$$E_{i \neq j}\big[\ln p(\mathbf{Y}, \mathbf{Z}) \big] = \int \ln p(\mathbf{Y}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i \tag{4.2.8}$$

Equation (4.2.6) is a negative KL between $q_j(\mathbf{Z}_j)$ and $\tilde{p}(\mathbf{Y}, \mathbf{Z}_j)$, its minimum occurs when $q_j(\mathbf{Z}_j) = \tilde{p}(\mathbf{Y}, \mathbf{Z}_j)$.

Therefore, the optimal solution $q_j^\star(\mathbf{Z}_j)$ is expressed as

$$\ln q_j^*(\mathbf{Z}_j) = E_{i \neq j}\big[\ln p(\mathbf{Y}, \mathbf{Z}) \big] + const \tag{4.2.9}$$

and the additive constant in (4.2.6), (4.2.7) and (4.2.9) is obtained by normalizing $q_j^*(\mathbf{Z}_j)$.

In other words, the log of the optimal distribution $q_j^*(\mathbf{Z}_j)$ is equivalent to the expectation of the joint distribution over the observed and hidden variables with respect to all other factors $q_i(\mathbf{Z}_i)$ for $i \neq j$. Taking the exponential of both sides and normalizing gives

$$q_j^\star(\mathbf{Z}_j) = \frac{exp\Big(E_{i\neq j}\big[\ln p(\mathbf{Y},\mathbf{Z})\big]\Big)}{\int exp\Big(E_{i\neq j}\big[\ln p(\mathbf{Y},\mathbf{Z})\big]\Big)d\mathbf{Z}_j} \tag{4.2.10}$$

The set of equations for $j = 1,..,M$ given by (4.2.9) do no represent an explicit solution since the optimum factor $q_j^\star$ depends on the expectations determined with respect to the factors q_i for $i \neq j$. A solution can be found through cycling between two steps similar to the EM algorithm. The factors q_i are initialized, then the algorithm cycles through the factors replacing each in turn with a new estimate determined by the equation (4.2.9) using the current estimates for all the other factors [67], [29]. The VB framework explained in this section is used to replace the EM framework for clustering IPDs in MESSL as explained in the next section.

4.3 VB for GMM in MESSL

Considering the version where only IPD cues are used for clustering the spectrogram points, MESSL relies essentially on maximizing the log likelihood function given by

$$\mathcal{L}(\Theta) = \sum_{\omega,t} \log \sum_{i,\tau} [\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega).\psi_{i\tau}] \tag{4.3.1}$$

where $\hat{\phi}(\omega,t;\tau)$ is the phase residual modelled by a Gaussian distribution [70] with frequency-dependent mean $\xi_{i\tau}(\omega)$ and precision $\lambda_{i\tau}(\omega) = \frac{1}{\sigma_{i\tau}^2(\omega)}$. The major problem with the EM algorithm as explained previously is the

potential unbounded property of the likelihood [31]. If one component has its mean exactly equal to one of the data points, its contribution to the likelihood function can be written as

$$\mathcal{N}(\hat{\phi}(\omega,t)|\xi_{i\tau}(\omega),\sigma^2_{i\tau}(\omega)) = \frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma_{i\tau}(\omega)} \tag{4.3.2}$$

as $\sigma_{i\tau}(\omega)$ tends to 0, the likelihood function tends to infinity. These singularities will always occur whenever one of the Gaussian components collapses onto a data point. Detection of such singularities and avoiding them is crucial when adopting MLE [67]. This difficulty does not occur if a VB approach is employed since the component parameters are not fixed but also considered as random variables with prior distributions. The probabilistic model of all the latent variables in the proposed approach is described below.

The latent variable model

For each observation $\hat{\phi}(\omega,t;\tau)$, there is a corresponding binary vector $\mathbf{z}(\omega,t)$ comprising the elements $z_{i\tau}(\omega,t)$; only one element is equal to unity with probability $\psi_{i\tau}(\omega)$, which represents the probability of belonging to a source i and delay τ such that $\sum_{i\tau} z_{i\tau}(\omega,t) = 1$. The number of the latent variables $\mathbf{z}(\omega,t)$ increases with the size of the data set. However, the size of the parameter set denoted by $\Theta = \{\xi_{i\tau}(\omega), \lambda_{i\tau}(\omega), \psi_{i\tau}(\omega)\}$ is fixed independent of the data size. Figure 4.3 depicts the graphical representation of the Bayesian GMM at each T-F point.

The Gaussian mixture distribution for the phase residual in MESSL can be written as [69]

$$p(\hat{\phi}(\omega,t;\tau) = \sum_{i,\tau}\psi_{i\tau}(\omega)\mathcal{N}\big(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega),\lambda^{-1}_{i\tau}(\omega)\big) \tag{4.3.3}$$

The distribution of the latent vector given the mixing coefficients can be expressed, as explained in Chapter 3, in terms of a multinomial distribution

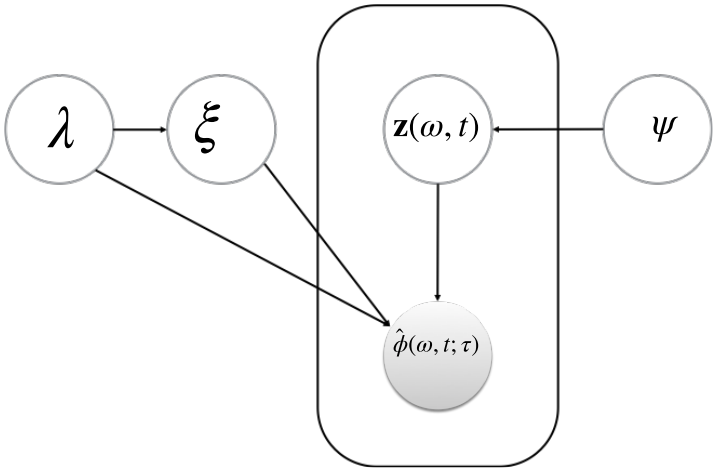


Figure 4.3: Directed graph of the Bayesian GMM at each spectrogram point. The shaded node represents the observed vector $\mathbf{y}(\omega, t; \tau)$. The arrow direction indicates dependencies between random variables. The component means $\zeta_{i\tau}(\omega)$ depend on the precision $\lambda_{i\tau}(\omega)$.

as follows

$$p(\mathbf{z}(\omega, t)|\boldsymbol{\psi}) = \prod_{i,\tau} \psi_{i\tau}(\omega)^{z_{i\tau}(\omega,t)} \tag{4.3.4}$$

where $\boldsymbol{\psi} = \{\psi_{i\tau}(\omega)\}$.

Since the conditional distribution of the residual phase given a specific value of $\mathbf{z}(\omega, t)$ is a single Gaussian [69], the conditional probability of the observed data $\hat{\phi}(\omega, t; \tau)$ given the latent variables and the component parameters is therefore expressed as

$$p(\hat{\phi}(\omega, t; \tau)|\mathbf{z}(\omega, t), \Theta) = \prod_{i,\tau} \mathcal{N}(\hat{\phi}(\omega, t; \tau)|\xi_{i\tau}(\omega), \lambda_{i\tau}^{-1}(\omega))^{z_{i\tau}(\omega,t)} \tag{4.3.5}$$

Parameter priors

To complete the Bayesian framework, the priors over the parameters are introduced. At each frequency ω, conjugate prior distributions are always

considered so that the posterior distributions have the same functional forms as their priors [67]. The concept of conjugate priors is briefly explained, followed by the priors used for modelling the mixing coefficients and that defining the component parameters.

Exponential family and conjugate priors

Given a continuous variable $\mathbf{x}$ with probability distribution $p(\mathbf{x}|\Theta_n)$ belonging to an exponential family having the form [91]

$$p(\mathbf{x}|\Theta_n) = h(\mathbf{x})f(\Theta_n)exp\{\Theta_n^T v(\mathbf{x})\} \tag{4.3.6}$$

where Θ_n is termed the set of natural parameters of the distribution, $v(\mathbf{x})$ is some function of $\mathbf{x}$ and $f(\Theta_n)$ is a normalization function to ensure that

$$f(\Theta_n)\int h(\mathbf{x})exp\{\Theta_n^T v(\mathbf{x})\}d\mathbf{x} = 1 \tag{4.3.7}$$

The likelihood function for a set of independent identically distributed data $X = \{x_1, ..x_n\}$ is expressed as

$$p(\mathbf{X}|\Theta_n) = \Big(\prod_N h(\mathbf{x}_n)\Big)f(\Theta_n)^N exp\Big\{\Theta_n^T \sum_N v(\mathbf{x}_n)\Big\} \tag{4.3.8}$$

For any member of the exponential family there exists a prior $p(\Theta_n)$ that is conjugate to the likelihood function such that the posterior distribution has the form of the prior [91]. This concept is explained next through the multinomial distribution and its corresponding conjugate prior.

The Dirichlet distribution

The multinomial distribution $p(\mathbf{z}(\omega,t)|\boldsymbol{\psi})$ is a member of the exponential family as it can be written in the form [67]

$$p(\mathbf{z}(\omega,t)|\boldsymbol{\psi}) = exp\Big\{\sum_{i\tau} z_{i\tau}(\omega,t)\ln\psi_{i\tau}(\omega)\Big\} \tag{4.3.9}$$

$$= exp(\Theta_n^T\mathbf{z}(\omega,t)) \tag{4.3.10}$$

where $\Theta_n = \{\ln\psi_{i\tau}(\omega)\}$. Comparing (4.3.9) to the general form (4.3.6), it can be deduced that

$$v(\mathbf{z}(\omega,t)) = z(\omega,t) \tag{4.3.11}$$

$$h(\mathbf{z}(\omega,t)) = 1 \tag{4.3.12}$$

$$f(\Theta_n) = 1 \tag{4.3.13}$$

By inspecting the form of the multinomial distribution $p(\mathbf{z}(\omega,t)|\boldsymbol{\psi})$, the conjugate prior is given by

$$p(\boldsymbol{\psi}|\alpha) \propto \prod_{i\tau}\psi_{i\tau}(\omega)^{\alpha_{i\tau}-1} \tag{4.3.14}$$

where $0 \leqslant \psi_{i\tau}(\omega) \leqslant 1$ and $\sum_{i\tau}\psi_{i\tau}(\omega) = 1$.
The normalized form of this distribution is called the Dirichlet distribution and is expressed as

$$Dir(\boldsymbol{\psi}|\alpha_0) = C(\alpha_0)\prod_{i\tau}\psi_{i\tau}(\omega)^{\alpha_0-1} \tag{4.3.15}$$

where α_0 is the distribution parameter assumed to be the same for all components and $C(\alpha_0)$ is the normalization constant [67]. Plots of the Dirichlet distribution for different values of α_0 are depicted in Figure 4.4.

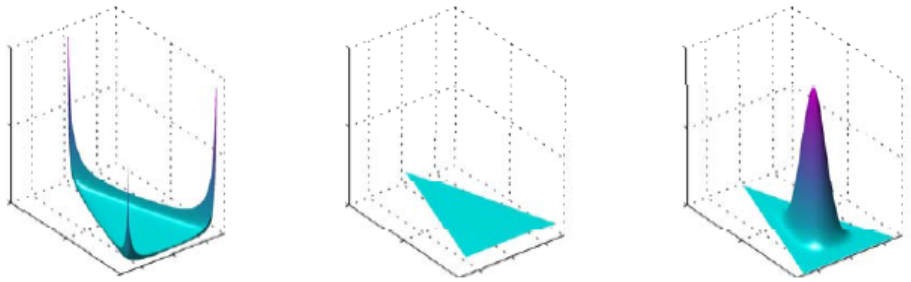


Figure 4.4: Plots of the Dirichlet distribution over three variables. The horizontal axes represent the coordinates in the plane of the simplex and the vertical axis is the value of the distribution. The left plot corresponds to $\alpha_0 = 0.1$, the middle is for $\alpha_0 = 1$ and the right for $\alpha_0 = 10$ [67].

The Gaussian-Wishart distribution

Since each component is modelled by a Gaussian distribution $\mathcal{N}(\hat{\phi}(\omega, t; \tau)|\xi_{i\tau}(\omega), \lambda_{i\tau}^{-1}(\omega))$, which belongs to the exponential family, the mean and the precision joint prior $p(\boldsymbol{\xi}, \boldsymbol{\lambda})$ exists and is represented by the Gaussian-Wishart distribution expressed as

$$p(\boldsymbol{\xi}, \boldsymbol{\lambda}) = p(\boldsymbol{\xi}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \tag{4.3.16}$$

$$= \prod_{i\tau} \mathcal{N}(\xi_{i\tau}(\omega)|m_0(\omega), (\beta_0\lambda_{i\tau}(\omega))^{-1}\mathcal{W}(\lambda_{i\tau}(\omega)|w_0, \nu_0) \tag{4.3.17}$$

where $\boldsymbol{\xi} = \{\xi_{i\tau}(\omega)\}$, $\boldsymbol{\lambda} = \{\lambda_{i\tau}(\omega)\}$ and $m_0(\omega)$, β_0, w_0, ν_0 are the Gaussian-Wishart distribution parameters [67]. $m_0(\omega)$ is chosen to be equal to the mean of the data [92] and hence is frequency dependent whereas β_0, w_0 and ν_0 are frequency independent and fixed a priori. The hyperparameters are generally chosen to give broad priors, and by symmetry are assumed equal for all components [67], [31].

Variational inference

Considering the data set $\mathbf{Y} = \{\hat{\phi}(\omega, t; \tau)\}$ and the latent variable set $\mathbf{Z} = \{\mathbf{z}(\omega, t; \tau)\}$, the joint distribution of the observed data and all the random variables can be decomposed as

$$p(\mathbf{Y}, \mathbf{Z}, \Theta) = p(\mathbf{Y}|\mathbf{Z}, \Theta).p(\mathbf{Z}|\Theta).p(\Theta) \tag{4.3.18}$$

and the evidence $p(\mathbf{Y})$ is expressed as

$$p(\mathbf{Y}) = \int_{\Theta} \sum_{\mathbf{Z}} p(\mathbf{Y}, \mathbf{Z}, \Theta) d\Theta \tag{4.3.19}$$

The evidence is intractable but its logarithm is lower bounded as explained in Section 4.2. The posterior distributions of all the hidden variables given the data set would be approximated by a distribution $q^*(\mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda})$ minimizing the Kullback-Leibler divergence functional [67] and satisfying the only assumption of the variational inference, namely

$$q^*(\mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = q^*(\mathbf{Z}) q^*(\boldsymbol{\psi}) q^*(\boldsymbol{\xi}, \boldsymbol{\lambda}) \tag{4.3.20}$$

The optimal distributions $q^*(\mathbf{Z})$, $q^*(\boldsymbol{\psi})$ and $q^*(\boldsymbol{\xi}, \boldsymbol{\lambda})$ have the same functional form as their priors [67].

Similarly to the EM algorithm, these variational posterior distributions are obtained in two steps. In the E-step, the current distributions are used to evaluate $E[z_{i\tau}(\omega, t)]$ followed by the M-step in which the parameters of the distributions are recomputed given the expected value of $z_{i\tau}(\omega, t)$. The VB EM update rules for GMMs are presented in the following section.

4.4 VB EM Update Rules

The posterior distributions of the latent variables and those of the component parameters can be obtained using the general result expressed in (4.2.9) as follows [67]:

$$\begin{aligned} \ln q^*(\mathbf{Z}) &= E_{\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\lambda}}[\ln p(\mathbf{Y},\mathbf{Z},\boldsymbol{\psi},\boldsymbol{\xi},\boldsymbol{\lambda}] + const \\ &= E_{\boldsymbol{\psi}}[\ln p(\mathbf{Z}|\boldsymbol{\psi}] + E_{\boldsymbol{\xi},\boldsymbol{\lambda}}[\ln p(\mathbf{Y}|\mathbf{Z},\boldsymbol{\xi},\boldsymbol{\lambda}] + const \end{aligned} \tag{4.4.1}$$

Substituting with the conditional distributions using (4.3.4) and (4.3.5) results into

$$\ln q^*(\mathbf{Z}) = \sum_t \sum_{i\tau} z_{i\tau}(\omega,t) \ln \rho_{i\tau}(\omega,t) \tag{4.4.2}$$

where

$$\begin{aligned} \ln \rho_{i\tau}(\omega,t) &= E[\ln \psi_{i\tau}(\omega)] + \frac{1}{2}E[\ln \lambda_{i\tau}(\omega)] - \frac{1}{2}\ln(2\pi) \\ &- \frac{1}{2}E_{\xi_{i\tau},\lambda_{i\tau}}[\big(\hat{\phi}(\omega,t;\tau) - \xi_{i\tau}(\omega)\big)^2] \end{aligned} \tag{4.4.3}$$

Taking the exponential of both sides of (4.4.2) gives

$$q^*(\mathbf{Z}) = \prod_t \prod_{i\tau} r_{i\tau}(\omega,t)^{z_{i\tau}(\omega,t)} \tag{4.4.4}$$

The posterior distribution $q^*(\mathbf{Z})$ has the same functional form of its prior $P(\mathbf{Z}|\boldsymbol{\psi})$. The $r_{i\tau}(\omega,t)$ terms which define the expectations of the latent variables $z_{i\tau}(\omega,t)$ are computed within the E-step using (4.3.3) as follows

$$\begin{aligned} E[z_{i\tau}(\omega,t)] &= r_{i\tau}(\omega,t) \\ &= \frac{\rho_{i\tau}(\omega,t)}{\sum_{i\tau}\rho_{i\tau}(\omega,t)} \end{aligned} \tag{4.4.5}$$

The following three statistics related to $r_{i\tau}(\omega, t)$ are defined as

$$N_{i\tau}(\omega) = \sum_t r_{i\tau}(\omega, t) \tag{4.4.6}$$

$$\bar{\phi}_{i\tau}(\omega) = \frac{1}{N_{i\tau}(\omega)} \sum_t \hat{\phi}(\omega, t; \tau) r_{i\tau}(\omega, t) \tag{4.4.7}$$

$$S_{i\tau(\omega)} = \frac{1}{N_{i\tau}(\omega)} \sum_t \big((\hat{\phi}(\omega, t; \tau) - \bar{\phi}_{i\tau}(\omega))^2 r_{i\tau}(\omega, t) \big) \tag{4.4.8}$$

and are also computed and used in the evaluation of the parameters of the variational posterior distributions in the M-step.

These statistics are similar to the quantities estimated in the classical EM for GMM (A.1.4)-(A.1.6), where $N_{i\tau}(\omega)$ is interpreted as the effective number of points associated to a source i and delay τ, $\bar{\phi}_{i\tau}(\omega)$ is the weighted mean of the data and $S_{i\tau(\omega)}$ is the corresponding weighted variance.

Similar to the posterior distribution of the latent variables, the log of the posterior distribution of the component parameters can be obtained as follows [67]

$$\ln q^*(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda}) = E_{\mathbf{Z}}[\ln p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda}] + const \tag{4.4.9}$$

$$\begin{aligned} &= \ln p(\boldsymbol{\psi}) + \sum_{i\tau} \ln p\big(\xi_{i\tau}(\omega), \lambda_{i\tau}(\omega)\big) + E_{\mathbf{Z}}[\ln p(\mathbf{Z}|\psi)] \\ &+ \sum_t \sum_{i\tau} E[z_{i\tau}(\omega, t)] \ln \mathcal{N}\big(\hat{\phi}(\omega, t; \tau) | \xi_{i\tau}(\omega), \lambda_{i\tau}^{-1}(\omega)\big) + const \end{aligned} \tag{4.4.10}$$

It can been seen from (4.4.10) that $\ln q^*(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\lambda})$ decomposes into the sum of terms involving only $\boldsymbol{\psi}$ and other terms depending only on $\xi_{i\tau}(\omega)$ and $\lambda_{i\tau}(\omega)$. Identifying the terms depending on ψ yields

$$\ln q^*(\boldsymbol{\psi}) = (\alpha_0 - 1) \sum_{i\tau} \ln \psi_{i\tau}(\omega) + \sum_{i\tau} \sum_t r_{i\tau}(\omega, t) \ln \psi_{i\tau}(\omega) + const \tag{4.4.11}$$

Hence, $q^*(\boldsymbol{\psi})$ is found to have a Dirichlet distribution

$$q^*(\boldsymbol{\psi}) = Dir(\boldsymbol{\psi}|\boldsymbol{\alpha}) \tag{4.4.12}$$

where $\boldsymbol{\alpha} = \{\alpha_{i\tau}(\omega)\}$.

The parameter of the updated Dirichlet distibution is given by

$$\alpha_{i\tau}(\omega) = \alpha_0 + N_{i\tau}(\omega) \tag{4.4.13}$$

By inspecting (4.4.9) and considering the terms involving $\xi_{i\tau}(\omega)$ and $\lambda_{i\tau}(\omega)$, their joint distribution at each frequency ω, is found to be a Gaussian-Wishart distribution

$$q^*(\boldsymbol{\xi}, \boldsymbol{\lambda}) = q^*(\boldsymbol{\xi}|\boldsymbol{\lambda})q^*(\boldsymbol{\lambda}) \tag{4.4.14}$$

$$= \prod_{i\tau} \mathcal{N}(\xi_{i\tau}(\omega)|m_{i\tau}(\omega), (\beta_{i\tau}(\omega)\lambda_{i\tau}(\omega))^{-1}\mathcal{W}(\lambda_{i\tau}(\omega)|w_{i\tau}(\omega), \nu_{i\tau}(\omega)) \tag{4.4.15}$$

The parameters of the updated Gaussian-Wishart distribution are defined as follows

$$\beta_{i\tau}(\omega) = \beta_0 + N_{i\tau}(\omega) \tag{4.4.16}$$

$$m_{i\tau}(\omega) = \frac{1}{\beta_{i\tau}(\omega)}(\beta_0 m_0(w) + N_k\bar{\phi}_{i\tau}(\omega)) \tag{4.4.17}$$

$$\begin{aligned} w_{i\tau}(\omega)^{-1} = {} & w_0^{-1} + N_{i\tau}(\omega)S_{i\tau}(\omega) \\ & + \frac{\beta_0 N_{i\tau}(\omega)}{\beta_0 + N_{i\tau}(\omega)}(\bar{\phi}_{i\tau}(\omega) - m_0)^2 \end{aligned} \tag{4.4.18}$$

$$\nu_{i\tau}(\omega) = \nu_0 + N_{i\tau}(\omega) \tag{4.4.19}$$

These parameters are then used to compute the set of expectations $E[\ln \psi_{i\tau}(\omega)]$, $E[\ln \lambda_{i\tau}(\omega)]$ and $E_{\xi_{i\tau},\lambda_{i\tau}}[(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega))^2]$ required for es-

timating $r_{i\tau}(\omega, t)$

$$\begin{aligned} &E_{\xi_{i\tau},\lambda_{i\tau}}[\big(\hat{\phi}(\omega,t;\tau)-\xi_{i\tau}(\omega)\big)^2] = 1/\beta_{i\tau}(\omega) \\ &+\nu_{i\tau}(\omega)\Big(\hat{\phi}(\omega,t;\tau)-m_{i\tau}(\omega)\Big)^2 w_{i\tau}(\omega) \end{aligned} \tag{4.4.20}$$

$$E[\ln \lambda_{i\tau}(\omega)] = \psi\big(\frac{\nu_{i\tau}(\omega)}{2}\big) + \ln 2 + \ln w_{i\tau}(\omega) \tag{4.4.21}$$

$$E[\ln \psi_{i\tau}(\omega)] = \psi(\alpha_{i\tau}(\omega)) - \psi\Big(\sum_{i\tau}\alpha_{i\tau}(\omega)\Big) \tag{4.4.22}$$

where $\psi(.)$ is the digamma function [67].

The variational optimization of the posterior distribution involves cycling between two stages as summarized in Table 4.1. In the variational equivalent of the E-step, the current estimates of the parameters are used to compute the moments in (4.4.20)-(4.4.21), required to determine the expectations of the latent variables $r_{i\tau}(\omega, t)$. These expectations are then used in the variational M-step to re-estimate the posterior distributions in (4.4.12) and (4.4.14). The posterior distributions have the same functional forms as their priors as a general result from the choice of conjugate distributions [67].

After convergence, the mask extracting each source i from the microphone signals $L(\omega, t)$ or $R(\omega, t)$ can be determined by summing the expectations of the latent indicators $r_{i\tau}(w, t)$ over the delay τ

$$M_i(\omega, t) \equiv \sum_{\tau} r_{i\tau}(\omega, t) \tag{4.4.23}$$

Table 4.1: VB EM update rules for GMMs

1. E-step:

Compute the expectations of the latent variables $z_{i\tau}(\omega, t)$, $r_{i\tau}(\omega, t)$ using the set of expectations $E[\ln \psi_{i\tau}(\omega)]$, $E[\ln \lambda_{i\tau}(\omega)]$ and $E_{\xi_{i\tau},\lambda_{i\tau}}[(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega))^2]$.

2. M-step:

Using $r_{i\tau}(\omega, t)$, estimate the parameters of the updated posterior distributions, $\alpha_{i\tau}(\omega)$, $\beta_{i\tau}(\omega)$, $m_{i\tau}(\omega)$, $w_{i\tau}(\omega)$ and $\nu_{i\tau}(\omega)$.

The main advantage of the VB framework in comparison to the classical EM algorithm is the absence of singularities arising from the likelihood optimization. Simulation studies presented in the following section confirm as well an improvement in the separation performance particularly for sources in close proximity which demonstrates the robustness of the Bayesian treatment in modelling uncertainties in real environments [93].

4.5 Experimental results

Two main experiments were performed in order to evaluate the performance of the proposed approach and compare it to MESSL. The first experiment examines the impact of the Dirichlet distribution hyperparameter choice on the separation performance. The second experiment compares the proposed algorithm with two versions of MESSL for the cases of two and three sources.

4.5.1 Data sources

Speech utterances of 2.5 s long were chosen randomly from the whole TIMIT database [76]. Mixtures were formed from different combinations of male and female signals sampled at 16 KHz. These signals were normalized to have the same energy. All experiments included either two or three speakers, one target and either one or two interferers.

4.5.2 Room Impulse responses

The BRIRs [78] described in Chapter 3 are used. Focusing on the case where sources are in close proximity, three different azimuthal positions for the interferer were tested [15°, 30°, 45°], in the case of two speakers. In the three-speaker case, the second interferer is located symmetrically with the same azimuth as shown in Figure 4.5. The target was always positioned facing the microphones.

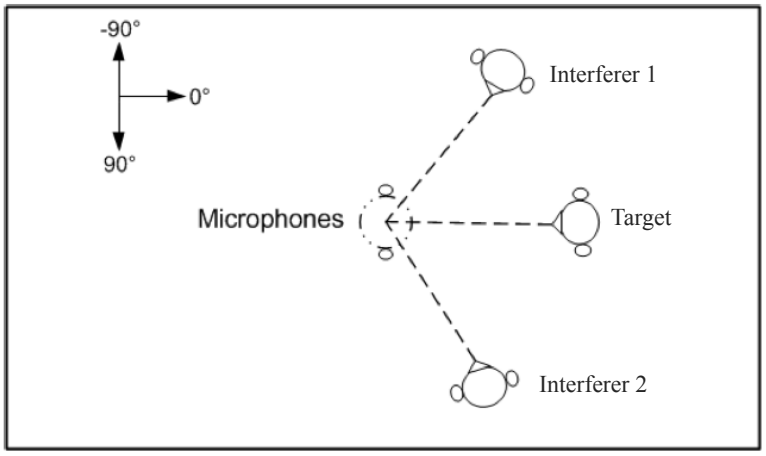


Figure 4.5: The room layout showing approximate positions of the sources and the microphones.

All speech sources are located at a distance of 1 m from the center of the microphones. The separation performance was evaluated objectively using

SDR [48].

4.5.3 Initialization

Let Θ_Ω denote the complexity in which IPD parameters vary with the frequency. In MESSL, this complexity results in a better separation than the frequency independent version but requires a bootstrapping approach to avoid local maxima [25]. The proposed approach also assumes frequency dependent parameters with less complexity as no bootstrapping is required [31]. The set of hyperparameters can be fixed a priori or can be inferred from the data. In all experiments, β_0 and $m_0(\omega)$ were set following [92], where $\beta_0 = 0.01$, $m_0(\omega)$ is equal to the mean of the data at each frequency and ν_0 was chosen empirically equal to 20 as smaller values resulted in slower convergence.

4.5.4 Dirichlet distribution hyperparameter

The Dirichlet distribution hyperparameter α_0 plays an important role in variational clustering as it can be seen as the effective prior number of observations associated with each component [67]. Solutions obtained for $\alpha_0 < 1$ correspond to the case where more mixing coefficients are equal to zero which better describes the sparseness of the speech sources in T-F domain.

Individual SDRs obtained for five mixtures using the proposed approach with different values of α_0 are shown in Table 1 and Table 2. Poor choice of prior distribution might affect the effectiveness of the VB approach as indicated in Table 2, where $\alpha_0 = 10$ and the average SDRs have been reduced by 0.9 dB, 1.1 dB and 1.2 dB for the three azimuthal separation angles respectively.

Table 4.1: SDR (dB) proposed approach Θ_Ω, $\alpha_0 = 0.1$

Azimuth angles	15°	30°	45°
mix1	2.53	3.49	2.79
mix2	3.57	3.1	5.1
mix3	3.55	3.99	4.95
mix4	3.16	3.08	3.3
mix5	2.98	2.17	3.35
Average	3.16	3.16	3.9

Table 4.2: SDR (dB) proposed approach Θ_Ω, $\alpha_0 = 10$

Azimuth angles	15°	30°	45°
mix1	1.43	2.62	1.5
mix2	2.62	2.37	4.29
mix3	2.9	2.99	4.04
mix4	1.97	1.09	1.54
mix5	2.31	1.01	2.05
Average	2.24	2.01	2.68

4.5.5 Comparison with MESSL

Ten different mixtures were randomly formed in total from the TIMIT database and the average SDR results comparing the proposed approach ($\alpha_0 = 0.1$) with two versions of MESSL are shown in Table 4.3 and Table 4.4 for two and three speakers, respectively. It can be seen that adding ILD cues for small separation angles does not improve the separation which is expected since both spatial cues get more similar as the sources move closer [25]. On the other hand, exploiting the VB clustering framework improves the estimation of the parameters of IPD cues for sources in close proximity, resulting in more accurate masks and a better separation.

Table 4.3: Separation performance comparison in terms of average SDR (dB) for the two-speaker case

Azimuth angles	15°	30°	45°	avg
MESSL IPD	2.38	2.62	3.67	2.89
MESSL IPD-ILD	1.92	2.22	3.47	2.54
Variational IPD	3.61	3.42	4.13	3.72

Table 4.4: Separation performance comparison in terms of average SDR (dB) for the three-speaker case

Azimuth angles	15°	30°	45°	avg
MESSL IPD	-0.67	0.09	2.11	0.51
MESSL IPD-ILD	-1.15	-0.02	2.22	0.35
Variational IPD	0.8	1.22	2.97	1.66

The average SDR improvement of the proposed approach decreases with the azimuthal separation. For the two-speaker case, the average SDR improvements obtained using the variational approach are 1.2 dB, 0.8 dB and 0.5 dB compared to the first version of MESSL. Whereas, compared to the second version MESSL IPD-ILD, the average SDR improvements obtained are 1.7 dB, 1.2 dB and 0.7 dB for the three azimuthal angles respectively. In Table 4, for the case of three speakers these improvements increased to 1.5 dB, 1.1 dB and 0.9 dB compared to the first version and 1.9 dB, 1.2 dB and 0.8 dB compared to the second version.

Using the same probabilistic modelling of the interaural cues and only a different clustering framework based on the VB methodology has shown a significant improvement in the average separation quality (over different azimuthal separation). For the case of two sources, an increase of 0.8 dB of the average SDR is obtained. This increase is equivalent to 28% improvement relative to the EM clustering algorithm employed in MESSL. The average SDR improvement increases to 1.2 dB in the case of three sources, which is equivalent to an increase of 225% relative to MESSL.

4.6 Summary

In this chapter, a variational Bayesian framework was proposed as an alternative to the EM algorithm based on likelihood maximization. The proposed framework is used for clustering spectrogram points depending only on their IPD cues. This elegant approach overcomes the drawbacks of the popular EM for GMMs as it avoids over-fitting and the presence of singularities associated with the likelihood optimization without requiring additional extensive computations. More importantly, with proper initialization and careful choice of hyperparameters values, experimental results confirmed a significant improvement of the separation performance particularly for nearby sources. In the next chapter, integrating the robust clustering resulting from the non-Gaussian modelling within the variational Bayesian framework, will be considered to cluster the spectrogram points based on both interaural phase and level difference cues.

Chapter 5

ROBUST VARIATIONAL BAYESIAN CLUSTERING FOR UNDERDETERMINED SPEECH SEPARATION

In this chapter, non-Gaussian modelling is integrated within a variational Bayesian framework to jointly cluster interaural spatial cues. At each spectrogram point, Student's t-distribution mixture models (SMMs) are used to define the probabilistic models of IPD and ILD cues. The parameters of these models are determined via a variational expectation-maximization (VEM) algorithm. As a result, soft probabilistic masks are generated for source separation. The proposed framework overcomes the limitations of other T-F masking algorithms employing the traditional EM framework to cluster interaural cues modelled using GMMs. Compared to GMMs commonly used, the heavy tailed SMMs are less sensitive to outlier values resulting in more accurate masks. More importantly, the variational inference overcomes the difficulties of the likelihood optimization especially the probable presence of singularities and over-fitting. The proposed statistical framework substantially improves the separation quality in the presence of reverberations particularly for sources in close proximity. An extensive set of simulation

studies evaluating the proposed approach using TIMIT database speech utterances and binaural real impulse responses (BRIRs) confirms its advantage in terms of both objective and subjective separation performance measures.

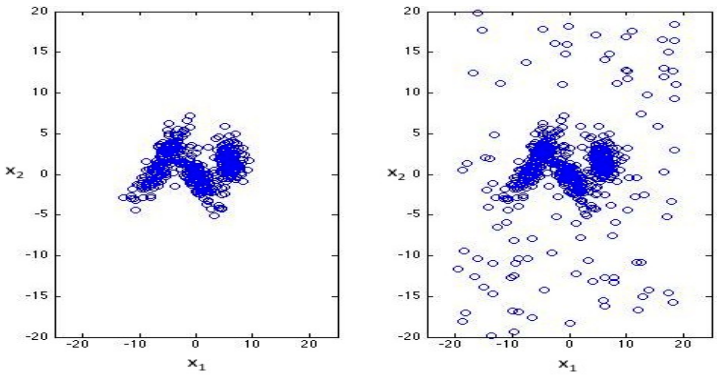
5.1 Introduction

The Bayesian approach for clustering has been used as previously discussed to avoid the limitations of maximum likelihood estimation. In addition to the probable presence of singularities occurring whenever a data point coincides with one of the Gaussian components, other convergence problems of the EM algorithm were thoroughly discussed in [93]. It was shown that although EM is guaranteed to converge theoretically to the value maximizing the likelihood function, in practice it fails particularly for sparse data. This convergence problem was associated to two situations, the presence of outliers in the data set or data repetitions. For practical situations, Bayesian estimation is known to be more appropriate as likelihood estimation often diverges [94]. Taking the sources of uncertainty into account explicitly when estimating model parameters better represents uncertainty than the frequentist likelihood maximization approach particularly for noisy environments.

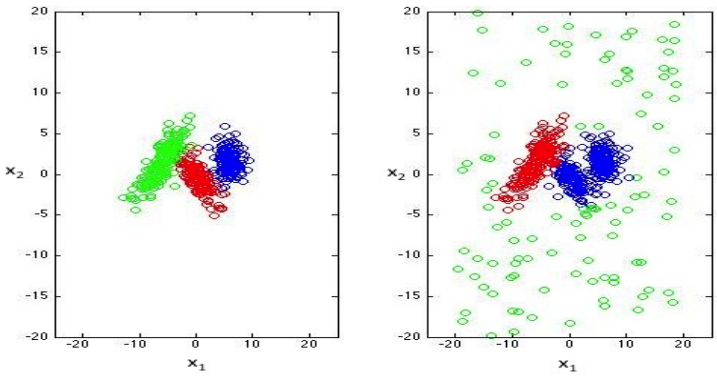
In [95], a variational inference algorithm was proposed by Svensén and Bishop to cluster a mixture of Student's t-distribution, which is considered an infinite mixture of concentric Gaussians and hence includes GMMs as a special case. Compared to other approaches for solving the Bayesian problem such as MCMC; the variational approach is found to be more computational efficient with relatively small overhead compared to the classical EM algorithm. In both EM and VEM the major computations result from the evaluation of the precision matrices [95]. A key adavantage for this algorithm was the robust estimation of the mean of each cluster compared to the estimates obtained assuming Gaussian modelling. This algorithm was

modified by Archambeau in [31] to improve the robustness by considering the dependence between the scaling variables of the Student's t-distribution and the binary latent indicators as it will be shown later in this chapter. This modified framework was tested on the Old faithful Geyser eruption data [95] and on the Enzymatic activity blood data [96] and has proven to be more robust to outliers. It was also applied in [97] as part of the European project OPTIVIP (optimization of the visual implantable prosthesis) to model the neurophysiological process linking certain stimulation parameters to their corresponding visual sensations with the ultimate goal of designing a system capable of restoring the vision for blind individuals.

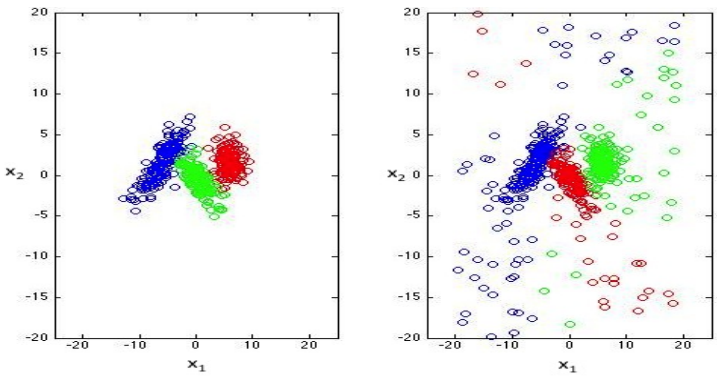
The robustness of the SMM approach employed in the VB framework can be illustrated in Figure 5.1, where both an SMM and a GMM are employed for clustering a mixture of three bivariate Gaussian distributions [31]. Two different sets are used, the first set on the left-hand side has no outliers while the second set contains 25% of outliers randomly chosen from a uniform distribution defined over the interval [-20 20], in each direction of the space [31]. In the absence of outliers, both GMM and SMM perform similarly and three different clusters corresponding to the original ones are clearly depicted by three different colours on the left-hand side of Figure 2(b) and Figure 2(c) respectively. However, in the presence of outliers, only SMM successfully selects the original clusters, whereas the GMM as seen in the right-hand side of Figure 2(b) falsely identifies three different clusters, only one cluster (in red) corresponds to the original one.



(a)



(b)



(c)

Figure 5.1: (a) Data on the left are obtained from mixing three bivariate Gaussian distributions with different means and precisions. On the right, the same data with 25% of outliers from a uniform distribution [-20 20]. (b) GMM successfully identifies the clusters in the case of no outliers but fails when data are corrupted. (c) SMM successfully identifies the clusters in both cases.

The reverberations present in a typical CPP context affect the auditory perception in different ways. While early reflections (within 50 ms-80 ms of the direct sound) improve the speech audibility, late reverberations distort the temporal information in the speech waveform and degrade intelligibility [78]. Various algorithms in reverberant environments were studied and compared in [25] at a separation distance $\geq$ 1 m. For this case, using ITDs and ILDs determines the source location within a "cone of confusion". However, for a distance $\leq$ 1 m, the location can be determined within a"torus of confusion" [98]. Typical CPP environments generally involve near by sources and small changes in the source location relative to the listener result in large variations in the direct-sound energy arriving at both ears/microphones compared to the case of distant sources. In other words, the correlation between source location and the impact of reverberation is maximized for nearby sources. Hence, the choice of a robust statistical framework that better models uncertainties in reverberant environments is crucial for the generation of accurate masks capable of extracting a speech of interest for sources in close proximity.

In order to achieve robust density estimation, robust clustering and improve the speech separation, a novel probabilistic T-F masking approach is proposed in this chapter. Based on the joint clustering of IPD and ILD cues, this approach integrates the non-Gaussian modelling of these cues into a VB framework.

In Section 5.2, the probabilistic modelling of a single source is explained. In Section 5.3, the variational Bayesian inference and the update rules for SMMs are presented. The performance of the proposed approach using real impulse responses is evaluated in Section 5.4. Finally, the chapter is summarized in Section 5.5.

5.2 Single source modelling

Focusing on the azimuthal cues, the interaural spectrogram is expressed as [25]

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{5.2.1}$$

where $L(\omega,t)$ and $R(\omega,t)$ are the spectrograms of a sound source arriving at two spatially distinct microphones; $\alpha(\omega,t)$ and $\phi(\omega,t)$ denote the ILD and the IPD, respectively. In order to avoid spatial aliasing [25], the phase residual $\hat{\phi}(\omega,t;\tau) = arg\Big(e^{j\phi(\omega,t)}e^{-j\omega\tau(\omega)}\Big)$ is used instead of $\phi(\omega,t)$. The frequency dependent delay $\tau(\omega)$ can be decomposed as follows [25]

$$\tau(\omega) = \tau + \omega^{-1}\zeta(\omega) \tag{5.2.2}$$

The set of delays τ is specified in advance, whereas the second term depends on the frequency dependent mean of $\hat{\phi}(\omega,t;\tau)$ denoted by $\zeta(\omega)$. The phase residual is the difference between the IPD resulting from a delay of τ samples and the actual IPD and is constrained to the interval $(-\pi,\pi)$.

The ILD and the residual IPD are jointly combined in a bivariate vector $\mathbf{y} \equiv \mathbf{y}(\omega,t;\tau) = [\hat{\phi}(\omega,t;\tau), \alpha(\omega,t)]^T$ which can be modelled by a Student's t-distribution

$$P(\mathbf{y}|\theta) = St(\mathbf{y}|\boldsymbol{\mu}(\omega), \boldsymbol{\Lambda}(\omega), \nu) \tag{5.2.3}$$

where $(.)^T$ is the transpose operator, $\theta \equiv \{\boldsymbol{\mu}(\omega), \boldsymbol{\Lambda}(\omega)\}$ is the set of model parameters and the degree of freedom ν is fixed a priori.

The multivariate Student's t-distribution is defined as [31]

$$St(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^\infty \mathcal{N}\big(\mathbf{y}|\boldsymbol{\mu},u\boldsymbol{\Lambda}\big)\mathcal{G}\big(u|\nu/2,\nu/2\big)du \tag{5.2.4}$$

$$= \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{d}{2}}}|\boldsymbol{\Lambda}|^{\frac{1}{2}}\big[1+\frac{1}{\nu}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})\big]^{-\frac{d+\nu}{2}} \tag{5.2.5}$$

The Gaussian and the Gamma distributions are given by [31]

$$\mathcal{N}\big(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Lambda}\big) = (2\pi)^{-\frac{d}{2}}e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\dagger\boldsymbol{\Lambda}(\mathbf{y}-\boldsymbol{\mu})} \tag{5.2.6}$$

$$\mathcal{G}\big(u|\kappa,\eta\big) = \frac{\eta^\kappa}{\Gamma(\kappa)}u^{\kappa-1}e^{-\eta u} \tag{5.2.7}$$

where $\boldsymbol{\mu}$ and Λ are the Gaussian mean vector and precision matrix, respectively. $\Gamma(.)$ is the Gamma function, κ and η are the Gamma distribution parameters and d is the dimension of the feature space.

In this work, underdetermined speech separation is based on the ILD and the residual IPD cues. However, the proposed probabilistic framework is flexible and other relevant cues such as the mixing vectors in [53] and [58] and monaural cues in [35] can be combined without enforcing any condition on their independence.

5.3 Bayesian Student's t-Distribution Mixture Models

Based on the approximate disjointness of speech in the T-F domain, each spectrogram point can be associated to a source i and delay $\tau(\omega)$. The number of sources I is assumed known. Spectrogram points belonging to the same source and delay are distributed identically. However, their corresponding model parameters can only be determined if the source dominating each spectrogram point and its delay were identified. This classical clustering

problem can be represented by a finite SMM defined as

$$P(\mathbf{y}|\Theta) = \sum_{i,\tau} \psi_{i\tau}(\omega) St(\mathbf{y}|\boldsymbol{\mu}_{i\tau}(\omega), \boldsymbol{\Lambda}_{i\tau}(\omega), \nu) \tag{5.3.1}$$

where $\Theta \equiv \{\boldsymbol{\mu}_{i\tau}(\omega), \boldsymbol{\Lambda}_{i\tau}(\omega), \psi_{i\tau}(\omega)\}$ denotes the set of the model parameters. By symmetry, ν is assumed equal for all components. The mixing coefficients $\psi_{i\tau}$ at each frequency, satisfy these conditions $\psi_{i\tau} \geq 0$ and $\sum_{i\tau} \psi_{i\tau} = 1$.

Each spectrogram point has a corresponding latent binary indicator vector $\mathbf{z}(\omega, t)$ consisting of $z_{i\tau}(\omega, t)$ which are equal to unity if the point originates from a source i and delay τ and equal to 0 otherwise, such that $\sum_{i,\tau} z_{i\tau}(\omega, t) = 1$.

Since the Student's t-distribution defined in (5.2.4) has an unobserved random scaling u, another set of latent scaling vectors $\mathbf{U} = \{\mathbf{u}(\omega, t)\}$ is defined; $\mathbf{u}(\omega, t)$ comprises the scaling variables $u_{i\tau}(\omega, t)$. The graphical representation of the Bayesian SMM is shown in Fig. 5.2.

The number of latent variables $\mathbf{z}(\omega, t)$ and $\mathbf{u}(\omega, t)$ increases with the size of the data set. In a Bayesian framework, the parameters are also considered as random variables with fixed size independent of the data size. This framework differs from that proposed in [95], in which it was assumed that the scaling variables and the binary indicators are independent. Independence between the component means and their corresponding precisions was also assumed. Avoiding unnecessary assumptions and taking the dependencies between random variables into account, improve the robustness of Bayesian estimation as demonstrated in [31]. The latent variable model and the parameter priors are presented next.

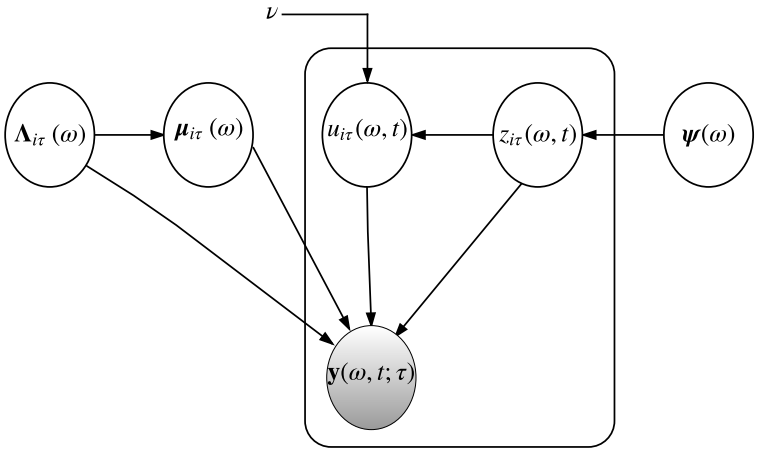


Figure 5.2: Directed graph of the Bayesian SMM at each spectrogram point. The shaded node represents the observed vector $\mathbf{y}(\omega, t; \tau)$. The arrow direction indicates dependencies between random variables. The scaling $u_{i\tau}(\omega, t)$ conditionally depend on the binary indicators $z_{i\tau}(\omega, t)$ and the component means $\boldsymbol{\mu}_{i\tau}(\omega)$ depend on the precision $\boldsymbol{\Lambda}_{i\tau}(\omega)$.

The latent variable model

For each data point $\mathbf{y}(\omega, t; \tau)$, denoted by $\mathbf{y}$ for convenience of notation, the latent variable model can be specified as follows [31]:

$$p(\mathbf{z}(\omega,t)|\Theta) = \prod_{i,\tau} \psi_{i\tau}(\omega)^{z_{i\tau}(\omega,t)} \tag{5.3.2}$$

$$p(\mathbf{u}(\omega,t)|\mathbf{z}(\omega,t),\Theta) = \prod_{i,\tau} \mathcal{G}\big(u_{i\tau}(\omega,t)\Big|\frac{\nu}{2},\frac{\nu}{2}\big)^{z_{i\tau}(\omega,t)} \tag{5.3.3}$$

$$\begin{aligned} &p(\mathbf{y}|\mathbf{z}(\omega,t),\mathbf{u}(\omega,t),\Theta) = \\ &\prod_{i,\tau} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i\tau}(\omega), u_{i\tau}(\omega,t)\boldsymbol{\Lambda}_{i\tau}(\omega))^{z_{i\tau}(\omega,t)} \end{aligned} \tag{5.3.4}$$

where $\boldsymbol{\psi} = \{\psi_{i\tau}(\omega)\}$.

Parameter priors

To complete the Bayesian framework, the priors over the parameters are introduced. At each frequency ω, conjugate prior distributions are always considered, so that the posterior distributions have the same functional forms as their priors [67], [31]. For the sake of notational simplicity, the frequency dependence is dropped in (5.3.5) and (5.3.6).

The conjugate prior of the multinomial distribution $p(\mathbf{z}(\omega,t)|\boldsymbol{\psi})$ is the Dirichlet density,

$$p(\boldsymbol{\psi}) = Dir(\boldsymbol{\psi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{i\tau}\psi_{i\tau}^{\alpha_0-1} \tag{5.3.5}$$

where α_0 is the Dirichlet parameter assumed equal for all components and $C(\boldsymbol{\alpha}_0)$ is the normalization constant.

Similarly, the conjugate of a Gaussian distribution is the Gaussian-Wishart prior, therefore the mean and the precision joint prior is given by

$$\begin{aligned} p(\boldsymbol{\mu},\boldsymbol{\Lambda}) &= p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ &= \prod_{i\tau}\mathcal{N}(\boldsymbol{\mu}_{i\tau}|\mathbf{m_0},\beta_0\boldsymbol{\Lambda}_{i\tau})\mathcal{W}(\boldsymbol{\Lambda}_{i\tau}|\mathbf{S}_0,\gamma_0) \end{aligned} \tag{5.3.6}$$

where $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_{i\tau}\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_{i\tau}\}$, and $\mathbf{m}_0$, β_0, $\mathbf{S}_0$, γ_0 are the Gaussian-Wishart distribution hyperparameters [31]. The hyperparameters are generally chosen to give broad priors, and by symmetry are assumed equal for all components [67], [31].

The full Bayesian solution for this SMM clustering problem is not analytically tractable and should be approximated [31]. Variational methods discussed in Chapter 4, provide an approximate solution [86]. They can be interpreted as optimization problems, where the quantity to be maximized/minimized is a functional and approximate solutions are obtained by optimizing the given functional over a restricted range of functions [90] as

explained in detail in the following section.

5.4 Variational Bayesian EM for the SMM

Considering the data set $\mathbf{Y} = \{\mathbf{y}(\omega, t; \tau)\}$, the goal is the estimation of the model evidence and the posterior distributions of all the latent variables given the data set. The model evidence is given by

$$p(\mathbf{Y}) = \int_{\Theta} \int_{\mathbf{U}} \sum_{\mathbf{Z}} p(\mathbf{Y}, \mathbf{U}, \mathbf{Z}, \Theta) d\mathbf{U} d\Theta \tag{5.4.1}$$

The joint distribution of the observed and hidden variables can be decomposed according to the dependencies shown in Figure 5.2 as

$$p(\mathbf{Y}, \mathbf{U}, \mathbf{Z}, \Theta) = p(\mathbf{Y}|\mathbf{U}, \mathbf{Z}, \Theta).p(\mathbf{U}|\mathbf{Z}, \Theta).p(\mathbf{Z}|\Theta).p(\Theta) \tag{5.4.2}$$

The first three factors are obtained from the latent variable model defined in (5.2.9)-(5.2.11), whereas $p(\Theta) = p(\boldsymbol{\psi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ can be factorized into $p(\boldsymbol{\psi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, which are defined in (5.2.12)-(5.2.13).

The model evidence is intractable but the lower bound of its logarithm can be expressed as follows

$$\ln p(\mathbf{Y}) \geq \ln p(\mathbf{Y}) - KL[q(\mathbf{U}, \mathbf{Z}, \Theta)||p(\mathbf{U}, \mathbf{Z}, \Theta|\mathbf{Y})] \tag{5.4.3}$$

where KL denotes the Kullback-Leibler divergence functional. Maximizing the lower bound of the evidence is equivalent to minimizing the KL between the approximate posterior and the true one. This leads to the estimated posterior distribution $q(\mathbf{U}, \mathbf{Z}, \Theta)$ satisfying the factorization assumption of the variational approach [31],

$$q(\mathbf{U}, \mathbf{Z}, \Theta) = q(\mathbf{U}, \mathbf{Z})q(\Theta) \tag{5.4.4}$$

The posterior distributions $q(\mathbf{U}, \mathbf{Z})$ and $q(\Theta)$ have the same forms as their priors [31]. Minimizing the KL functional leads to the VB EM update rules. In the E-step, the current distributions are used to evaluate $q(\mathbf{U}, \mathbf{Z})$ followed by the M-step in which $q(\Theta)$ is recomputed given the distributions estimated in the E-step.

5.4.1 VB EM Update Rules

Within the E-step, $q(\mathbf{u}(\omega,t), \mathbf{z}(\omega,t))$ can be obtained as follows [31]

$$q(\mathbf{u}(\omega,t), \mathbf{z}(\omega,t)) \propto exp\big(E_{\Theta}\{\ln p(\mathbf{y}, \mathbf{u}(\omega,t), \mathbf{z}(\omega,t)|\Theta\}\big) \tag{5.4.5}$$

where $E_{\Theta}\{.\}$ denotes the expectation taken with respect to the posterior distribution $q(\Theta)$. Details of the derivation can be found in Appendix B, only the update rules for the latent variable expectations are shown below.

E-step

The expected value of the latent indicator $z_{i\tau}(\omega,t)$ denoted $r_{i\tau}(\omega,t)$ is computed as

$$r_{i\tau}(\omega,t) = \frac{\rho_{i\tau}(\omega,t)}{\sum_{i\tau}\rho_{i\tau}(\omega,t)} \tag{5.4.6}$$

where

$$\begin{aligned}\rho_{i\tau}(\omega,t) \propto & \frac{\Gamma(\frac{d+\nu}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{d}{2}}}\tilde{\psi}_{i\tau}(\omega)\tilde{\Lambda}_{i\tau}^{\frac{1}{2}}(\omega) \\ .\Big[1 + & \frac{\gamma_{i\tau}(\omega)}{\nu}\Big(\mathbf{y}-\mathbf{m}_{i\tau}(\omega)\Big)^{T}\mathbf{S}_{i\tau}^{-1}(\omega)\Big(\mathbf{y}-\mathbf{m}_{i\tau}(\omega)\Big) \\ & + \frac{d}{\nu\beta_{i\tau}(\omega)}\Big]^{-\frac{d+\nu}{2}}\end{aligned} \tag{5.4.7}$$

and $\beta_{i\tau}(\omega)$, $\mathbf{m}_{i\tau}(\omega)$, $\mathbf{S}_{i\tau}(\omega)$ together with $\gamma_{i\tau}(\omega)$ denote the parameters of the posterior Gaussian-Wishart distribution.

The special quantity $\tilde{\psi}_{i\tau}(\omega)$ is estimated as follows

$$\begin{aligned} \ln \tilde{\psi}_{i\tau}(\omega) &= E[\ln \psi_{i\tau}(\omega)] \\ &= \psi(\alpha_{i\tau}(\omega)) - \psi\Big(\sum_{i\tau} \alpha_{i\tau}(\omega)\Big) \end{aligned} \tag{5.4.8}$$

where $\psi(.)$ is the digamma function [67].

Similarly,

$$\begin{aligned} \ln \tilde{\Lambda}_{i\tau}(\omega) &= E\big[\ln |\Lambda_{i\tau}(\omega)|\big] \\ &= \sum_{j=1}^{d} \psi\big(\frac{\gamma_{i\tau}(\omega)+1-j}{2}\big) + d\ln 2 - \ln |\mathbf{S}_{i\tau}(\omega)| \end{aligned} \tag{5.4.9}$$

The posterior distribution of the scaling random variables $u_{i\tau}(\omega, t)$ given the indicator variables has the form of their prior, which is a Gamma distribution [31]. Therefore, in addition to $r_{i\tau}(\omega, t)$ the parameters κ and $\eta_{i\tau}(w, t)$ of the Gamma distribution and the expected value of the scaling variables are computed as follows:

$$\kappa = \frac{d+\nu}{2} \tag{5.4.10}$$

$$\begin{aligned} \eta_{i\tau}(w) = \gamma_{i\tau}(w)\Big(\mathbf{y} - \mathbf{m}_{i\tau}(\omega)\Big)^T \mathbf{S}_{i\tau}^{-1}(\omega)\Big(\mathbf{y} - \mathbf{m}_{i\tau}(\omega)\Big) \\ + \frac{d}{2\beta_{i\tau}(w)} + \frac{\nu}{2} \end{aligned} \tag{5.4.11}$$

$$\bar{u}_{i\tau}(\omega, t) = \kappa/\eta_{i\tau}(w) \tag{5.4.12}$$

The following four statistics depending on $r_{i\tau}(\omega, t)$ are defined as

$$\bar{\psi}_{i\tau}(\omega) = \sum_{t} r_{i\tau}(\omega, t) \tag{5.4.13}$$

$$\bar{w}_{i\tau}(\omega) = \sum_{t} r_{i\tau}(\omega, t)\bar{u}_{i\tau}(\omega, t) \tag{5.4.14}$$

$$\bar{\boldsymbol{\mu}}_{i\tau}(\omega) = \frac{1}{\bar{w}_{i\tau}(\omega)} \sum_t r_{i\tau}(\omega, t)\bar{u}_{i\tau}(\omega, t)\mathbf{y} \tag{5.4.15}$$

$$\begin{aligned}\bar{\boldsymbol{\Sigma}}_{i\tau}(\omega) = \frac{1}{\bar{w}_{i\tau}(\omega)} \sum_t \Big[& r_{i\tau}(\omega, t)\bar{u}_{i\tau}(\omega, t) \\ & \big(\mathbf{y} - \bar{\boldsymbol{\mu}}_{i\tau}(\omega)\big)\big(\mathbf{y} - \bar{\boldsymbol{\mu}}_{i\tau}(\omega)\big)^T \Big]\end{aligned} \tag{5.4.16}$$

and are used in the M step.

Within the M-step, $q(\Theta)$ can be determined as follows [31]

$$q(\Theta) \propto p(\Theta) \times exp\big(E_{\mathbf{U},\mathbf{Z}}\{\ln \mathcal{L}_c(\mathbf{Y}, \mathbf{U}, \mathbf{Z})\}\big) \tag{5.4.17}$$

where $p(\Theta) = p(\boldsymbol{\psi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, $\mathcal{L}_c$ is the complete data likelihood and $E_{\mathbf{U},\mathbf{Z}}\{.\}$ is the expectation with respect to the posterior joint distribution of the latent variables. Details of this derivation can also be found in Appendix A.1, whereas the update rules for the hyperparameters of the posterior distributions are given below.

M-step

$$\alpha_{i\tau}(\omega) = \alpha_0 + \bar{\psi}_{i\tau}(\omega) \tag{5.4.18}$$

$$\beta_{i\tau}(\omega) = \beta_0 + \bar{w}_{i\tau}(\omega) \tag{5.4.19}$$

$$\gamma_{i\tau}(\omega) = \gamma_0 + \bar{\psi}_{i\tau}(\omega) \tag{5.4.20}$$

$$\mathbf{m}_{i\tau}(\omega) = \frac{1}{\beta_{i\tau}(\omega)}\Big(\beta_0\mathbf{m}_0(\omega) + \bar{w}_{i\tau}(\omega)\bar{\mu}_{i\tau}(\omega)\Big) \tag{5.4.21}$$

$$\begin{aligned}\mathbf{S}_{i\tau}(\omega) = \mathbf{S}_0 + \bar{w}_{i\tau}(\omega)\bar{\mathbf{\Sigma}}_{i\tau}(\omega) \\ + \frac{\beta_0 \bar{w}_{i\tau}(\omega)}{\beta_{i\tau}(\omega)}(\bar{\boldsymbol{\mu}}_{i\tau}(\omega) - \mathbf{m}_0(\omega))(\bar{\boldsymbol{\mu}}_{i\tau}(\omega) - \mathbf{m}_0(\omega))^T\end{aligned} \tag{5.4.22}$$

where $\alpha_{i\tau}(\omega)$ is the parameter of the updated Dirichlet distibution.

Table 4.1 summarizes the VB EM algorithm. After convergence, the probabilistic mask extracting each source i from the microphone signals $L(\omega, t)$ or $R(\omega, t)$ is determined by summing the expectations of the latent indicators $r_{i\tau}(w, t)$ over the delay τ

$$M_i(\omega, t) \equiv \sum_{\tau} r_{i\tau}(\omega, t) \tag{5.4.23}$$

Table 4.1: VB EM update rules for SMMs

1. E-step:

Compute the expectations of the latent variables $r_{i\tau}(\omega, t)$ and that of the scaling variables $\bar{u}_{i\tau}$, using the parameter estimates of the posterior distributions.

2. M-step:

Using $r_{i\tau}(\omega, t)$ and $\bar{u}_{i\tau}(\omega, t)$, update the parameters of the posterior Dirichlet distribution $\alpha_{i\tau}(\omega)$, and the parameters of the Gaussian-Wishart distribution $\beta_{i\tau}(\omega)$, $\gamma_{i\tau}(\omega)$, $\mathbf{m}_{i\tau}(\omega)$ and $\mathbf{S}_{i\tau}(\omega)$.

5.5 Experimental Evaluation

Four main experiments were performed in order to evaluate the proposed approach and compare it with other algorithms. The first experiment examines the impact of the degree of freedom parameter choice on the separation performance. The second experiment compares the proposed approach with three other underdetermined T-F masking approaches, DUET [24],

MESSL [25] and the modified version of MESSL (MESSL with SMM) proposed in Chapter 3. In the third experiment, the separation performance is evaluated for two scenarios; nearby speech sources and different reverberation times to emphasize the robustness of the VB framework compared to the classical EM employed in MESSL and MESSL with SMM. The fourth experiment examines the performance in comparison with a particular version of MESSL, where an additional source named "garbage" is used to account for the reverberations from different sources independent of their locations. The garbage source acts as an outlier detection method to improve the accuracy of the soft masks by avoiding poorly fitted spectrogram points [25].

5.5.1 Experimental Set-up

Data sources

Speech utterances of 2.5 s long were chosen randomly from the whole TIMIT acoustic-phonetic continuous speech corpus [76]. Mixtures were formed from different combinations of male and female signals sampled at 16 KHz. These signals were normalized to have the same energy and convolved with real BRIRs described in the following section. All experiments included either two or three speakers, one target and either one or two interferers. The target was always positioned facing the microphones and the interferers were located at various azimuthal angles. For three simultaneous speakers, the second interferer was symmetrically positioned.

Binaural Impulse Responses

Four different sets of BRIRs were used in the experiments. The first two sets [78] were measured in a real classroom named Room A, of dimensions 5 m$\times$9 m$\times$3.5 m and a reverberation time (RT60) of 565 ms. The room A set 1 BRIRs corresponds to a separation distance of 1 m between the sources and the center of the sensors and the other set corresponds to a separation

distance of 0.4 m. The other dataset of BRIRs [71] was captured in four different sized rooms and different reverberation times. Only two sets of measurements were used, the first set was recorded in a medium sized office, Room B, of dimensions 5.72 m×6.64 m×2.31 m and an RT60 of 320 ms; whereas, the second set was recorded in a medium size seminar room, Room C, of dimensions 8.02 m×8.72 m×4.25 m and an RT60 of 890 ms. In both rooms, the BRIRs were recorded using a head and torso simulator (HATS) with the speech sources always located at a distance of 1.5 m from the HATS and at different azimuths in the interval $[-90^\circ\ 90^\circ]$ with 5° sampling. The layouts of Room B and Room C are shown in Fig. 5.3 and Fig. 5.4. The different BRIRs are summarized in Table 5.1.

Table 5.1: Binaural real impulse responses

Rooms	RT60	Separation Distance
Room A set 1	565 ms	1 m
Room A set 2	565 ms	0.4 m
Room B	320 ms	1.5 m
Room C	890 ms	1.5 m

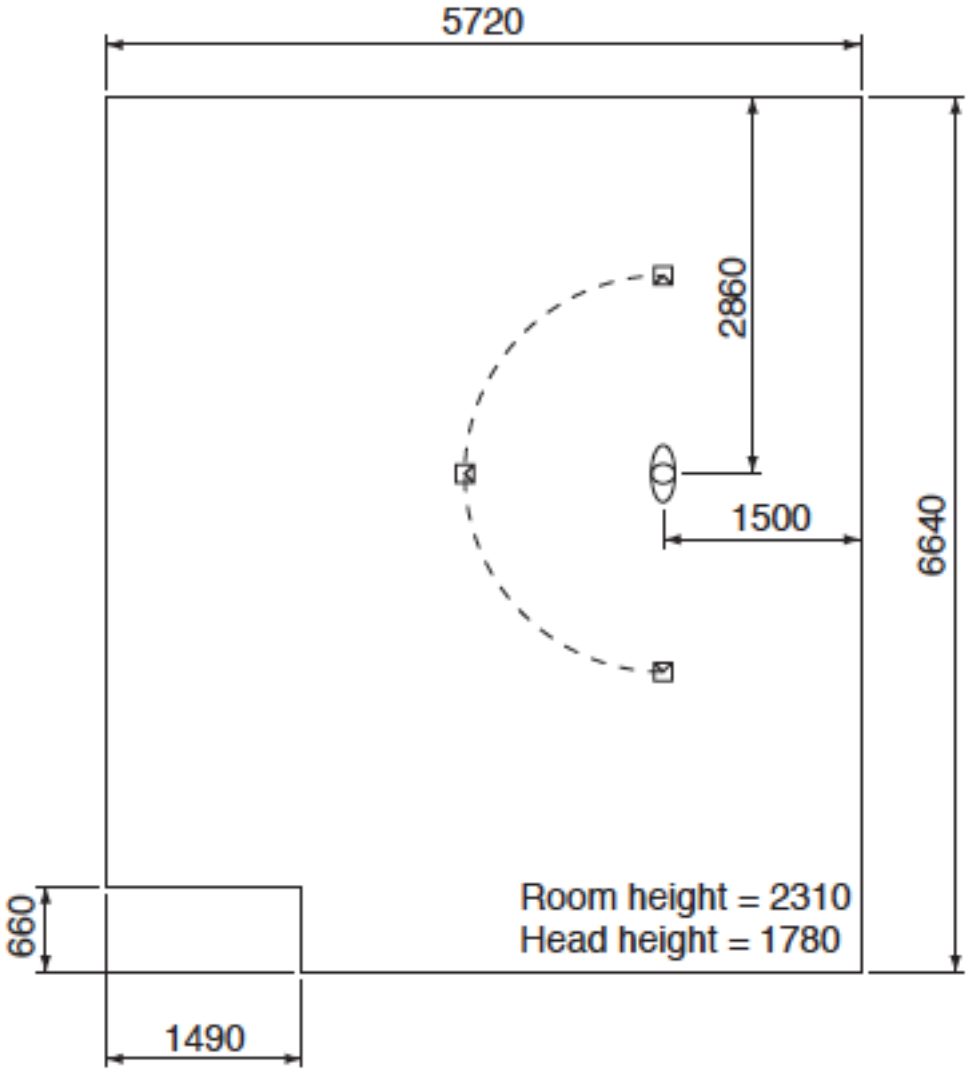


Figure 5.3: Layout of Room B, dimensions 5.72 m×6.64 m×2.31 m, RT60 = 320 ms.

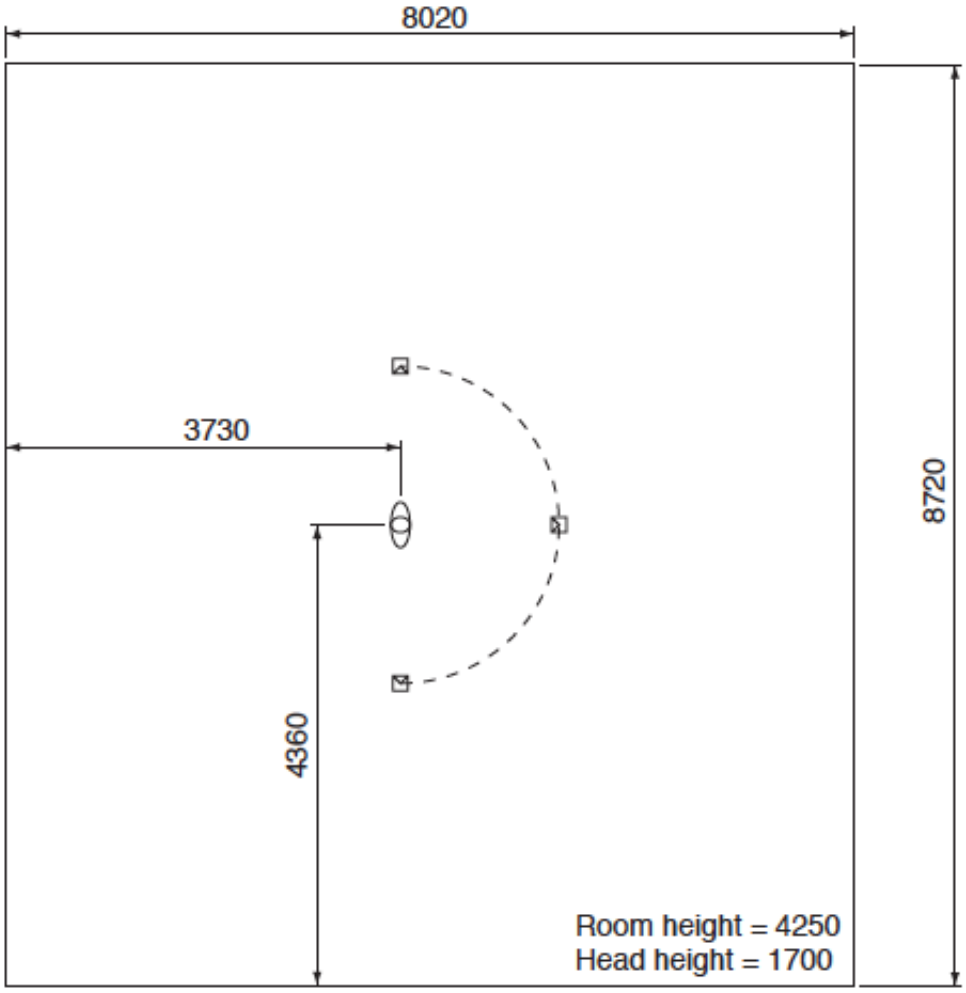


Figure 5.4: Layout of Room C, dimensions 8.02 m×8.72 m×4.25 m, RT60 = 890 ms.

Separation Performance Measures

The separation performance was evaluated objectively by the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [48]. The Matlab function bss_eval_sources.m recommended in the SignalSeparation Evaluation Campaign (SiSEC 2008) [99] was used for this evaluation. The SDR as defined in Chapter 3 is the ratio of the energy of the target signal to the energy of other signals including noise, interferers as well as other artifacts. The SIR is instead the ratio of the target to the interference energy excluding noise and artifacts. In the estimated signal, any energy resulting from the target source or a linear combination of its delayed versions (up to 32ms) is considered as target energy. On the other hand, energy from the masker or its delayed versions represents the interfering energy. Any energy that cannot be explained by any of these is considered artifacts caused mainly by reverberations. The speech quality was also evaluated using the Perceptual Evaluation of Speech Quality (PESQ), which is used in the standard ITU P.862 for assessing the quality of speech transmitted over communication channels [100]. It is highly correlated with subjective perceived quality measured using a mean opinion score (MOS). The MOS is a test undertaken by human listeners to evaluate speech quality, its correlation with PESQ was found greater than 0.90 [100]. PESQ score varies between -0.5 and 4.5, with 4.5 being the best possible quality.

Initialization

The set of hyperparameters was set following [92], where $\beta_0 = 0.01$, this value was experimentally determined, $\gamma_0 = 3$, $\mathbf{m}_0(\omega)$ is set equal to the mean of the data and $\mathbf{S}_0$ is a dxd identity matrix. The delay τ associated with each speech source is estimated using PHAT [25]. The delay τ is assumed to vary in the interval [-15 15] in steps of 0.5 equivalent to [-940 ms: 940 ms] in steps of 30 ms [61], i.e. τ is a grid of 61 elements. Only the values of

τ estimated by PHAT are considered and $\psi_{i\tau}$ is initialized as a frequency independent diagonal matrix with $\psi_{ii} = 1/I$, where I is the total number of speech sources. For the VB EM steps, an iteration number of 8 was used in all the experiments, which was found empirically to be an approximate trade off between convergence and complexity.

5.5.2 Impact of the Degree of Freedom

In this experiment, the impact of varying the degree of freedom on the separation performance was studied for the case of two speakers. Room A set 1 BRIRs were used and four angular positions for the interferer were tested [15°, 30°, 45°, 60°]. SDR and SIR results were averaged over ten different mixtures at each angular position. Smaller values of ν result in heavier tails of the non-Gaussian distribution, while as ν increases the Student's t reduces to a Gaussian distribution. The degree of freedom affects the robustness of the algorithm, for large values of ν it becomes more sensitive to outliers, hence the estimation of the T-F masks is less accurate and the speech separation is negatively affected as illustrated in Fig. 5.5.

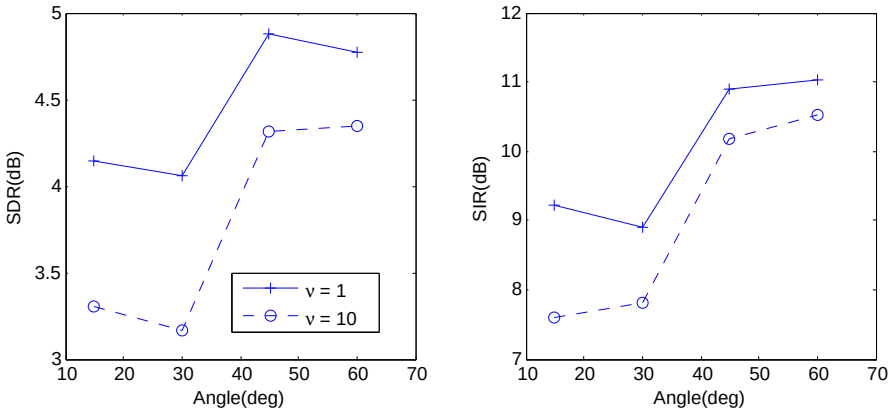


Figure 5.5: SDR and SIR as a function of the separation angles for two speakers. Room A set 1 BRIRs were used. The results were averaged over ten random mixtures at each of the four angles.

The SDR and SIR values are depicted as a function of the separation angles for $\nu = 1$ and $\nu = 10$, it is clear from the graphs that the performance of the algorithm decreases at all separation angles particularly at small separations as ν increases. At $\nu = 10$, SDR and SIR are both reduced by an average of 0.7 dB and 1 dB, respectively.

5.5.3 Comparison with Other Algorithms

This experiment compares the proposed variational Bayesian source separation (VBSS) approach, with three other underdetermined speech separation algorithms, DUET [24], MESSL [25] and MESSL with SMM. A quick overview of each algorithm is presented below, followed by the experimental results.

DUET

Based on the W-disjoint orthogonality, DUET estimates the mixing parameters by constructing a two dimensional weighted histogram of the attenuation-delay pairs resulting from the ratio of the T-F representation of the left and right channels. The number of peaks indicates the number of sources and their locations determine the corresponding mixing coefficients estimates. Accordingly, each spectrogram point is assigned to the peak location that is nearest using the likelihood function as a measure of closeness. Given these peak centers $(\tilde{\alpha}_i, \tilde{\delta}_i)$, $i = 1, .., I$, T-F points are clustered, via $J(\omega, t)$ defined as

$$J(\omega, t) = \frac{|\tilde{a}_i exp^{-j\tilde{\delta}_i \omega} L(\omega, t) - R(\omega, t)|}{1 + \tilde{a}_i^2} \tag{5.5.1}$$

where $\tilde{\delta}_i$ is the estimated relative delay of source i, $\tilde{a}_i$ is the corresponding relative attenuation and $\alpha_i = a_i - \frac{1}{a_i}$ is the symmetrical attenuation used instead of a_i to construct the histogram [24]. As a by product, DUET generates binary masks to extract the sources from their mixtures.

MESSL

In contrast to DUET, MESSL generates probabilistic masks by independently modelling the interaural phase and level differences with Gaussian distributions. Spectrogram points are clustered based on their interaural parameters which are determined through likelihood maximization via the EM algorithm applied to Gaussian mixture models. The total log likelihood maximized in MESSL as explained previously is expressed as follows [25]

$$\mathcal{L}_{GMM}(\Theta) = \sum_{\omega,t} \log \sum_{i,\tau} \Big[\Big(\mathcal{N}(\hat{\phi}(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega) \Big) \\ .\mathcal{N}\Big(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega) \Big) .\psi_{i\tau} \Big] \tag{5.5.2}$$

where $\Theta_{\Omega\Omega} \equiv \{\xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\}$ denotes the set of model parameters. $\xi_{i\tau}(\omega)$ and $\sigma^2_{i\tau}(\omega)$ are the means and the variances of IPD respectively and $\mu_i(\omega)$ and $\eta_i^2(\omega)$ correspond to the means and variances of the ILD. For localization, all possible values of τ are used in the EM framework and the initial value of $\psi_{i\tau}$ is approximated by a Gaussian distribution with its mean at each cross-correlation peak and a standard deviation of one sample. Hence, $\boldsymbol{\psi}$ is a matrix of dimension Ix61. This assumption increases the dimensionality of the latent space as well as the total computational complexity which is proportional to the number of sources, number of the discrete values of τ, number of spectrogram points and the number of iterations [25]. In contrast to EM, the variational Bayesian approach inherently assumes that the mixing coefficients $\psi_{i\tau}$ are random variables with a Dirichlet prior, only estimates of τ corresponding to the cross correlation peaks (equal to the number of sources) are considered in the VB EM update rules which reduces the computational complexity.

MESSL with SMM

In [101], the GMM used in MESSL was replaced by an SMM and both interaural cues were independently clustered by maximization of the log likelihood expressed as

$$\mathcal{L}_{SMM}(\Theta) = \sum_{\omega,t} \log \sum_{i,\tau} \Big[\Big(\mathcal{S}t(\hat{\phi}(\omega,t;\tau)|\mu_{pi\tau}(\omega), \lambda_{pi\tau}(\omega) \Big) .\mathcal{S}t\Big(\alpha(\omega,t)|\mu_{li}(\omega), \lambda_{pi\tau}(\omega) \Big).\psi_{i\tau} \Big] \tag{5.5.3}$$

where $\mu_{pi\tau}(\omega)$, $\lambda_{pi\tau}$ are the means and precisions of IPD and $\mu_{li}(\omega)$, $\lambda_{li\tau}(\omega)$ are the the means and precisions of ILD respectively. The degree of freedom was assumed fixed a priori and equal to unity. Similar to MESSL, the maximum likelihood estimation of the SMM parameters is also obtained in two steps. The main difference between the EM algorithm of the SMM and that applied to the GMM is the estimation of the additional phase and level scaling expectations [101].

Comparison results

Fifteen different mixtures were generated and Room A set 1 BRIRs were used for this comparison. The average SDR and SIR results comparing the VBSS algorithm with the other three algorithms are depicted in Figure 5.6 for two speakers and Figure 5.7 for three speakers. The performance of all the algorithms decreases for small azimuthal separations. A decrease of 2.3 dB and 2.1 dB can be seen in the average SDR between the largest and the smallest angles in DUET and MESSL, respectively. However, for the algorithms employing SMM the performance is improved substantially at small angular separation and this difference reaches 1.1 dB for MESSL with SMM and only 0.9 dB for the proposed approach. The proposed approach, improves both SDR and SIR results at all angles. For two speakers, the average SDR improvement over the four azimuthal angles obtained using

the proposed approach compared to DUET is 2.5 dB. For three speakers, this improvement is 2.8 dB. Compared to MESSL, the improvement is 1.4 dB for the two speakers case and 1.6 dB for the case of three speakers. These improvements decrease to 0.7 dB and 0.8 dB compared to MESSL with SMM for the cases of two and three speakers respectively. For the case of two speakers, the SIR results confirm the advantage of the VBSS algorithm with an average improvement (over the four azimuthal separations) of 4.2 dB, 3.3 dB and 1.8 dB relative to DUET, MESSL and MESSL with SMM, respectively. Adding a second interferer reduces the improvements to 3 dB, 2.6 dB, 1.4 dB compared to the three algorithms.

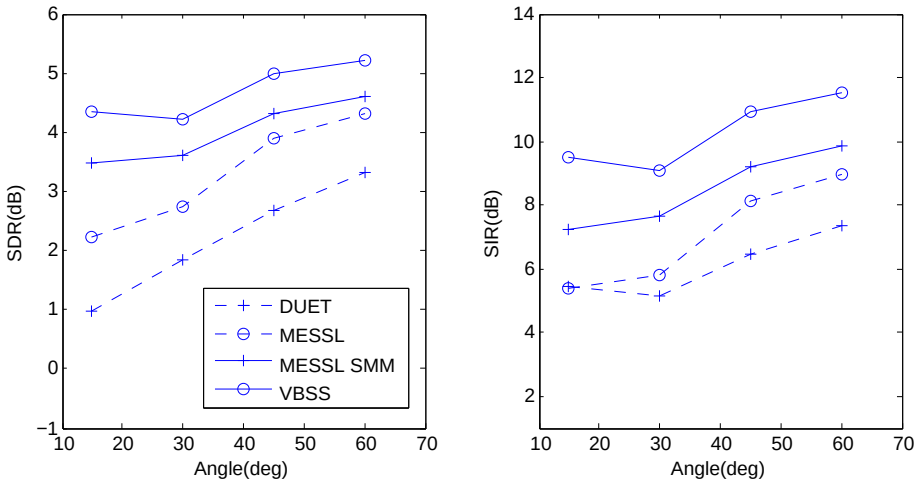


Figure 5.6: SDR and SIR as a function of the azimuthal separation for two speakers. Room A set 1 BRIRs was used. The results were averaged over 15 random mixtures at each of the four angles.

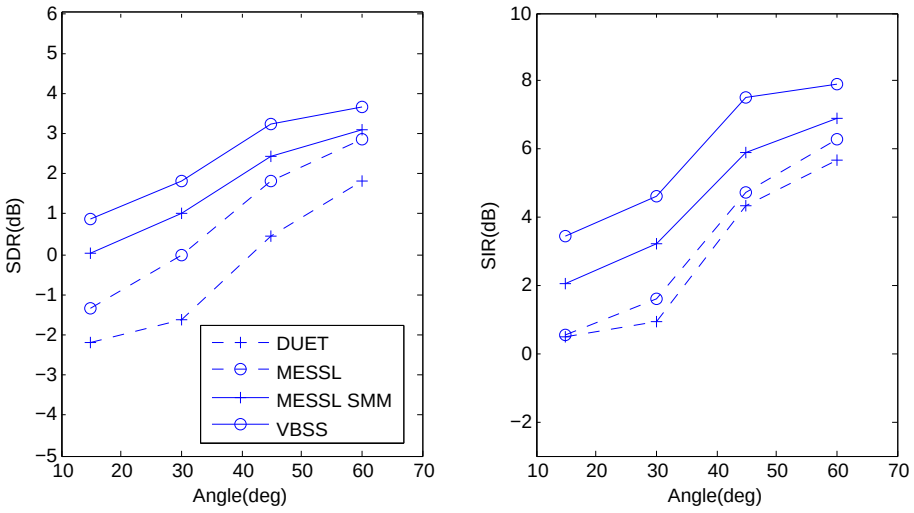


Figure 5.7: SDR and SIR as a function of the azimuthal separation for three speakers. Room A set 1 BRIRs was used. The results were averaged over 15 random mixtures at each of the four angles.

5.5.4 Sources in Close Proximity and Different Reverberation Times

Nearby sources

Although a typical cocktail party problem usually involves sources relatively near the listener, previous approaches such as DUET [24] and MESSL [25] were experimentally tested for a separation distance of a meter or more between the speech sources and the microphones/ears. This experiment focuses on nearby sources (separation distance < 1 m). Unlike the case of distant sources, small variations in the speaker location relative to the listener largely affect the direct-sound energy reaching the microphones which maximizes the interaction between the azimuthal variation of the source location and the effects of reverberation [78]. In [25], the average SDR and PESQ improvement obtained by MESSL in comparison with other separa-

tion algorithms were 1.6 dB and 0.27 MOS units, respectively.

Therefore, the first part of this experiment focuses on comparing the performance of the VBSS approach with only the state-of-the-art MESSL and MESSL with SMM, to demonstrate the gradual improvement in the quality of speech separation obtained through enhancing the statistical framework. The robust probabilistic modelling of the interaural cues via SMM [101] improves partly the separation quality, followed by an additional improvement due to the VB clustering framework as shown in Table 5.2, for the case of two speakers and Table 5.3, for the case of three speakers. Room A set 2 BRIRs corresponding to a distance of 0.4 m between speech sources and the listener and a reverberation time of 565 ms, were used in this experiment and the results were averaged over 15 different speech mixtures.

For the two-speaker case, the average SDR improvements obtained using the proposed approach compared to MESSL and MESSL with SMM are 2.8 dB and 1.3 dB, respectively. The average SIR improvements are 4.9 dB and 2.9 dB, respectively. The PESQ results follow the objective measures and the VBSS algorithm performs better than MESSL by 0.24 MOS units and better than MESSL with SMM by 0.11 MOS units. For the three speakers case, the SDR and SIR improvements compared to MESSL are 3 dB and 4.7 dB respectively and the PESQ scores are higher than MESSL by an average of 0.21 MOS units over the four angles respectively. Compared to MESSL with SMM, the SDR, SIR and PESQ improvements are 1.4 dB, 2.7 dB and 0.11 MOS, respectively.

Table 5.2: Comparison between MESSL, MESSL with SMM and VBSS in terms of average SDR, SIR and PESQ for two speakers, Room A set 2

SDR in dB				
Azimuth angles	15°	30°	45°	60°
MESSL	4.77	5.58	8.76	7.86
MESSL with SMM	6.17	7.76	9.97	9.04
VBSS	8.93	8.86	10.63	9.70
SIR in dB				
Azimuth angles	15°	30°	45°	60°
MESSL	6.87	6.84	11.44	9.81
MESSL with SMM	7.55	9.73	13.76	11.88
VBSS	12.71	12.28	15.90	13.70
PESQ in MOS units				
Azimuth angles	15°	30°	45°	60°
MESSL	1.94	2.13	2.20	2.21
MESSL with SMM	2.06	2.32	2.32	2.29
VBSS	2.27	2.42	2.42	2.34

Table 5.3: Comparison between MESSL, MESSL with SMM and VBSS in terms of average SDR, SIR and PESQ for three-speakers, Room A set 2

SDR in dB				
Azimuth angles	15°	30°	45°	60°
MESSL	1.37	3.04	6.21	5.55
MESSL with SMM	3.13	5.41	7.49	6.37
VBSS	5.35	6.80	8.63	7.21
SIR in dB				
Azimuth angles	15°	30°	45°	60°
MESSL	2.20	3.82	8.03	7.12
MESSL with SMM	3.92	6.89	10.26	8.04
VBSS	7.30	9.29	12.96	10.23
PESQ in MOS units				
Azimuth angles	15°	30°	45°	60°
MESSL	1.63	1.89	1.96	1.96
MESSL with SMM	1.78	2.05	2.03	2.01
VBSS	1.94	2.14	2.13	2.07

Different RT60s

As explained in [71] the reverberation structure consists of early reflections and dense late reverberations. The early reflections generally affect speech positively by amplifying it. On the other hand, the late reverberation reflections are poorly correlated with the speech and act as additive noise (outliers). In reverberant conditions, the distribution of interaural cues are broadened which deteriorates the speech intelligibility and the human ability to use these cues for the separation of multiple speech sources [72]. In [102], the binaural listening was investigated by measuring the intelligibility of speech against its spatially separated speech masker. An intelligibility gain of 4 to 5 dB was measured for anechoic conditions, this gain was decreased by 2 to 3 dB for an RT60 of 400 ms. Similarly, in the case of concurrent speech sources, the separation performance degrades in reverberant environments. In [25], when MESSL was tested to separate two and three speech sources, the SDR was reduced by 5 to 6 dB in an RT60 of 565 ms compared to the anechoic conditions.

In this experiment, the BRIRs of Room B and Room C were used to investigate the performance of the VBSS algorithm in comparison to MESSL and MESSL with SMM. For these BRIRs, the separation distance is 1.5 m. Since sources in close proximity are more common in a typical CPP environment, the azimuthal separation between the target and the interferer was chosen in the interval [10° 40°]. The SDR results averaged over 15 different speech mixtures at each of the azimuthal separation, are shown in Figure 5.8 for two speakers and Figure 5.9 for three speakers.

It is clear from the graphs that SDR values decrease with the increase of the reverberation time. For separating two speakers, the average SDR over the four separation angles in Room B scored by MESSL was 6.9 dB and decreased to 2.7 dB for Room C. Similarly, for the case of three speakers the average SDR decreased from 5 dB in Room B to 1.8 dB in Room C.

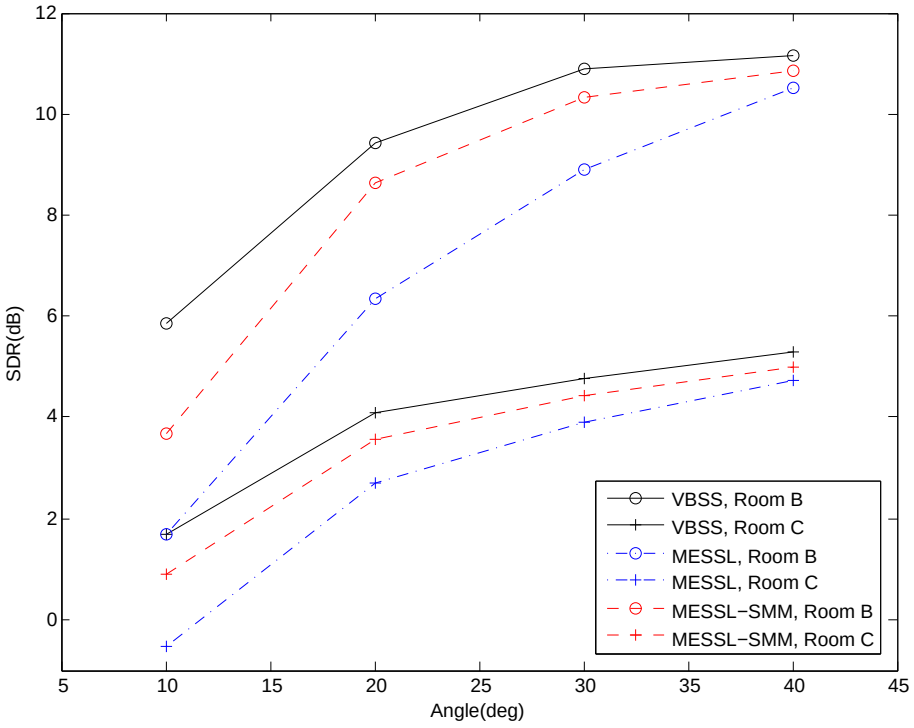


Figure 5.8: SDR as a function of separation angles for the case of two speakers. Room B BRIRs and Room C BRIRs were used. The results were averaged over 15 random mixtures at each of the four angles.

The proposed approach performs better than MESSL and MESSL with SMM at all separation angles especially for smaller values of physical separation at which the similarity between spatial interaural cues is higher. Compared to MESSL, average improvements of 2.4 dB and 1.3 dB were obtained in Room B and Room C, respectively, for the case of two speakers. These improvements were 1.7 dB and 1.4 dB for Room B and Room C in the case of three speakers. Average values of SIR confirmed the separation performance improvement as shown in Table 5.4 and Table 5.5. For the case of two speakers, average improvements of 4.8 dB and 4.1 dB are obtained for Room B and Room C, respectively. For three speakers, these improvements were reduced, respectively, to 3 dB and 2.6 dB. Compared to MESSL with SMM, average SDR improvements of 0.95 dB and 0.5 dB were obtained

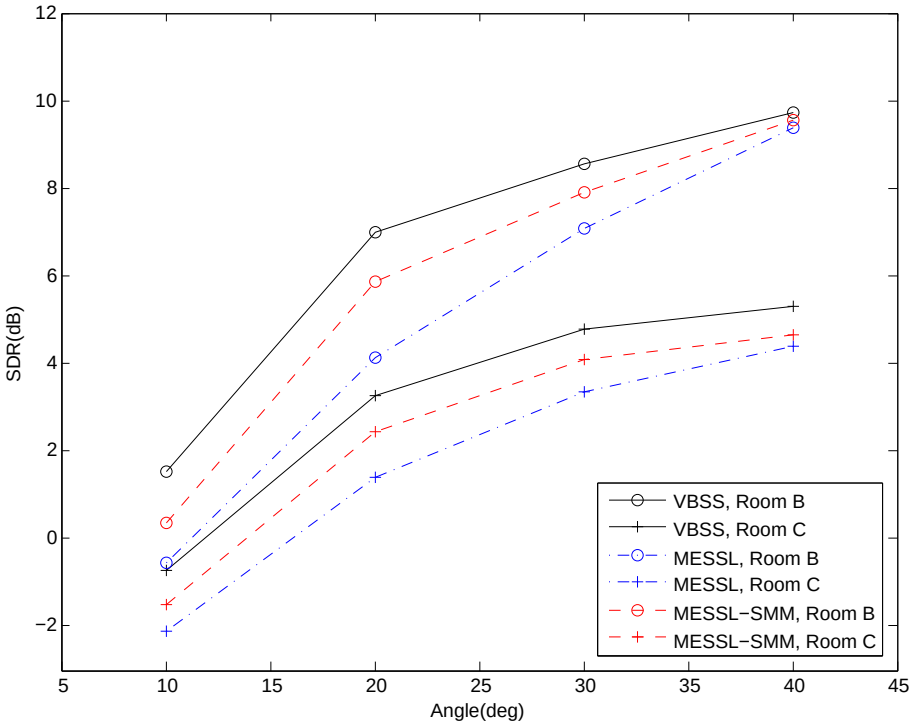


Figure 5.9: SDR as a function of separation angles for the case of three speakers. Room B BRIRs and Room C BRIRs were used. The results were averaged over 15 random mixtures at each of the four angles.

in Room B and Room C, respectively, for the case of two speakers. These improvements were 0.77 dB and 0.73 dB for Room B and Room C in the case of three speakers. Similarly, the SIR average values for the case of two speakers, increased by 2.6 dB and 2.3 dB for Room B and Room C, respectively. For three speakers, these improvements were reduced to 1.6 dB and 1.4 dB. PESQ scores were closely related to the SDR results and are shown in Table 5.6 and Table 5.7.

It can be seen in Figure 5.9, that for the case of three speakers and at the smallest azimuthal separation (10°) all algorithms perform poorly. This can be explained partly by the similarity of spatial cues but it was also observed that initialization via PHAT failed totally due to high reverberations. Integrating video modality in [57] has shown to improve speech separation

compared to audio-only based methods. Assuming video information is used for initialization instead of PHAT, the average SDR over 15 mixtures at 10° for VBSS improved by 2.2 dB.

Table 5.4: SIR for two and three speakers in Room B

Two speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	2.77	7.66	11.65	14.70
MESSL with SMM	4.96	10.94	14.57	4.94
VBSS	8.57	13.79	16.60	16.91
Three speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	-0.082	5.28	8.96	12.07
MESSL with SMM	0.79	7.63	10.19	13.02
VBSS	2.76	9.83	11.73	13.85

Table 5.5: SIR for two and three speakers in Room C

Two speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	3.31	9.15	11.53	15.19
MESSL with SMM	5.04	11.162	13.95	16.56
VBSS	7.87	13.81	15.70	18.37
Three speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	0.1	4.76	8.41	12.11
MESSL with SMM	0.78	6.62	10.15	12.58
VBSS	2.12	8.14	11.74	13.71

Table 5.6: PESQ in MOS units for two and three speakers in Room B

Two speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	1.80	2.21	2.40	2.51
MESSL with SMM	1.92	2.38	2.57	2.57
VBSS	2.03	2.42	2.68	2.63
Three speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	1.58	1.94	2.14	2.32
MESSL with SMM	1.66	2.07	2.22	2.33
VBSS	1.72	2.15	2.31	2.39

Table 5.7: PESQ in MOS units for two and three speakers in Room C

Two speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	1.78	2.15	2.33	2.42
MESSL with SMM	1.88	2.23	2.42	2.51
VBSS	1.96	2.23	2.43	2.54
Three speakers				
Azimuth angles	10°	20°	30°	40°
MESSL	1.55	1.89	2.13	2.11
MESSL with SMM	1.57	1.94	2.16	2.20
VBSS	1.55	1.95	2.19	2.19

5.5.5 MESSL with garbage source

In the previous experiments, the proposed approach was compared with other T-F CASA based algorithms, relying on clustering the same binaural cues and differing solely in either the probabilistic modelling of these cues

or the clustering framework or both.

In order to minimize the impact of reverberations, a "garbage" source was proposed in MESSL to account for spectrogram points which are not fitted by other source models. This added source allows a better estimation of the actual source parameters by avoiding atypical points. As shown in Table 5.8 and Table 5.9, the proposed approach through the non-Gaussian modelling improves the robustness against outliers and outperforms MESSL with the garbage source (MESSLG), without the need for any additional dereverberation method. Room A set 2 BRIRs were used in this experiment and the results were averaged over 15 different speech mixtures. For the two-speaker case, the average SDR, SIR and PESQ improvements obtained using the proposed approach compared to MESSLG are 1.7 dB, 3.6 dB and 0.17 MOS units, respectively. For three speakers, these improvements compared to MESSLG are 3 dB, 4.7 dB and 0.2 MOS units, respectively.

Table 5.8: Comparison between MESSLG and VBSS in terms of average SDR, SIR and PESQ for two speakers, Room A set 2

SDR in dB				
Azimuth angles	15°	30°	45°	60°
MESSLG	5.12	7.93	9.48	8.87
VBSS	8.93	8.86	10.63	9.70
SIR in dB				
Azimuth angles	15°	30°	45°	60°
MESSLG	6.43	9.95	12.35	11.36
VBSS	12.71	12.28	15.90	13.70
PESQ in MOS units				
Azimuth angles	15°	30°	45°	60°
MESSLG	1.97	2.25	2.26	2.28
VBSS	2.27	2.42	2.42	2.34

Table 5.9: Comparison between MESSLG and VBSS in terms of average SDR, SIR and PESQ for three speakers, Room A set 2

SDR in dB				
Azimuth angles	15°	30°	45°	60°
MESSLG	1.69	4.58	7.71	6.79
VBSS	5.35	6.80	8.63	7.21
SIR in dB				
Azimuth angles	15°	30°	45°	60°
MESSLG	2.20	3.82	8.03	7.12
VBSS	7.30	9.29	12.96	10.23
PESQ in MOS units				
Azimuth angles	15°	30°	45°	60°
MESSLG	1.67	1.91	1.95	1.96
VBSS	1.94	2.14	2.13	2.07

5.5.6 Computational Complexity

There is a close similarity between the variational Bayesian solution and the EM algorithm for maximum likelihood [67]. The dominant computational cost of the variational algorithm and the conventional EM results from the evaluation of the expected values of the latent variables, together with the evaluation and inversion of the data covariance matrices. The computational overhead in using this approach as compared to the traditional MLE is compensated by many advantages. In addition to the absence of singularities which occur in maximum likelihood whenever a Gaussian component collapses onto a specific data point, the Bayesian treatment avoids the use of other techniques associated with the EM framework such as bootstrapping [31] [103]. Bootstrapping was used in MESSL [25] to avoid source permutations and other local maxima and to ensure consistency of parameters estimation across frequency. The average running time estimated to separate two 2.5 s speech sources from their convolutive mixtures, in a real room reverberant environment (Room A set 1), using VBSS on a 2.6 GHz Intel Core i7 is approximately 16 s. Under the same conditions, the running time required to separate 3 sources reaches 32 s.

5.6 Summary

In this chapter, a general probabilistic approach for T-F masking speech separation was proposed. Non-Gaussian modelling was integrated into a variational Bayesian framework for the joint clustering of IPD and ILD cues. This approach overcomes the shortcomings of the traditional EM algorithm for GMMs as it avoids the probable unbounded behaviour and the convergence problems associated with the likelihood maximization. The robust clustering resulting from employing the heavy tailed Student's t-distribution for modelling interaural cues has improved the estimation of the soft proba-

bilistic masks at various reverberation times particularly for nearby sources. Additionally, the joint modelling of the interaural cues inherently considers their dependence avoiding thus unnecessary assumptions or additional effort to model their correlation. Comparative studies of the proposed approach with other T-F masking algorithms under different scenarios have confirmed a significant improvement in terms of objective and subjective performance measures. Compared to the state-of-art MESSL algorithm, the average SDR improvement over these scenarios (different separation distances and reverberation times) is 1.98 dB, which is equivalent to an improvement of 40% , for the case of two speakers. For the case of three speakers, the VBSS algorithm produces SDRs which are 1.87 dB (62%) higher than those obtained when applying MESSL. The PESQ average results confirm as well these improvements with an average increase of 0.17 MOS for the case of two sources and 0.12 MOS for the case of three sources. Conclusions and suggestions for future research are finally presented in the following chapter.

Chapter 6

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

This study has provided a substantial improvement in the statistical framework used to achieve underdetermined T-F masking blind separation of speech. The proposed framework is a step towards the creation of a machine or a computer system capable of solving the cocktail party problem in realistic reverberant conditions. The contributions can be summarized as follows:

- A novel approach to the probabilistic modelling of the spatial interaural cues used in T-F masking speech separation algorithms.
- Exploiting variational Bayesian inference clustering as an alternative to the traditional EM algorithm to avoid the limitations of maximum likehood optimization.
- Integration of non-Gaussian modelling into a variational Bayesian clustering framework to improve the separation performance for sources in close proximity.
- Multivariate modelling of interaural cues to avoid unnecessary assumptions of independence which as well improves the robustness in mod-

elling uncertainties.

In the first contribution, non-Gaussian modelling based on the Student's t-distribution was proposed as an alternative to the Gaussian distribution used to model the interaural cues in CASA based T-F masking algorithms. Gaussian mixture models commonly used for analytical tractability are known to be very sensitive to outliers. Their tails are often lighter than required which affects the estimates of the means and variances of the components and hence the estimation of the probabilistic T-F masks. The Student's t-distribution, on the other hand, whose heavy tails better reflect outlier values, provides a generalization to the Gaussian distribution. This approach was applied to the state-of the-art MESSL algorithm and has proven to improve the robustness in reverberant environments without the need for any reverberation detection method. Using the EM algorithm as a clustering framework and frequency dependent SMMs for fitting interaural cues has significantly improved the speech separation, an average SDR improvement of 1.3 dB, which is equivalent to 57% improvement, was obtained for clustering both interaural phase and level differences, compared to the Gaussian modelling employed in MESSL.

In the second contribution, a variational Bayesian framework was proposed as an alternative to the EM algorithm and was used for clustering spectrogram points depending only on their IPD cues. This approach avoids the drawbacks of the traditional EM algorithm for GMMs, particularly the probable presence of singularities associated with the likelihood optimization, without requiring additional extensive computations. More importantly, experimental results have also shown an improvement in the quality of speech separation. For the case of two speakers, an average SDR improvement of 0.8 dB (28%) was obtained relative to the EM clustering algorithm employed in MESSL. The average SDR improvement increased to 1.2 dB in the case of three sources, which is equivalent to an increase of 225%.

The major contributions of this thesis are the multivariate modelling of the interaural phase and level differences and the integration of their non-Gaussian modelling within a variational Bayesian framework. This approach combines the advantages of the robust estimation provided by the Student's t-distribution and the robust clustering inherent in the Bayesian approach when modelling uncertainties. In other words, this general approach avoids the probable unbounded behaviour and the convergence problems of the likelihood maximization and makes use of the heavy tailed Student's t-distribution for modelling interaural cues to improve the estimation of the soft probabilistic masks at various reverberation times particularly for sources in close proximity. Additionally, the joint modelling of the interaural cues inherently considers their dependence thus avoiding unnecessary assumptions or additional effort to model their correlation. Comparative studies of the proposed approach with other T-F masking algorithms under different scenarios have confirmed a significant improvement in terms of objective and subjective performance measures. Compared to the state-of-art MESSL algorithm, the average SDR improvement over these scenarios (different separation distances and reverberation times) is 1.98 dB, which is equivalent to an improvement of 40%, for the case of two speakers. For the case of three speakers, the VBSS algorithm produces SDRs which are 1.87 dB (62%) higher than those obtained when applying MESSL. The PESQ average results also confirm these improvements with an average increase of 0.17 MOS for the case of two speakers and 0.12 MOS for the case of three speakers.

6.2 Future research

Based on this work, different directions can be exploited in the future. The proposed framework achieves T-F masking speech separation from a batch of

stereo mixtures. However, a realistic system imitating the human capability should be able to perform real time blind source separation. An adaptive online version of the proposed algorithm would allow real time separation and would also be applicable for time varying mixing conditions. Since humans exploit audio and visual cues to solve the cocktail party problem, integrating the VBSS in a multimodal system where video information is available would surely further improve the speech separation. For instance, video cues can be used for the source localization required to initialize the separation algorithm thereby improving the performance quality especially when audio based localization algorithms fail to properly localize nearby speech sources in high reverberant environments. Finally, only binaural cues were used in the proposed framework, other monaural cues such as harmonicity and amplitude modulation might be exploited and combined for a better speech separation.

Appendix A

A.1 Maximum Likelihood and Expectation Maximization

Assuming a data set of independent observations $\{y_1, \ldots, y_N\}$ and using GMM for independent and identically distributed random variables, the log likelihood function can be expressed by [67]

$$\ln p(\mathbf{y}|\psi, \mu, \sigma^2) = \sum_{n=1}^{N} \ln \Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(y_n|\mu_k, \sigma_k^2) \Big) \tag{A.1.1}$$

where $\psi = \{\psi_k\}$ is the set of mixing coefficients μ_k and σ_k^2 denote the component mean and variance, respectively.

Maximizing this log likelihood function is more complex compared to the case of a single Gaussian, due to the summation over k. Setting the derivatives of the log likelihood to zero in order to estimate the parameters will no longer result in closed form solutions. An alternative approach of solving this problem is the EM algorithm [67].

A.1.1 Expectation Maximization (EM) for GMM

Setting the derivatives of the log likelihood in (A.1.1) with respect to the means μ_k of the Gaussian components to zero, gives

$$0 = \sum_{n=1}^{N} \frac{\psi_k \mathcal{N}(y_n|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_n|\mu_j, \sigma_j^2)} \sigma_k^2 (y_n - \mu_k) \tag{A.1.2}$$

The posterior probability ν_{nk} can be seen on the right-hand side and is expressed as

$$\nu_{nk} = \frac{\pi_k \mathcal{N}(y_n|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_n|\mu_j, \sigma_j^2)} \tag{A.1.3}$$

Substituting in equation (A.1.2) and multiplying by the inverse of the variance yields to the update equation of the component mean

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \nu_{nk} y_n \tag{A.1.4}$$

where

$$N_k = \sum_{n=1}^{N} \nu_{nk} \tag{A.1.5}$$

Similarly, by setting the derivatives of $\ln p(\mathbf{y}|\pi, \mu, \sigma)$ with respect to σ_k to zero, the update equation of the component variance is given by

$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^{N} \nu_{nk} (y_n - \mu_k)^2 \tag{A.1.6}$$

In order to obtain the update equation for the mixing coefficients psi_k, the loglikelihood function should be maximized under the following constraint $\sum_{k=1}^{K} \psi_k = 1$. This can be done using Lagrange multiplier as follows

$$\ln p(\mathbf{y}|\pi, \mu, \sigma) + \lambda \Big(\sum_{k=1}^{K} \pi_k - 1 \Big) \tag{A.1.7}$$

which yields to

$$0 = \sum_{n=1}^{N} \frac{\mathcal{N}(y_n|\mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_n|\mu_j, \sigma_j)} + \lambda \tag{A.1.8}$$

Multiplying both sides by ψ_k and summing over K results in $\lambda = -N$ and

the update equation of the mixing coefficients is given by

$$\psi_k = \frac{N_k}{N} \tag{A.1.9}$$

The equations (A.1.4), (A.1.6) and (A.1.9) are not closed form solutions for the parameters of the models since the responsibilities ν_{nk} required to estimate these parameters actually depend on them through equation (A.1.3). The EM iterative scheme is therefore an appropriate approach, in which some initial values for the means, variances and mixing coefficients are chosen, followed by alternating between the E step and the M step. In the E step, the current values are used to estimate the posteriori probabilities or responsibilities, followed by the M step, where these responsibilities are used to re-estimate the means, variances and mixing coefficients.

Appendix B

B.1 VB EM update rules

Maximizing the lower bound of the log evidence with respect to $q(\mathbf{U}, \mathbf{Z})$ and $q(\Theta)$ under the variational main assumption expressed in (20) results in the following expectation and maximization steps at each frequency [31]:

VB E-step:

$$q(\mathbf{u}(\omega,t), \mathbf{z}(\omega,t)) \propto exp\big(E_{\Theta}\{\ln p(\mathbf{y}, \mathbf{u}(\omega,t), \mathbf{z}(\omega,t)|\Theta\}\big) \tag{B.1.1}$$

where $E_{\Theta}\{.\}$ denotes the expectation taken with respect to the posterior distribution $q(\Theta)$.

VB M-step:

$$q(\Theta) \propto p(\Theta) \times exp\big(E_{\mathbf{U},\mathbf{Z}}\{\ln \mathcal{L}_c(\mathbf{Y}, \mathbf{U}, \mathbf{Z})\}\big) \tag{B.1.2}$$

where $p(\Theta) = p(\boldsymbol{\psi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, $\mathcal{L}_c$ is the complete data likelihood and $E_{\mathbf{U},\mathbf{Z}}\{.\}$ is the expectation with respect to the posterior joint distribution of the latent variables.

B.1.1 VB E-step

Since the priors on the parameters are chosen conjugate to the likelihood terms, the variational posteriors have the same functional form as their pri-

ors, therefore

$$q(\Theta) = Dir(\boldsymbol{\psi}|\boldsymbol{\alpha}) \prod_{i\tau} \mathcal{N}(\boldsymbol{\mu}_{i\tau}|\mathbf{m}_{i\tau}, \beta_{i\tau}\boldsymbol{\Lambda}_{i\tau}) \mathcal{W}(\boldsymbol{\Lambda}_{i\tau}|\mathbf{S}_{i\tau}, \gamma_{i\tau}) \tag{B.1.3}$$

Taking the expectation with respect to the posterior distribution of the parameters expressed in (44) leads to

$$\begin{aligned} &E_\Theta\big\{ \ln p(\mathbf{y}, \mathbf{u}(\omega,t), \mathbf{z}(\omega,t)\big\} = \sum_{i\tau} z_{i\tau}(\omega,t) \\ &\times \Big\{ \ln \tilde{\psi}_{i\tau}(\omega) - \frac{d}{2}\ln 2\pi + \frac{d}{2}\ln u_{i\tau}(\omega,t) + \frac{1}{2}\ln \tilde{\Lambda}_{i\tau}(\omega) \\ &- \frac{u_{i\tau}(\omega,t)\gamma_{i\tau}}{2}\Big(\mathbf{y} - \mathbf{m}_{i\tau}(\omega)\Big)^T \mathbf{S}_{i\tau}^{-1}(\omega)\Big(\mathbf{y} - \mathbf{m}_{i\tau}(\omega)\Big) \\ &- \frac{u_{i\tau}(\omega,t)d}{2\beta_{i\tau}(\omega)} + \frac{\nu}{2}\ln\frac{\nu}{2} - \ln\Gamma(\frac{\nu}{2}) \\ &+ \Big(\frac{\nu}{2} - 1\Big)\ln u_{i\tau}(\omega,t) - \frac{\nu}{2}u_{i\tau}(\omega,t)\Big\} \end{aligned} \tag{B.1.4}$$

The two special quantities $\ln \tilde{\psi}_{i\tau}(\omega)$ and $\ln \tilde{\Lambda}_{i\tau}(w)$ are defined in equations (23) and (24), respectively. By substituting (45) into (42) and integrating out the scale variables, the quantities $\rho_{i\tau}(\omega,t)$ equivalent to $q(z_{i\tau}(\omega,t)=1)$ are estimated in (22).

Additionally, substituting (45) into (42) results in the posterior distribution of the scale variables

$$q\big(\mathbf{u}_{i\tau}(\omega,t)|\mathbf{z}_{i\tau}(\omega,t)=1)\big) = \mathcal{G}\big(u_{i\tau}(\omega,t)|\kappa, \eta_{i\tau}(\omega)\big) \tag{B.1.5}$$

and leads to the parameter estimates of the scale variables defined in equations (25) to (27).

B.1.2 VB M-step

$$
\begin{aligned}
&E_{\mathbf{U},\mathbf{Z}}\{\ln \mathcal{L}_c(\Theta|\mathbf{Y},\mathbf{U},\mathbf{Z})\} = \sum_t \sum_{i\tau} r_{i\tau}(\omega,t) \\
&\times \Big\{ \ln \psi_{i\tau}(\omega) - \frac{d}{2}\ln 2\pi + \frac{d}{2}\ln(\tilde{u}_{i\tau}(\omega,t)) + \frac{1}{2}\ln|\Lambda_{i\tau}(\omega)| \\
&- \frac{\bar{u}_{i\tau}(\omega,t))}{2}\big(\mathbf{y}-\boldsymbol{\mu}_{i\tau}(\omega)\big)^T \boldsymbol{\Lambda}_{i\tau}(\omega)\big(\mathbf{y}-\boldsymbol{\mu}_{i\tau}(\omega)\big) \\
&+ \frac{\nu}{2}\ln\frac{\nu}{2} - \ln\Gamma\big(\frac{\nu}{2}\big) + \big(\frac{\nu}{2}-1\big)\ln\tilde{u}_{i\tau}(\omega,t) - \frac{\nu}{2}\bar{u}_{i\tau}(\omega,t)\Big\}
\end{aligned}
\tag{B.1.6}
$$

The special quantities $\bar{u}_{i\tau}(\omega,t) = E_{\mathbf{U}}\big\{u_{i\tau}(\omega,t)\big\}$ and $\ln\tilde{u}_{i\tau}(\omega,t) = E_{\mathbf{U}}\ln\big\{u_{i\tau}(\omega,t)\big\}$ can be found using the properties of the Gamma distribution [67]. Substituting (47) into (43) leads to the update equations (32) to (36) used to estimate the hyperparameters of the posterior distributions.

References

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

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, 2005.

[3] D. J. Cook, J. C. Augusto, and V. R. Jakkula, "Ambient intelligence: Technologies, applications, and opportunities," *Pervasive and Mobile Computing 5*, pp. 277–298, 2009.

[4] S. Young, "Emerging technology," *The Ingenia Magazine*, no. 54, pp. 41–46, 2013.

[5] C. Jutten and J. Herault, "Blind separation of sources, part I-III," *Signal Processing*, vol. 24, pp. 1–29, 1991.

[6] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[7] A. K. Nandi, *Blind estimation using higher order statistics*. Kluwer Academic Plublishers, 1999.

[8] L. Castedo, C. J. Escudero, and A. Dapena, "A blind signal separation method for multiuser communications," *IEEE Transactions on Signal Processing*, vol. 45, no. 5, pp. 1343–1348, 1997.

[9] Y. Li and K. R. Liu, "Adaptive blind source separation and equalization for multiple-input/multiple-output systems," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2864–2876, 1998.

[10] J. T. Chien and B. C. Chen, "A new independent component analysis for speech recognition and separation," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1245–1254, 2006.

[11] T. P. Jung *et al.*, "Imaging brain dynamics using independent component analysis," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.

[12] A. D. Back and A. S. Weigend, "A first application of independent component analysis to extracting structure from stock returns," *International journal of neural systems*, vol. 8, no. 4, pp. 473–484, 1997.

[13] J.-F. Cardoso, J. Delabrouille, and G. Patanchon, "Independent component analysis of the cosmic microwave background," *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA03)*, 2003.

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

[15] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[16] M. Pedersen, J. Larsen, U. Kjems, and L. Parra, "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Processing and Speech Communication*, pp. 1–34, 2007.

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

[18] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley & Sons, 2004.

[19] T. Kim, I. Lee, S. Lee, and T. Lee, "Independent vector analysis: definition and algorithms," *Signals, Systems and Computers*, pp. 1393–1396, 2006.

[20] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Sys. and Tech.*, vol. 15, pp. 18–33, 2005.

[21] M. Cooke and D. P. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communications*, vol. 35, no. 3, pp. 141–147, 2001.

[22] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–351, 2008.

[23] A. S. Bregman, *Auditory Scene Analysis.* Cambridge: MIT Press, 1990.

[24] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

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

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

[27] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Mag.*, vol. 13, pp. 47–60, 1996.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stastical Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[29] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Mag.*, vol. 25, pp. 131–146, 2008.

[30] D. Peel and G. J. McLachlan, "Robust mixture modeling using the t-distribution," *Statistics and Computing*, pp. 339–348, 2000.

[31] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Networks*, pp. 127–138, 2007.

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

[33] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 7, pp. 1640–1655, 2005.

[34] D. L. Wang and J. G. Brown, *Computational scene analysis: Principles, algorithms and applications*. Wiley-IEEE Press, 2006.

[35] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1856–1866, 2010.

[36] P. C. Loizou, *Speech enhancement: Theory and practice.* Boca Raton, FL: CRC Press, 2007.

[37] G. J. Brow and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[38] K. Torkkola, "Blind separation for audio signals - are we there yet?," *ICA'99*, pp. 239–244, 1999.

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

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

[41] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," *Proc. of the 1997 Conference on Advances in Neural Information Processing Systems 10*, pp. 273–279, 1998.

[42] D. Pham and P. Garat, "Blind separation of mixture of independent sources through a quazi- maximum likelihood approach," *IEEE Trans. Signal processing*, vol. 45, pp. 1712–1725, 1997.

[43] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, 1997.

[44] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*, vol. 8, pp. 757–763, 1996.

[45] T. Eltoft, T. Kim, and T. Lee, "On the multivariate Laplace distribution," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 300–303, 2006.

[46] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," pp. 722–727, 2001.

[47] C. Colomes, C. Schmidmer, T. Thiede, and W. C. Treurniet, "Perceptual quality assessment for digital audio: PEAQ - The new ITU standard for objective measurement of the perceived audio quality," 1999.

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

[49] A. Cichocki and S. Amari, *Adaptive blind signal and image processing: learning algorithms and applications*. John Wiley, 2002.

[50] S. M. Naqvi, *Multimodal Methods for Blind Source Separation for Audio Sources*. PhD thesis, Loughborough University, 2009.

[51] S. Araki, S. S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.*, vol. 3, pp. 881–884, 2004.

[52] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica," *Proceedings of the Fifth International Congress, ICA 2004*, pp. 898–905, 2004.

[53] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[54] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," *Proc. ICA (LNCS 3195)*, vol. 15, pp. 430–436, 2004.

[55] P. D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, pp. 1–17, 2008.

[56] J. Mouba, "MOSPALOSEP, a platform for the binaural localization and separation of spatial sound using models of interaural cues and mixture models," *Proc. of the 13th Int. conference on Digital Audio Effects*, 2010.

[57] M. S. Khan, S. M. Naqvi, A. Rehman, W. Wang, and J. Chambers, "Video-aided model-based source separation in real reverberant rooms," *IEEE Trans. Speech and Audio Processing*, vol. 21, no. 9, pp. 1900–1921, 2013.

[58] A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 9, pp. 1434–1448, 2014.

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

[60] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 2233–2243, 2014.

[61] M. I. Mandel, *Binaural Model-Based Source Separation and Localization*. PhD thesis, Columbia University, 2010.

[62] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," *Proc. Interspeech*, pp. 419–422, 2008.

[63] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[64] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. Syst., Man, Cybern. Part C*, vol. 32, no. 4, pp. 474–484, 2002.

[65] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 4, pp. 320–327, 1976.

[66] G. Mclachlan and D. Peel, *Finite mixture models*. J. Wiley & Sons, 2000.

[67] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[68] D. Böhning, *Computer-assisted analysis of mixtures and applications: Meta-analysis, disease mapping and others*. Chapman& Hall/CRC, 1999.

[69] M. I. Mandel and D. P. W. Ellis, "A probability model for interaural phase difference," *Proc. ISCA Workshop Statist. Percept. Audio Process. (SAPA)*, pp. 1–6, 2006.

[70] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust*, pp. 275–278, 2007.

[71] C. Hummersone, *A Psychoacoustic Engineering Approach to Machine Sound Source Separation in Reverberant Environments*. PhD thesis, University of Surrey, 2011.

[72] J. F. Culling, K. I. Hodder, and C. Toh, "Effects of reverberation on perceptual segregation of competing voices," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2871–2876, 2003.

[73] D. M. Rocke and D. L. Woodru, "Identification of outliers in multivariate data," *J. Am. Statist. Assoc. 91*, pp. 1047–1061, 1996.

[74] A. S. Kosinski, "A procedure for the detection of multivariate outliers," *Computational Statistics & Data Analysis 29*, pp. 145–161, 1999.

[75] D. Gerogiannis, C. Nikou, and A. Likas, "The mixtures of Student's t-distribution as a robust framework for rigid registration," *Image and Vision Computing*, pp. 1285–1294, 2008.

[76] J. S. Garofolo *et al.*, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[77] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[78] B. S. Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, pp. 3100–3115, 2005.

[79] S. M. Stigler, "Thomas Bayes's Bayesian inference," *Journal of the Royal Statistical Society. Series A (General)*, vol. 145, no. 2, pp. 250–258, 1982.

[80] E. S. Fienberg, "When did Bayesian inference become "Bayesian"?," *Bayesian Analysis*, pp. 1–41, 2006.

[81] S. M. Stigler, "Laplace's 1774 memoir on inverse probability," *Statistical Science*, pp. 359–363, 1986.

[82] V. Šmídl and A. Quinn, *The variational Bayes method in signal processing.* Springer Science & Business Media, 2006.

[83] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Speech and Audio Processing*, vol. 14, no. 6, pp. 2174–2188, 2006.

[84] C. M. Bishop, "Variational principal components," *Ninth International Conference on Artificial Neural Networks, ICANN*, vol. 1, pp. 509–514, 1999.

[85] H. Attias, "A variational Bayesian framework for graphical models," *Advances in neural information processing systems*, vol. 12, no. 1-2, pp. 209–215, 2000.

[86] G. Parisi, *Statistical Field Theory*. Addison-Wesley, 1988.

[87] N. Nasios and A. G. Bors, "Variational learning for gaussian mixture models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 849–862, 2006.

[88] A. L. J. Taghia, N. Mohammadiha, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," *Proc. of Int. Conference on Acoustics, Speech and Signal Processing*, pp. 253–256, 2012.

[89] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Bayesian unification of sound source localization and separation with permutation resolution," *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pp. 2038–2045, 2012.

[90] T. S. Jaakkola, *Tutorial on Variational Approximation Methods*. MIT Press, 2000.

[91] J. M. Bernardo and A. F. M. Smith, *Bayesian theory*. Wiley, 1994.

[92] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2007, pp. 151–181.

[93] C. Archambeau, J. A. Lee, M. Verleysen, *et al.*, "On convergence problems of the EM algorithm for finite Gaussian mixtures.," vol. 3, pp. 99–106, 2003.

[94] K. Yamazaki and S. Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *Neural networks*, vol. 16, no. 7, pp. 1029–1038, 2003.

[95] M. Svensén and C. M. Bishop, "Robust Bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[96] S. Richardson and J. P. Green, "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society: series B (statistical methodology)*, vol. 59, no. 4, pp. 731–792, 1997.

[97] C. Archambeau, *Probabilistic models in noisy environments: and their application to a visual prosthesis for the blind.* PhD thesis, UCL, 2005.

[98] B. S. Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, 2000.

[99] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," *International Conference on Independent Component Analysis and Signal Separation*, pp. 734–741, 2009.

[100] L. D. Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.

[101] Z. Zohny and J. A. Chambers, "Modelling interaural level and phase cues with Student's t-distribution for robust clustering in MESSL," *Proc. Dig. Sig. Process.*, pp. 59–62, 2014.

[102] R. Plomp, "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," *Acustica*, vol. 34, pp. 200–211, 1976.

[103] S. Chandna and W. Wang, "Improving model-based convolutive blind source separation techniques via bootstrap," *Proc. IEEE Statistical Signal Processing Workshop (SSP 2014)*, 2014.