Gain Adapted Optimum Mixture Estimator for Single Channel Speech Separation

Thesis submitted in partial fulfillment of the requirement for the award of the degree of

MASTER OF ENGINEERING

In

ELECTRONICS & COMMUNICATION ENGINEERING

Submitted by

Divneet Singh Kapoor

Roll no. 800961007

Under the Guidance of

Dr. Amit Kumar Kohli

Assistant Professor

Electronics and Communication Engineering Department

Thapar University, Patiala-147004 (PUNJAB)

June 2011

CERTIFICATE

I, **Divneet Singh Kapoor**, hereby certify that the work which is being presented in this thesis entitled **"Gain Adapted Optimum Mixture Estimator for Single Channel Speech Separation"** by me in partial fulfillment of the requirements of the award of the degree of Masters of Engineering in Electronics and Communication Engineering from Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Amit Kumar Kohli**.

The matter presented in this thesis has not been submitted in any other University/Institute for the award of the degree of Masters of Engineering.

Divneet Singh Kapoor

Date..01/07/2011.....

This is certified that the above statement made by the candidate is correct to the best of my knowledge.

Dr. Amit Kumar Kohli

Assistant Professor, ECED,

Thapar University, Patiala

Date.....1/7/2011.....

Countersigned by:

Professor & Head, ECED,

Thapar University, Patiala

Date..1/7/11.....

Dr. S. K. Mohapatra

Dean of Academic Affairs

Thapar University, Patiala

Date.....1/7/2011.....

ACKNOWLEDGEMENT

No volume of words is enough to express my gratitude towards my guide, **Dr. Amit Kumar Kohli**, Assistant Professor, Electronics and Communication Engineering Department, Thapar University, who has been very concerned and has aided for all the material essential for the preparation of this thesis report. He has helped me to explore this vast topic in an organized manner and provided me with all the ideas on how to work towards a research-oriented venture.

I am also thankful to **Dr. A. K. Chatterjee**, Head of Department, ECED and **Ms. Alpana Aggarwal**, P.G. Coordinator, for the motivation and inspiration that triggered me for the thesis work.

I would also like to thank the staff members and my colleagues who were always there in the need of the hour and provided with all the help and facilities, which I required, for the completion of my thesis.

Most importantly, I would like to thank my parents and the Almighty for showing me the right direction out of the blue, to help me stay calm in the oddest of the times and keep moving even at times when there was no hope.

Divneet Singh Kapoor

Roll No. - **800961007**

ABSTRACT

While automatic speech recognition has become useful and convenient in daily life as well as an important enabler for other modern technologies, speech recognition accuracy is far from sufficient to guarantee a stable performance. It can be severely degraded when speech is subjected to additive noises. Though speech may encounter various types of noises, the work described in this thesis concerns one of the most difficult problems in robust speech recognition, corruption by an interfering speech signal with only a single channel of information. This thesis deals with the separation of mixed speech signals from a single acquisition channel; a problem that is commonly referred to as *Single Channel Speech Separation (SCSS).* This problem is especially difficult because the acoustical characteristics of the desired speech signal are easily confused with those of the interfering masking signal, and because useful information pertaining to the location of the sound sources is not available with only a single channel. The phenomenon of single channel speech commonly occurs due to the combination of speech signals from simultaneous and independent sources into one signal at the receiving microphone, or when two speech signals are transmitted simultaneously over a single channel. An efficient single channel speech separation system is an important front-end component in many applications such as Automatic Speech Recognition (ASR), Speaker Identification (SID), and hearing aids.

The separation process of single channel speech consists, mainly, of three stages: *Analysis, Separation,* and *Reconstruction.* The central separation stage represents the heart of the system in which the target speech is separated from the interfering speech. At the front, since the separation process works on one segment of single channel speech at a time, a mean must be found in the analysis stage to accurately classify each segment into single or multi-speaker before separation. Precise estimation of each speaker's speech model parameters is another important task in the analysis stage. The speech signal of the desired speaker is finally synthesized from its estimated parameters in the reconstruction stage. In order to have a reliable overall speech separation system, improvements need to be achieved in all three stages.

The goal of the thesis is to recover the target component of speech mixed with interfering speech, and to improve the recognition accuracy that is obtained using the recovered speech signal. Various techniques are employed to separate the sources' signals from the linear mixture of the sources, which include Model-based SCSS, Blind Source Separation, and Computational Auditory Scene Analysis. Usually model based separation is used in which Sources are modelled using Composite Source Modelling making the use of Gaussian Models. The thesis introduces the optimum mixture estimator for the estimation of the mixture, of the two sources' signals underlying, with different ratios. The mixture is also estimated by various estimators, namely MixMax, Quadratic estimators. Various estimators are compared on the basis of Mean Squared Error, and finally the sources' signals are estimated from the mixture estimate in Minimum Mean Squared Error sense.

Keywords: *Single Single channel speech separation (SCSS), optimum mixture estimator, mixture-maximization (MIXMAX), quadratic estimator, gain adaption.*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

ASA	Auditory Scene Analysis
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CDF	Cumulative Distribution Function
DFT	Discrete Fourier Transform
EVD	Eigen Value Decomposition
ICA	Independent Component Analysis
IID	Interaural Intensity Difference
IRM	Ideal Ratio Mask
ITD	Interaural Time Difference
MAP	Maximum A-posteriori
MCSS	Multi Channel Speech Separation
MIXMAX	Mixture-Maximization
ML	Maximum Likelihood
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PDF	Probability Density Function
SCSS	Single Channel Speech Separation
SE	Speech Enhancement
SID	Speaker Identification
SSR	Signal to Signal Ratio
STFT	Short Term Fourier Transform
WER	Word Error Rate

CHAPTER 1
INTRODUCTION

1.1 Motivation

In many speech processing applications such as speech recognition, speaker identification, and speech enhancement, the input speech signal is often corrupted by the surrounding acoustic noise. This in turn deteriorates the perceived quality and intelligibility of the speech and consequently degrades the overall performance of the speech processing algorithm. When the acoustic interference consists of competing speech signals from other talkers (known commonly as the cocktail-party effect), then further degradation results due to the similarity in nature between the desired and undesired signals. In such scenarios, therefore, a speech separation algorithm represents an essential front-end component to enhance speech quality and intelligibility for further processing. If we can separate the desired speech signal prior to its processing, this will help in enhancing the overall performance of the speech processing algorithm.

Speech recognition has been the object of extensive research for many decades. While recognition accuracy in clean environments improved substantially, recognition in noisy environments still suffers due to many reasons such as the mismatch between clean training and noisy testing conditions. Among many different types of interference (including but not limited to white noise, colored noise, background music, and speech babble), competing speech has been considered to be among the most challenging type of interference. The high correlation of temporal structures between the speech from target and masking speakers is one major reason for poor recognition accuracy. Nevertheless, in daily communication among humans, competing speech is among one of the most commonly-encountered noises. For example, speech by news anchors is sometimes overshadowed by background speakers and multiple speakers talk simultaneously in teleconferences. In both examples mentioned above, the target speech is more or less corrupted by interfering speech. While machines still do a very poor job of recognizing combined speech correctly, human beings are impressive

in their ability to either extract the target speech, suppress interfering speech sources or both to achieve reasonably good recognition accuracy in communicating with each other.

In some practical situations where only a single acquisition channel is available, single channel separation techniques must be used. This may be imposed by the system used (as telephone based applications) or by the availability of the desired signal (as pre-recorded applications). They are especially interesting due to the simplicity in microphone installation but the major constraint of single channel methods is that there is no reference signal for the interference available. Therefore the power spectral density of the interfering speech has to be estimated based on the available single channel speech signal only and this is what makes it a challenging task. This problem is commonly referred to as the ***single channel speech separation*** problem.

1.2 Why is Single-Channel Speech Separation Important?

In the real world, speech activity is collected by a single microphone or by multiple microphones and sent to computers for further processing. During the collection procedure, if conditions permit, multiple microphones are naturally preferred. In this case, spatial information can be preserved and used as additional cues to separate combined speech. However, in cocktail-party environments with multiple sound sources, if the target speaker is not predetermined, microphone arrays may not be used to good advantage. Even worse, an environment that facilitates multiple microphones is not always available. In many scenarios using one microphone is the only choice.

One good example of single-channel speech processing is automatic speech recognition of radio broadcasts. In this case, speech activity is transmitted and collected from radio channels, and there is no spatial information available. In many speech segments, the news anchor's voice is corrupted by background speakers. Directly sending this simultaneous speech into a speech recognizer results in poor accuracy. Another example is speech recognition in teleconferences. The presence of more than one interfering speaker presents a

very difficult task for any state-of-the-art recognizer. The examples above are frequently the first step of some very complex systems. Usually, these systems apply further processing to the output of the recognition system, such as news summarization, categorization, question answering, dialogue systems and text-to-speech systems. All these applications require a good single-channel speech system to achieve reasonably good speech recognition accuracy, as poor speech recognition accuracy may lead to a serious accumulation of errors. For these reasons solutions to the problem of single-channel speech recognition in interfering noise are very important.

1.3 Why is Single-Channel Speech Separation Challenging?

It is widely believed that single-channel speech separation (SCSS) is a very challenging task. Unlike multi-channel speech separation (MCSS), spatial information cannot be utilized. Only those intrinsic acoustic features, such as pitch, harmonic structure, local time or frequency proximity can be exploited to separate speech. Due to highly correlated temporal structures, it is very difficult to extract many good inherent acoustic features accurately from combined speech. Pitch information has been widely considered to be a good way to extract harmonic structure. But it is very difficult to estimate accurately pitch contours from target speech while interfering speech is present. Since competing speech contains many similar human speech characteristics, unlike the case of other noise types such as white / colored noise or mechanical noise, the usage of time/frequency proximity, amplitude modulation and other features provides only limited improvement. The major goal of this thesis is to address the speech-on-speech problem by developing a model that separates speech sources based on mixture estimation of the signals mixed, when only one microphone is available.

CHAPTER 2
BACKGROUND & LITERATURE REVIEW

In a natural world, a speech signal is frequently accompanied by other sound sources upon reaching auditory systems, yet listeners are capable of holding conversations in a wide range of conditions. Sounds are created by a wide range of acoustic sources, such as several people talking during a cocktail party. The typical source generates complex acoustic energy that has many frequency components. In a quiet environment, it is usually easy to understand what a person is saying. In many listening situations however, different acoustic sources are active at the same time, and only the sum of those spectra will reach the listener's ears. Therefore, for individual sound patterns to be recognized – such as those arriving from a particular human voice among a mixture of many – the incoming auditory information must be partitioned, and the correct subset of elements must be allocated to individual sounds so that a veridical description may be formed for each. This is a complicated task because each ear has access only to a single pressure wave that is the sum of the pressure waves from all individual sound sources. This phenomenon is well known as the "cocktail party" effect.

2.1 Cocktail Party Effect

The cocktail party effect describes the ability to focus one's listening attention on a single talker among a mixture of conversations and background noises, ignoring other conversations. The effect enables people to talk in a noisy place. For example, when conversing in a noisy crowded party, people can still listen and understand the person they are talking with, and can simultaneously ignore background noise and conversations. Nevertheless, if someone calls out their name from across the room, people will immediately notice. Another aspect of the cocktail party effect is de-reverberation. In a normal room a listener perceives much less echo and reverberation than a microphone recording does. The human auditory system seems to ignore most of the reflected sound, because it arrives from other directions than the direct sound. The auditory system can also switch the direction of attention and turn from one sound source to another.

2.1.1 Binaural Processing

The cocktail party effect is a **binaural effect**, which requires hearing with both ears. Persons with only one functional ear are much more disturbed by interfering noise than people with two healthy ears. The cocktail party effect is related to the localization of sound sources. Experiments have shown that the auditory system is able to localize at least two sound sources simultaneously and assign the correct sound source characteristics to these sound sources simultaneously too. In other words, as soon as the auditory system has localized a sound source, it can extract the signals of this sound source out of a mixture of interfering sound sources. It is assumed that the auditory system performs a kind of cross-correlation function between both ear signals. A cross correlation function projects signals onto an axis, which corresponds to the time difference between both ear signals. For example, sound with an interaural time difference of 0.3 ms is projected onto the 0.3 ms position of the correlation axis. If multiple sound sources are present, then complex correlation patterns appear. The statistical parameters of these patterns, like mean value and variance, depend on the directions and levels of the sound sources.

The auditory system is obviously able to analyze these patterns and determine the signals of a dedicated sound source. Attempts have been made to simulate the cocktail party effect by technical means. Cocktail party processors have been constructed which can extract the signal of a single sound source out of a mixture of sound sources. There are cocktail party processors, which are based on correlation functions, evaluating interaural time differences, but there are also cocktail party processors for interaural level differences. However, the principles of the human cocktail party effect are not yet fully investigated. Technical cocktail party processors do not yet reach the capabilities of the human auditory system.

2.1.2 Monaural Processing

The auditory system does not only use methods for a direction specific signal processing, it also uses monaural effects for noise reduction. If the characteristics of a desired signal are known (like the characteristics of speech) or can be estimated (like expected phonemes at

observed mouth movements), then all signal components which do not match the expected characteristics can be suppressed and the disturbing effect of this noise can be reduced. The human pinna (the external flap of skin and cartilage of the ear) is a directionally-dependent filter that selectively removes particular frequencies, based on the direction from which sound comes. This filter can distinguish sounds from above vs. below, and from front vs. back, even when only a single ear is used.

It is valuable to make a computer have the ability of a human being to segregate the object source from other interfering sources. An effective separation system can greatly facilitate many applications, including automatic speech recognition (ASR), speaker identification, audio retrieval, digital content management, etc. Therefore, the research on speech separation gradually catches the researchers' attentions, and it has become an increasingly popular topic in the field of signal processing.

2.2 Single Channel Speech Separation

Source separation problems in digital signal processing are those in which several signals have been mixed together and the objective is to find out what the original signals were. The classical example is the "cocktail party problem", where a number of people are talking simultaneously in a room (like at a cocktail party), and one is trying to follow one of the discussions. The human brain can handle this sort of auditory source separation problem, but it is a very tricky problem in digital signal processing. This was first analyzed by Colin Cherry [1].

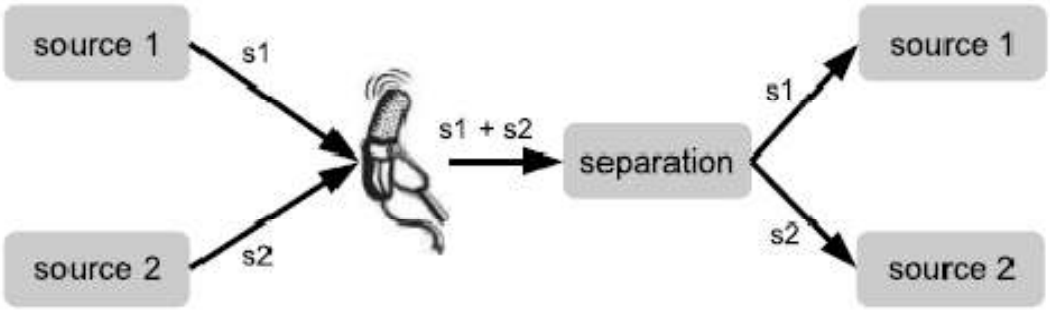

Fig. 2.1. *Single channel speech separation*

The idea of a single channel speech separation is to automatically process the mixed signal in order to recover each talker's original speech. Minimizing artifacts in the processed speech is a key concern, especially if the final goal is to use the recovered speech in machine-based applications such as automatic speech recognition and speaker identification systems. The goal therefore of single channel speech separation algorithm is to:

• Improve the perceptual aspects of a degraded speech signal.
• Improve the performance of the final speech processing system.
• Increase the robustness of machine-based speech processing systems.

Several approaches have been proposed for the solution of this problem but development is currently still very much in progress. Some of the more successful approaches are principal components analysis and independent components analysis, which work well when there are no delays or echoes present; that is, the problem is simplified a great deal. It can be categorized in several ways: Depending upon the amount of available information about the mixing process and sources, it can be divided into Blind Source Separation (BSS) and semi-BSS; According to the relation of ***'n'*** (number of sources) and ***'m'*** (number of sensors), it falls into the categories of an under-determined problem ($m < n$), even-determined problem ($m = n$), and over-determined problem ($m > n$); Based upon relation between sources, it is either a problem with independent sources, or a problem with dependent sources.

Most of the source separation algorithms are based on the assumption that the sources are statistically independent, which holds in most cases. The very few algorithms dealing with dependent sources include both semi-BSS techniques and BSS techniques. Compared with even-determined and over-determined problems, under-determined problems are much more difficult due to lack of constraints. Additional constraints are normally applied by making strong assumptions on the source characteristics, incorporating source models or providing prior knowledge on the mixing process and/or the signals.

The field of computational auditory scene analysis attempts to achieve auditory source separation using an approach that is based on human hearing. The human brain must also

solve this problem in real time. In human perception this ability is commonly referred to as auditory scene analysis or the Cocktail party effect. While the detailed mechanism of exactly how humans separate signals is still not truly clear, one popular computational model, auditory scene analysis (ASA) suggests that the one-dimensional speech signal is projected onto a two-dimensional time-frequency space for further processing. After this projection, some capabilities that humans use in image detection (e.g. edge detection) can also be used to separate speech sources from one another. Although it is clearly evident that the human auditory system is very proficient at focusing on a particular speaker or speakers in a mixture, computer algorithms, on the other hand, designed to do the same task have demonstrated only a limited degree of success. Therefore, exploring new strategies to combat this problem has been of great interest in recent years.

2.2.1 Speech Separation Systems

Researchers have developed various types of systems to address the problem of minimizing the negative effects from competing speech, which can be characterized in a number of different ways. For example, systems with inputs from multiple microphones have the advantage of being able to take advantage of exploiting spatial information such as interaural time difference (ITD) and interaural intensity difference (IID). Nevertheless, multiple microphones are not always available, so there will always be a role for single-channel systems, which have only the intrinsic information carried by speech itself, in the absence of spatial cues. Systems can be also characterized as being knowledge-based versus statistically-based. This is actually a continuum that depends on the extent to which the structure of the system is developed manually through background knowledge about speech versus statistical learning from a large database. Knowledge-based systems have the advantage of requiring fewer constraints and in principle are more easily adapted to unknown speakers. However, due to the lack of precise fundamental knowledge of how human perceptual processing really works, imperfect modeling makes this type of system tend not to be able to achieve the same level of word error rate (WER) that statistically-based systems enjoy. In contrast, statistically-based systems frequently require significant computational resources as well as a pool of speakers for training and testing. In many cases, the WER from speech recognition

systems becomes worse due to changes in the testing environment. Speech separation systems can also be characterized as being directed toward speech enhancement (which refers to improving the quality of speech for human listeners) versus recognition-based systems, which are designed to improve automatic speech recognition accuracy.

2.3 Review of Techniques Used For SCSS

With the development of advanced machine learning, great interest to solve open problems in speech processing has emerged in recent years. One such problem is monaural (or equivalently single channel) speech separation in which the goal is to estimate the underlying speech signals $\boldsymbol{X_i(t)}$ from the observed mixture signal $\boldsymbol{Y(t)}$, which for the two speaker case is often expressed as

$$\boldsymbol{Y(t) = X_1(t) + X_2(t)}$$

Mathematically, there are infinitely many solutions for the SCSS problem unless we impose some constraints on the sources, such as statistical independence. In the big picture, approaches proposed to solve the SCSS problem are categorized into three branches:

1. **Under-determined Blind Source Separation (BSS)**
2. **Computational Auditory Scene Analysis (CASA)**
3. **Model-based SCSS**

The two branches, namely BSS and Model based SCSS, mainly rely on *a priori* knowledge of sources obtained during a training phase. In contrast, CASA-based approaches seek discriminative features in the observation signal to separate the speech signals.

2.3.1 Blind Source Separation (BSS)

Blind Source Separation is a recent, and remarkable, chapter in the development of signal processing. The field began at a neural networks conference in Utah in 1986 where Jutten and Herault presented a paper entitled “Space or time adaptive signal processing by neural network models” [2].

Source separation has long been a topic of interest in electrical engineering. Many algorithms have been developed to perform separation, but prior to BSS major assumptions were always required on the nature of the sources. Jutten and Herault's technique was revolutionary in that it did not require assumptions on the nature of the signals being separated. Despite this, however, their technique did not initially attract much attention. This is primarily because in the 1980s, neural network research focused on Hopfield networks and Jutten and Herault's work went largely unnoticed. It was only with a much clearer formulation of BSS by Comon in 1994 [3] that BSS became a mainstream topic of research.

A single audio signal can be modelled with its probabilistic representation, the time varying structure, and its decomposition into fundamental basis functions that produce an efficient coding scheme. The generative model for a single source can be extended into a multiple source observation problem. The problem is then to understand the relationship between sources and how to model their interaction with little a priori knowledge. Blind Source Separation (BSS) consists in estimating ***n*** signals (the sources) from the sole observation of ***m*** mixtures of them (the observations). Depending upon the number of sources and the sensors, the BSS can be specified into three types of problems:-

1. **Over Determined Problem (m > n)**
2. **Even Determined Problem (m = n)**
3. **Under Determined Problem (m < n)**

Blind source separation is a prime example for modelling multiple sources in an environment. Furthermore, the model is realistic since audio signals do not occur isolated but are active simultaneously. Multiple source models may be given for single channel observations as well as multiple channel observations. To model the interactions with changing environments, this multiple source model needs to be further extended to include contextual changes due to the environment or non-stationary character of the sources. The model should be able to make inference about the environmental dynamics, possibly track signal sources, and understand the structure of the interacting source signals. In its simplest form, a source can be a random variable with a fixed probability density function. A non-linear function such as the *sigmoid* function or the *tanh* function could represent the

cumulative density for the source signal. This non-linear function was used to separate super-Gaussian sources. This was a sufficient model because the goal was to estimate an un-mixing matrix and the observation model was linear, deterministic (no sensor noise) and there were as many sources as given observation channels. There are many ways to extend this source model to include other density functions such as sub-Gaussian sources and more complicated source densities that can be modelled with a mixture of Gaussians.

Underdetermined BSS is another branch in which the SCSS is studied. In underdetermined BSS, as opposed to ordinary BSS techniques [3]-[6], we are unable to exploit the "spatial diversity" [7] of the sources since the number of observations is less than the number of sources. In underdetermined BSS techniques, using independent component analysis (ICA) [8], nonnegative matrix factorization (NMF) [9], or sparse coding [10], the sources are projected onto a set of basis functions whose coefficients are as sparse as possible. Then given the basis functions and coefficient distributions, techniques such as maximum *a-posteriori* (MAP) estimation are used to estimate the sources. These techniques do not work well when the trained basis functions of two sources overlap; mixture of two speech signals is such a case [11]. The goal of BSS is to determine the original sources given mixtures of those sources. The adjective **`blind'** emphasizes that the sources signals are not observed and no information is available with regard to the mixture process.

Algorithms developed to perform blind separation of sources were given the name Independent Component Analysis (ICA) algorithms and the term ICA is often used interchangeably with BSS. In this thesis, BSS refers to the entire body of knowledge relevant to blindly separating signals, whereas the term ICA is reserved more specifically for algorithms that perform this separation. Separation techniques were named ICA to highlight the fact that independent components were being separated from mixtures of signals, but also to emphasise a close link with the classical signal processing technique of Principal Component Analysis (PCA). PCA can be used to separated mixtures of signals using decorrelation. A well-known fact from elementary statistics, however, is that for non-Gaussian signals, uncorrelated signals are not necessarily independent. To de-correlate, it is

only necessary to consider second order statistics, whereas, in general, independence requires higher order statistics.

As a result, it is common to consider ICA to be an extension of PCA that is able to separate non-Gaussian signals. This partly explains the late development of ICA as until fairly recently, Gaussian sources were assumed in most signal processing research. Despite being limited to second-order statistics, PCA is still a powerful technique and has many uses, including feature extraction and data compression [12]. Following Comon's seminal paper, there was a rapid proliferation of ICA algorithms. Algorithms were formulated based on a wide variety of principles, including mutual information [5], maximum likelihood [13] and higher order statistics [3], to name just a few of the more popular approaches. Despite such wide variety, all ICA algorithms are fundamentally similar. ICA algorithms invariably obtain estimates of the independent signals by adopting a numerical approach (e.g. gradient descent) to maximizing an "independence metric", i.e. a measure of the signals' independence. The main difference between different ICA algorithms is in the metric that is used. These ideas form the basis of ICA and they are explored more thoroughly in Chapter 3.
On publication of their algorithm in 1995, Bell and Sejnowski's [5] approach to ICA for details) became the most popular choice due to its simplicity and its favourable convergence properties (see Chapter 3). However, the algorithm involved matrix inversion which significantly hindered efficiency. Amari discovered an important improvement (using "natural" gradient descent, see [14]) to the algorithm of Bell and Sejnowski which eliminated the matrix inversion. This gave a significant performance improvement and made ICA more practical for real world problems, especially in separating large numbers of sources. Another important ICA algorithm, called FastICA [15], was developed in 1997 by Oja and Hyv¨arinen of the Helsinki University of Technology. FastICA is examined in detail in Chapter 3. It was shown to be a very good alternative to Bell and Sejnowski's algorithm, and is probably currently the most widely used ICA algorithm.

With the explosion of interest in BSS, there came many different approaches to solving the source separation problem. A great deal of progress was made in showing that seemingly unrelated approaches were, in fact, equivalent. Bell and Sejnowski made a major contribution

to this movement by proposing a unifying framework for BSS based on information theoretic considerations. Continuing on the work of Bell and Sejnowski, BSS researchers soon showed the equivalence of many different approaches to BSS and as the field of Blind Source Separation has matured, research has led to a convergence onto a small set of well understood principles.

2.3.2 Computational Auditory Scene Analysis (CASA)

Computational Auditory Scene Analysis (CASA) has broad application to source separation. Generally speaking, CASA is a wide collection of various computational implementations of auditory scene analysis (ASA). It is necessary to briefly introduce ASA and its major application. The CASA community have focused on both multiple and single microphone source separation problems under highly realistic acoustic conditions, but have used almost exclusively hand designed systems which include substantial knowledge of the human auditory system and its psychophysical characteristics (e.g. [16], [17]). Unfortunately, it is difficult to incorporate large amounts of detailed statistical knowledge about the problem into such an approach. On the other hand, machine learning researchers, especially those working on independent components analysis (ICA) and related algorithms, have focused on the case of multiple microphones in simplified mixing environments and have used powerful "blind" statistical techniques. These "un-mixing" algorithms (even those which attempt to recover more sources than signals) cannot operate on single recordings.

Furthermore, since they often depend only on the joint amplitude histogram of the observations they can be very sensitive to the details of filtering and reverberation in the environment. Un-mixing algorithms such as ICA and its extensions recover sources by reweighting multiple observation sequences, and thus cannot operate when only a single observation signal is available. A technique called *re-filtering* which recovers sources by a non-stationary reweighting ("masking") of frequency sub-bands from a single recording, and argue for the application of statistical algorithms to learning this masking function. Many scientists believe that audition shares many similarities with vision. The human auditory

system transforms speech into a neural representation which is then presumed to be processed in a fashion that is similar to image processing.

Biological Foundation of Auditory Scene Analysis

The neurobiological foundation of auditory scene analysis has received considerable attention over the last decade [18]. Evidence from single cell recordings shows that frequency periodicity, upon which concurrent sound segregation is partly based, is reflected within the patterns of afferent spike trains. Multi-unit recordings in nonhuman primates have also revealed a distinct pattern of neural activity in primary auditory cortex associated with conditions that promote sequential auditory stream formation. This suggests that both spectral and temporal transitions between successive stimuli are represented within the primary auditory cortex. Although these results suggest early bottom–up (stimulus–driven) processes in auditory scene analysis, representations of incoming acoustic information in the ascending auditory pathway are probably not sufficient for the detection and identification of different sound objects. It is argued that the discrimination and identification of auditory objects requires additional computations that follow the initial processing in the ascending pathway and primary auditory cortex, suggesting that these computations might be carried out in the planum temporale.

In addition, it has been proposed that identifying the content (what) and the location (where) of sound in the environment may be functionally segregated in a manner analogous to the ventral (what) and dorsal (where) pathways in the visual modality. In comparison, most studies involving the processing of sound identity (e.g., phoneme discrimination, pitch discrimination etc.) reported activation in the anterior portion of the temporal lobe and the inferior frontal gyrus. While neuro-imaging studies in humans have identified a number of regions that may contribute to auditory scene analysis, little is known about the time course of these neural events and how they relate to phenomenological experience.

Procedure of CASA

The entire ASA procedure can be separated into two stages: segregation and regrouping. In the first stage, speech is decomposed into a higher-dimensional space (such as a spectro-temporal two-dimensional representation) where similar units (*e.g.*, time-frequency cells in the previous 2D representation example) are collected together into different regions. In the second stage, these regions are grouped together into different streams based on the values of various acoustic cues or other information. Finally, the target speech or interfering speech or both can be reconstructed for different purposes. These functions will be discussed below in detail. In general, CASA uses computational methods to generate a machine perception system which may have similar functionality to that of humans. We consider primarily either one or two microphones (the two ears of human audition). In this scenario, CASA is defined as the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene."

Fig. 2.2 below shows a simplified diagram of a typical CASA system, where input speech is first going through a domain transformation function. Most of the time, this function transforms the one-dimensional speech signal into the very popular two dimensional time-frequency representation, either by standard short time Fourier transformation (STFT) or a Gammatone Filter bank. Following the domain transformation, the next procedure is speech component segregation. In this part, all time-frequency cells are segmented into different regions. All cells sitting in the same region are believed to be from the same speech source. Various feature extraction algorithms are proposed in this stage in order to optimize the segmentation results. The next step in the figure is described as speech component regrouping. This stage is processed in an utterance-based format to extract all the segments believed to be from the same speech source while suppressing all others. In this stage, the most popular method used is speaker identification based on the training data, from which each speaker's acoustic characteristics are learned by the system. In the reconstruction procedure, the *a-posteriori* probability of each speaker in the training pool is calculated for each time frame to get the best match. Based on speaker identification results, further extraction or suppression is performed to generate different speech streams. The most

popular such method is speaker identification. Finally, all extracted time-frequency cells are used to reconstruct the re-synthesized speech.

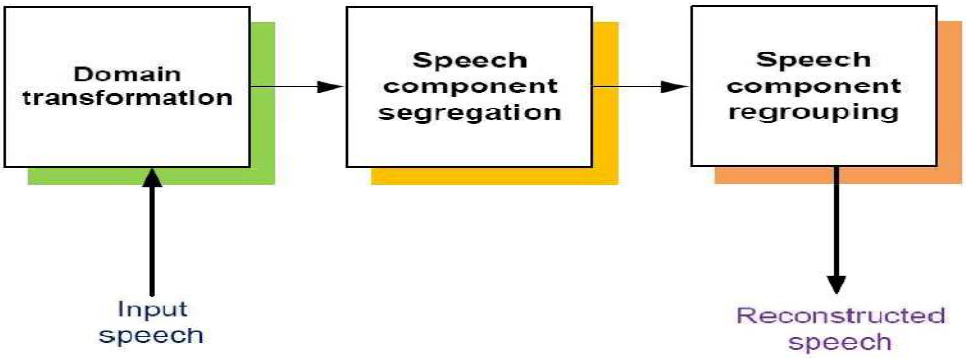

Fig. 2.2. *A simplified end-to-end CASA speech separation system, where the most three important parts are domain transformation, speech-component segregation, and speech-component regrouping.*

CASA systems have been widely used in many different applications. Below is a limited collection of several examples [19].

- **Robust automatic speech and speaker recognition**
- **Hearing prostheses**
- **Automatic music transcription**
- **Audio information retrieval**
- **Auditory scene reconstruction**

Projecting time-domain speech into the time-frequency domain is generally considered to be the first step towards solving the problem. The short-time Fourier Transform (STFT) and Gammatone Filter-bank are both considered to be proper vehicles to implement the transform. In STFT, the centre frequencies of each frequency channel are separated by the same difference in frequency. For each fixed time frame, the frequency response is the Fourier transform of that given time frame. While each frequency channel is fixed, the output of each channel can be considered as a filter output of a specific bandpass filter. In the application of Gammatone Filtering, the spacing of the centre frequency of each channel varies with frequency. This provides better frequency resolution for low frequencies at the

expense of worse spectral resolution for high frequencies. After the transformation is done, an attempt is made to determine which time-frequency cells are believed to have similar characteristics. This step is usually done by applying different intrinsic acoustic cues.

2.3.3 Model Based SCSS

Model-based SCSS techniques are spiritually similar to model-based single-channel speech enhancement (SE) techniques. In this case, SCSS can be considered as an SE problem in which both the target and interference, i.e., and, which are non-stationary sources with similar probabilistic characteristics, must be estimated. Concisely speaking, the following procedures are commonly applied in model-based SCSS techniques. First, patterns of the sources are obtained in the training phase. Then, those patterns whose combinations model the observation signal are chosen. Finally, the selected patterns are either directly used to estimate the sources or used to build filters which when imposed on the observation signal result in an estimate of the sources. A model-based monaural speech separation technique consists of three fundamental modules:

- **Feature Selection,**
- **Modelling, and**
- **Estimation.**

Accordingly, based on the choice of feature, model, or estimation method, wide varieties of model-based monaural speech separation techniques have been proposed. In monaural speech separation, signals are separated by means of processing features such as time waveform, log spectrum, modulated lapped transform, or discrete cosine transform, to specify a few. Among these features, the log spectrum domain is perceptually more important than the others such that it is often used in separation techniques. The relation between sources and observation in the log spectrum domain is, however, non-linear which makes the modelling and estimation stages complicated. Fortunately, a simple but effective approximation method known as Mix-Max can be applied which reduces the complexity dramatically. The short-term spectrum of speech is vulnerable when the short-term spectral

values are modified by the frequency response of the communication medium. As a result, the feature selection is an important issue and is often considered as an effective tool to obtain a reasonable performance in a certain application. Selecting better feature types renders the statistical models to obtain high quality separation results.

In order to model the statistical characteristics of the features, vector quantization, Gaussian mixture models, hidden Markov models, non-negative matrix factorization, and independent component analysis have been deployed. The modelling method should be not only accurate but also computationally economical; otherwise the estimation process becomes untraceable. Usually Gaussian Mixture Modelling is used to statistically model the sources and hence the mixture is modelled. The model is fully determined by the *a priori* probability of sub-sources and the probability of the observation occurring given the sub-source. Fig. 2.3 shows the modelling of the source using different Gaussian sub-sources.

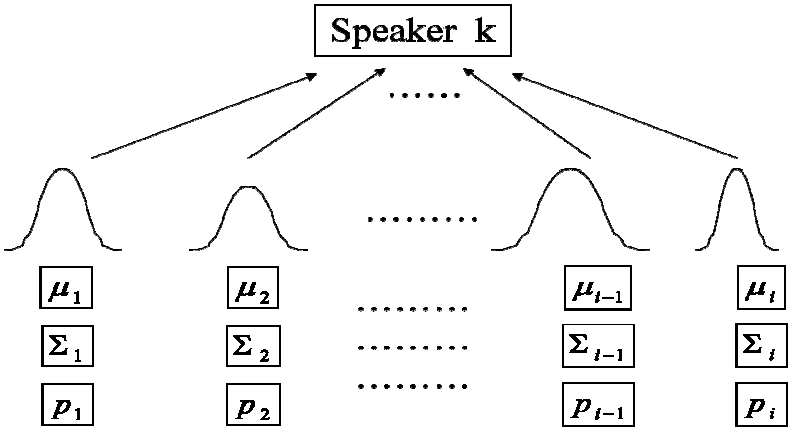

Fig. 2.3 *Gaussian Mixture Modelling*

Finally, ***estimation***, which forms the core of separation processing, is carried out using well-known probabilistic approaches such as maximum likelihood (ML), maximum *a-posteriori* and minimum mean square error (MMSE). A new monaural speech separation technique, in which the log spectrum, a Gaussian mixture model, and MMSE are used as the feature, the statistical model, and the estimation approach, respectively is used recently.

CHAPTER 3
BLIND SOURCE SEPARATION

Consider a situation in which we have a number of sources emitting signals which are interfering with one another. Familiar situations in which this occurs are a crowded room with many people speaking at the same time, interfering electromagnetic waves from mobile phones or crosstalk from brain waves originating from different areas of the brain. In each of these situations the mixed signals are often incomprehensible and it is of interest to separate the individual signals. This is the goal of Blind Source Separation.

This chapter presents the basic mathematical framework of BSS. The signal mixing model is presented in Section 3.1. A general description of the approach to achieving separation via BSS is given in Section 3.2. Section 3.3 details the underlying assumptions of the BSS framework and important ambiguities that are inherent to BSS are discussed in Section 3.4. A description of specific details of BSS algorithms is given in Sections 3.5 and 3.6 and the chapter concludes with two simple examples of applying Blind Source Separation in Section 3.7.

3.1 Mathematical Description of Source Mixing

The first step in deriving a solution to the source separation problem is to adequately model source mixing. BSS can be applied to a collection of statistically independent sources which are emitting signals that interfere with each other and the interfering signals are recorded using a number of spatially separated sensors. In this chapter, for the purpose of clarity, the simplest case where the number of sources is equal to the number of sensors is considered (an assumption that is frequently made in the literature).

Suppose we have N statistically independent signals, $s_i(t), i = 1, \dots\dots, N$. We assume that the sources themselves cannot be directly observed and that each signal, $s_i(t)$, is a realization of some fixed probability distribution at each time point t. Also, suppose we observe these signals using N sensors, then we obtain a set of N observation signals $x_i(t), i = 1, \dots\dots, N$

that are mixtures of the sources. A fundamental aspect of the mixing process is that the sensors must be spatially separated (e.g. microphones that are spatially distributed around a room) so that each sensor records a different mixture of the sources. With this spatial separation assumption in mind, we can model the mixing process with matrix multiplication as follows:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) \tag{3.1}$$

where $\boldsymbol{A} \in R^{N \times N}$ is an unknown matrix called the mixing matrix and $\boldsymbol{x}(t), \boldsymbol{s}(t) \in R^N$ are the two vectors representing the observed signals and source signals respectively. Incidentally, the justification for the description of this signal processing technique as blind is that we have no information on the mixing matrix, or even on the sources themselves.

The objective is to recover the original signals, $s_i(t)$, from only the observed vector $\boldsymbol{x}(t)$. We obtain estimates for the sources by first obtaining the "un-mixing matrix" $\boldsymbol{W}$, where:

$$\boldsymbol{W} = \boldsymbol{A}^{-1} \tag{3.2}$$

This enables an estimate, $\boldsymbol{y}(t)$, of the independent sources to be obtained:

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} \tag{3.3}$$

where the time index t has been omitted for notational simplicity.

The diagram in Fig. 3.1 illustrates both the mixing and un-mixing process involved in BSS. The independent sources are mixed by the matrix $\boldsymbol{A}$ (which is unknown in this case). We seek to obtain a vector $\boldsymbol{y}$ that approximates $\boldsymbol{s}$ by estimating the un-mixing matrix $\boldsymbol{W}$. If the estimate of the unmixing matrix is accurate, we obtain a good approximation of the sources.

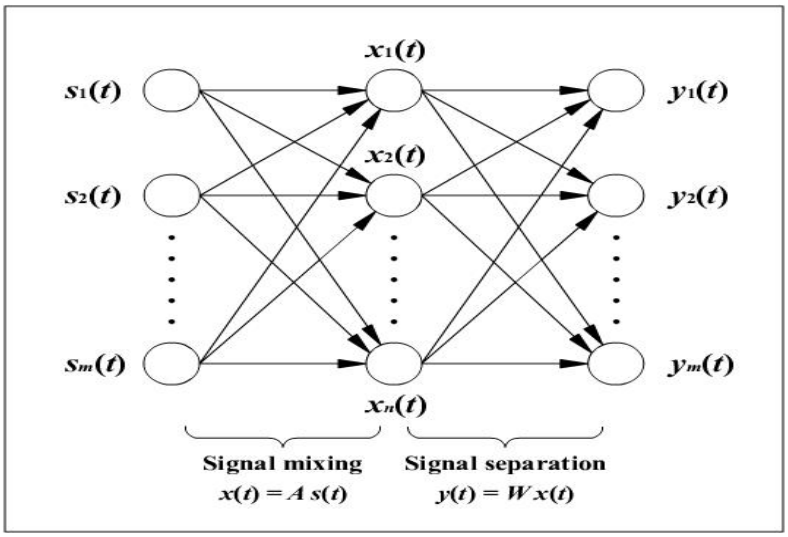

Fig. 3.1. *Sources* s *have been linearly mixed by the unknown mixing matrix* A *and we estimate the sources by estimating the un-mixing matrix* W.

3.2 Independent Component Analysis

As mentioned in Chapter 2, algorithms that perform Blind Source Separation are known as Independent Component Analysis (ICA) algorithms. Intuitively, ICA involves estimating the linear transformation that maximizes the independence of the signals. This linear transform is referred to as the un-mixing matrix, $\boldsymbol{W}$. Since the original sources, $s_i(t)$, were assumed to be independent, we know that maximising the independence of the components of $\boldsymbol{y}$ from Eq.(3.3) we will obtain estimates of the original sources. At this point it is important to place this intuitive notion of ICA on a more rigorous foundation. Suppose the observation vector $\boldsymbol{x}$ is formed according to Eq.(3.1), that is, $\boldsymbol{x}$ is a linear combination of independent components. To estimate one of the independent components we consider a linear combination of the x_i terms. Let y_i be one of the estimates, then we have:

$$y_i = \boldsymbol{W}^T \boldsymbol{X} \tag{3.4}$$

The crucial point is that if w was one of the rows of the inverse of the mixing matrix, then yi would actually be one of the original independent components. At this point, however, it is not clear how we can determine such a vector $\boldsymbol{W}$ when we have no information about the

mixing matrix $\boldsymbol{A}$. Independent Component Analysis has been developed to solve this problem. ICA depends fundamentally on the independence of the original sources, $s_i(t)$. The following equation, derived from Eq.(3.4), illustrates this link:

$$y_i = \boldsymbol{W}^T\boldsymbol{A}\boldsymbol{s} = \boldsymbol{Z}^T\boldsymbol{s} \tag{3.5}$$

From Eq.(3.5) it is clear that y_i is a linear combination of the independent components, $s_i(t)$.

The approach to Independent Component Analysis considered here is to consider the "Gaussianity" of the signals s_i of Eq.(3.5). From the Central Limit Theorem we know that the distribution of a sum of independent random variables approaches a Gaussian distribution [26]. From Eq.(3.5) it is clear that the estimate y_i is a sum of independent random variables, thus we expect its distribution to be "more Gaussian" than the distribution of the independent components s_i. Assuming none of the independent components are Gaussian distributed (see Section 3.3), we obtain the most "non-Gaussian" distribution for y_i when it is exactly one of the independent components s_i. As a result, by choosing the vector $\boldsymbol{W}$ to maximize the non-Gaussianity of y_i, we will obtain an estimate of one of the independent components s_i. This methodology gives an approach to separating the independent components, but to proceed further we require a method of measuring non-Gaussianity.

Perhaps the most straightforward approach to measuring non-Gaussianity is Kurtosis. Kurtosis is the fourth order cumulant [20] defined by:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 \tag{3.6}$$

The kurtosis of a Gaussian random variable is zero, and is non-zero for (almost) any non-Gaussian random variable. As a result, by maximizing the absolute value of the kurtosis we can maximize non-Gaussianity. While kurtosis offers a simple approach to measuring non-Gaussianity, it has been established that it is not robust as a basis for numerical optimization [20].

A more practical approach is to measure Gaussianity by calculating the entropy of the sources since a Gaussian random variable has the greatest entropy of all random variables of equal variance [20]. Entropy (or negentropy to be precise) is the fundamental independence metric used in deriving the FastICA algorithm (see Section 3.6.1). Once an appropriate

measure of non-Gaussianity has been established (kurtosis and entropy are two possibilities) then this measure can be used to define a cost function which when minimised yields the independent components, $s_i(t)$.

This section provided a broad overview of the development of ICA algorithms using non-Gaussianity as an independence metric. Details of specific ICA algorithms are given in Section 3.6.

3.3 BSS Assumptions

Blind Source Separation is distinguished from other approaches to source separation in that it requires relatively few assumptions on the sources and on the mixing process. The assumptions essential to BSS are discussed here [20]:

1. The sources being considered are statistically independent
2. The independent components have non-Gaussian distribution
3. The mixing matrix is invertible

The first assumption is fundamental to ICA. As discussed in Section 3.2, statistical independence is the key feature that enables estimation of the independent components $y_i(t)$ from the observations $x_i(t)$.

The second assumption is necessary because of the close link between Gaussianity and independence. It is impossible to separate Gaussian sources using the ICA framework described in Section 3.2 because the sum of two or more Gaussian random variables is itself Gaussian [21]. That is, the sum of Gaussian sources is indistinguishable from a single Gaussian source in the ICA framework, and for this reason Gaussian sources are forbidden. This is not an overly restrictive assumption as in practice most sources of interest are non-Gaussian.

The third assumption is straightforward. If the mixing matrix is not invertible then clearly the un-mixing matrix we seek to estimate does not even exist. If these three assumptions are satisfied, then it is possible to estimate the independent components modulo some trivial

ambiguities (discussed in Section 3.4). It is clear that these assumptions are not particularly restrictive and as a result we need only very little information about the mixing process and about the sources themselves.

3.4 BSS Ambiguities

There are two inherent ambiguities in the ICA framework. Firstly there is a scaling ambiguity. Since both s and $\boldsymbol{A}$ are unknown, any scalar multiplier of s_i could be cancelled by dividing the i^{th} column of $\boldsymbol{A}$ by the same factor.

This is expressed in the following equation:

$$x_i = \sum_i \left(\frac{1}{\beta_i} a_i\right)(s_i \beta_i) \tag{3.7}$$

where β_i is some arbitrary constant. Due to Eq. (3.7), it is impossible to know the variances of the independent components. Since the variance is arbitrary, ICA algorithms can only estimate independent components up to some arbitrary scaling factor.

Secondly, there is a permutation ambiguity in BSS in that the source estimate vector $\boldsymbol{y}$ will be an arbitrary permutation of $\boldsymbol{s}$. This can be stated formally as a permutation matrix, $\boldsymbol{P}$, and its inverse, $\boldsymbol{P}^{-1}$, can be substituted into Eq.(3.1) to give:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{s} \tag{3.8}$$

Here the elements of $\boldsymbol{P}\boldsymbol{s}$ are the original sources, except in a different order, and $\boldsymbol{A}\boldsymbol{P}^{-1}$ is another unknown mixing matrix. Eq. (3.8) is indistinguishable from Eq.(3.1) within the ICA framework, demonstrating that the permutation ambiguity is inherent to Blind Source Separation. This ambiguity is to be expected – in separating the sources we do not seek to impose any restrictions on the order of the separated signals. Thus all permutations of the sources are equally valid.

3.5 Preprocessing

Before examining specific ICA algorithms, it is instructive to discuss preprocessing steps that are generally carried out before ICA.

3.5.1 Centering

A simple preprocessing step that is commonly performed is to "center" the observation vector $\boldsymbol{x}$ by subtracting its mean vector $\boldsymbol{m} = E\{\boldsymbol{x}\}$. That is then we obtain the centered observation vector, $\boldsymbol{x_c}$, as follows:

$$\boldsymbol{x_c} = \boldsymbol{x} - \boldsymbol{m} \tag{3.9}$$

This step simplifies ICA algorithms by allowing us to assume a zero mean. Once the un-mixing matrix has been estimated using the centered data, we can obtain the actual estimates of the independent components as follows:

$$\boldsymbol{y} = \boldsymbol{A}^{-1}(\boldsymbol{x_c} + \boldsymbol{m}) \tag{3.10}$$

From this point on, all observation vectors will be assumed centered.

3.5.2 Whitening

Another step which is very useful in practice is to pre-whiten the observation vector $\boldsymbol{x}$. Whitening involves linearly transforming the observation vector such that its components are uncorrelated and have unit variance [20]. Let $\boldsymbol{x_w}$ denote the whitened vector, then it satisfies the following equation:

$$E\{\boldsymbol{x_w}\boldsymbol{x_w^T}\} = \boldsymbol{I} \tag{3.11}$$

where $E\{\boldsymbol{x_w}\boldsymbol{x_w^T}\}$ is the covariance matrix of $\boldsymbol{x_w}$. Also, since the ICA framework is insensitive to the variances of the independent components, we can assume without loss of generality that the source vector, $\boldsymbol{s}$, is white, i.e. $E\{\boldsymbol{s}\boldsymbol{s^T}\} = \boldsymbol{I}$.

A simple method to perform the whitening transformation is to use the eigenvalue decomposition (EVD) [20] of $\boldsymbol{x}$. That is, we decompose the covariance matrix of $\boldsymbol{x}$ as follows:

$$E\{\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}} \tag{3.12}$$

where $\boldsymbol{V}$ is the matrix of eigenvectors of $E\{\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\}$, and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues, i.e. $\boldsymbol{D} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$.

The observation vector can be whitened by the following transformation:

$$E\{\boldsymbol{x}_{\boldsymbol{w}}\} = \boldsymbol{D}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}}\boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{x} \tag{3.13}$$

We can confirm that this yields a whitened vector $\boldsymbol{x}_{\boldsymbol{w}}$:

$$\begin{aligned} E\{\boldsymbol{x}_{\boldsymbol{w}}\boldsymbol{x}_{\boldsymbol{w}}^{\boldsymbol{T}}\} &= \boldsymbol{D}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}}\boldsymbol{V}^{\boldsymbol{T}}(E\{\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}}\})\,\boldsymbol{V}\,\boldsymbol{D}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}} \\ &= \boldsymbol{D}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}}\boldsymbol{V}^{\boldsymbol{T}}\,(\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{\boldsymbol{T}})\boldsymbol{V}\,\boldsymbol{D}^{-\frac{\boldsymbol{1}}{\boldsymbol{2}}} \\ &= \boldsymbol{I} \end{aligned} \tag{3.14}$$

since $\boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{I}$ as the matrix of eigenvectors is orthogonal.

By pre-whitening the data, we transform the mixing matrix $\boldsymbol{A}$ into a new matrix $\boldsymbol{A}_{\boldsymbol{w}}$. The usefulness of whitening is that the new mixing matrix is orthogonal, as shown by the following chain of equations:

$$\boldsymbol{I} = E\{\boldsymbol{x}_{\boldsymbol{w}}\boldsymbol{x}_{\boldsymbol{w}}^{\boldsymbol{T}}\} = \boldsymbol{A}_{\boldsymbol{w}}E\{\boldsymbol{s}\boldsymbol{s}^{\boldsymbol{T}}\}\boldsymbol{A}_{\boldsymbol{w}}^{\boldsymbol{T}} = \boldsymbol{A}_{\boldsymbol{w}}\boldsymbol{I}\boldsymbol{A}_{\boldsymbol{w}}^{\boldsymbol{T}} = \boldsymbol{A}_{\boldsymbol{w}}\boldsymbol{A}_{\boldsymbol{w}}^{\boldsymbol{T}} \tag{3.15}$$

An orthogonal matrix contains only $n(n-1)/2$ degrees of freedom compared to n^2 degrees of freedom for an unconstrained matrix. As a result, pre-whitening the vector $\boldsymbol{x}$ effectively reduces the number of parameters that need to be estimated by ICA by about half. This is a very useful step as whitening is a simple and efficient process that significantly reduces the computational complexity of ICA.

An example of the effect of ICA on the joint statistical distribution of two signals is given in Section 3.7. The observed signals in Fig. 3.6 are not independent, and the result of applying ICA yields two independent signals as shown in Fig. 3.7 (see Section 3.7). The joint distribution that results from whitening the signals of Fig. 3.6 is shown in Fig. 3.2. Whitening is an intermediate step before ICA is applied, and by comparing the whitened data of Fig. 3.2

with Fig. 3.7, we can see that, in this case, pre-whitening reduces ICA to finding an appropriate rotation to yield independence. This is a simplification as a rotation is an orthogonal transformation which requires only one parameter.

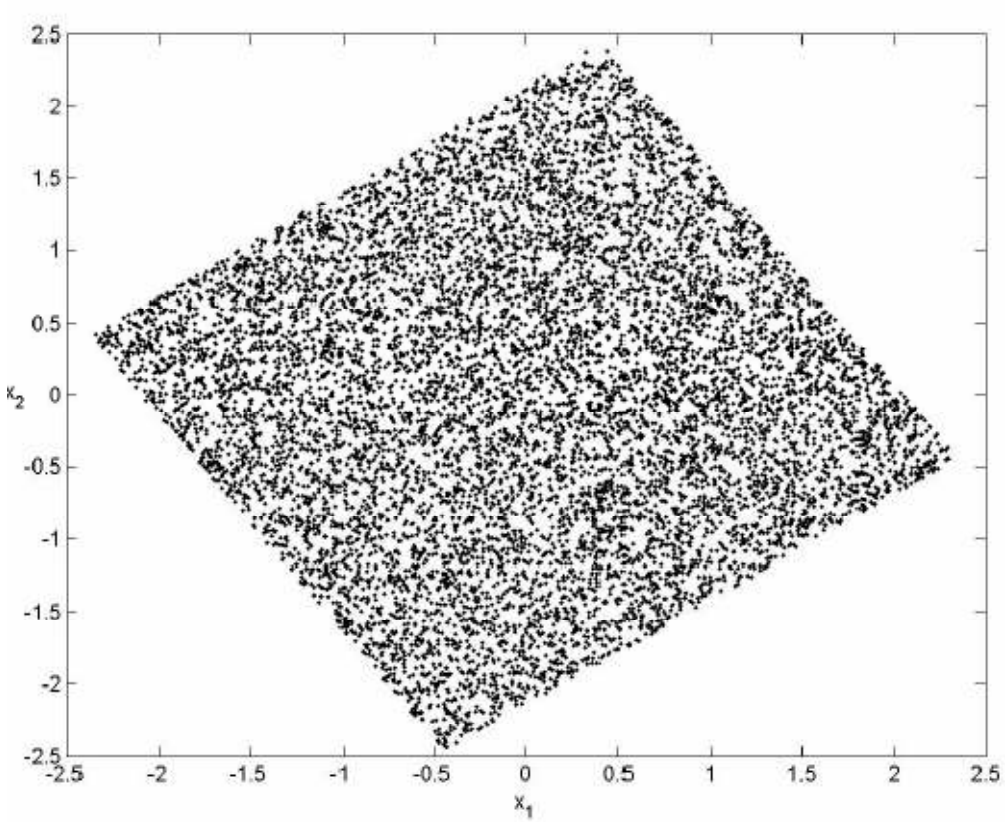

Fig. 3.2. *Joint density of whitened signals obtained from whitening the signals.*

From this point on all observation vectors will be assumed centered and whitened.

3.6 ICA Algorithms

In this section, two of the most important ICA algorithms are presented in some detail. Oja and Hyvarinen's FastICA algorithm is presented in Section 3.6.1 and Bell and Sejnowski's information maximization algorithm in Section 3.6.2. These two algorithms are probably the most widely used and they each illustrate the important principles of ICA.

3.6.1 Fast ICA

Oja and Hyvarinen's FastICA algorithm is described in [22]. The algorithm is based on using "non-Gaussianity" as a metric for independence as discussed in Section 3.2. FastICA is based on using entropy as a measure of non-Gaussianity. A fundamental result of information theory is that a Gaussian random variable has the greatest entropy of all random variables of equal variance. As a result, entropy can be used as a measure of non-Gaussianity. To be precise, FastICA is not based on entropy, but rather on negentropy. Negentropy is a related concept defined by:

$$J(y) = H(y_{gauss}) - H(y) \tag{3.16}$$

where $H(\cdot)$ denotes the entropy of a random variable, $J(\cdot)$ denotes negentropy and y_{gauss} is a Gaussian random vector with the same covariance matrix as y. Negentropy is always non-negative since $H(y_{gauss}) \geq H(y)$.

FastICA is a fixed point algorithm that maximizes negentropy using Newton's iterative method, for details of the derivation of FastICA refer to [22]. The FastICA algorithm is given below. The algorithm determines the unmixing matrix one column at a time, with the update rule for each column defined by:

$$w^{+} = E\{xg(w^{T}x)\} - E\{g'(w^{T}x)\}w \tag{3.17}$$

The function $g(.)$ can be almost any non-quadratic function, but hyperbolic tangent functions have been shown to behave well in practice [22].

3.6.2 Bell and Sejnowski

Bell and Sejnowski's algorithm (henceforth referred to as the BS algorithm) is based on the neural network principle of Information Maximization [5] but is essentially one of the family of maximum-likelihood (ML) ICA algorithms. That is, independence is maximised by estimating the un-mixing matrix that maximises the probability of the observation vector x.

Based on Eq.(3.1), we can find the probability density of the observation vector, x, in terms of the density of s:

$$p_x(x) = |\det \boldsymbol{W}| p_s(s) = |\det \boldsymbol{W}| \prod_i p_i(s_i) \tag{3.18}$$

where $\boldsymbol{W} = \boldsymbol{A}^{-1}$, and $p_i(s_i)$ denotes the density of the i^{th} component of s. The second equality in Eq.(3.18) follows because of the independence of the components of s.

Let us assume we have a vector w_i (as discussed in Section 3.2) that satisfies $w_i^T x = s_i$. We can then rewrite Eq.(3.18) as:

$$p_x(x) = |\det \boldsymbol{W}| \prod_i p_i(w_i^T x) \tag{3.19}$$

This expression can be used to define a likelihood for the un-mixing matrix $\boldsymbol{W}$ since for a given observation vector x we can determine the vector w_i that maximises the likelihood that x would be observed. Suppose we have T observations of x, then we can estimate the likelihood of the un-mixing matrix $\boldsymbol{W}$ with the log-likelihood expression:

$$\log \mathrm{L}(\boldsymbol{W}) = \sum_{t=1}^{T} \sum_{i=1}^{n} log\ p_i\left(w_i^T w(t)\right) + T \log |\det \boldsymbol{B}| \tag{3.20}$$

Bell and Sejnowski's algorithm maximizes this likelihood expression by performing gradient ascent. For details of the derivation of the BS algorithm see [4], but the actual algorithm is given here:

$$\Delta \boldsymbol{W} \propto [\boldsymbol{W^T}]^{-1} + E\{g(\boldsymbol{W}x)x^T\} \tag{3.21}$$

Technically this algorithm is only semi-blind because the derivation requires that the non-linear activation function $g(.)$ must approximate the CDF of s, thus some assumptions must be made about the distribution of each independent component, s_i. It is fortunate, however, that this algorithm is quite insensitive to the accuracy of this approximation and that in most cases of interest the independent components are Gaussian-like in nature. Hyperbolic tangent and logistic functions are good approximations of the CDF of Gaussian-like random variables and either can be used effectively for g_i.

3.7 Simple Illustrations of ICA

To clarify the concepts discussed in the preceding sections two simple illustrations of ICA are presented here (more advanced applications of ICA are examined in subsequent chapters). The results presented below were obtained using the FastICA algorithm of Section 3.6.1 but could equally well have been obtained from any of the numerous ICA algorithms that have been published in the literature (including the BS algorithm).

3.7.1 Separation of Two Signals

In this illustration two independent signals, s_1 and s_2, are generated. These signals are shown in Fig. 3.3. The independent components are then mixed according to Eq.(3.1) using an arbitrarily chosen mixing matrix A, where

$$A = \begin{bmatrix} -0.3784 & 0.8537 \\ 0.8600 & 0.5936 \end{bmatrix} \tag{3.22}$$

The resulting signals from this mixing are shown in Fig. 3.4. Finally, the mixtures x_1 and x_2 are separated using ICA to obtain y_1 and y_2, shown in Fig. 3.5. By comparing Fig. 3.5 to Fig. 3.3, it is clear that the independent components have been estimated accurately and that the independent components have been estimated without any knowledge of the components themselves or the mixing process. This example also provides a clear illustration of the scaling and permutation ambiguities discussed in Section 3.4. The amplitudes of the corresponding waveforms in Figs. 6 and 8 are different and the sawtooth waveform in Fig. 3.5 has been reflected vertically with respect to the sawtooth waveform in Fig. 3.3. Thus the estimates of the independent components are some multiple of the independent components of Fig. 3.3, and in the case of s_1, the scaling factor is negative. The permutation ambiguity is also demonstrated as the order of the independent components has been reversed between Fig. 3.3 and Fig. 3.5.

3.7.2 Illustration of Statistical Independence in ICA

The previous example was a simple illustration of how ICA is used; we start with mixtures of signals and use ICA to separate them. However, this gives no insight into the mechanics of ICA and the close link with statistical independence. The statistical basis of ICA is illustrated more clearly in this example. We assume that the independent components can be modelled

as realizations of some underlying statistical distribution at each time instant (e.g. a speech signal can be accurately modelled as having a Laplacian distribution [28]). One way of visualizing ICA is that it estimates the optimal linear transform to maximise the independence of the joint distribution of the signals x_i.

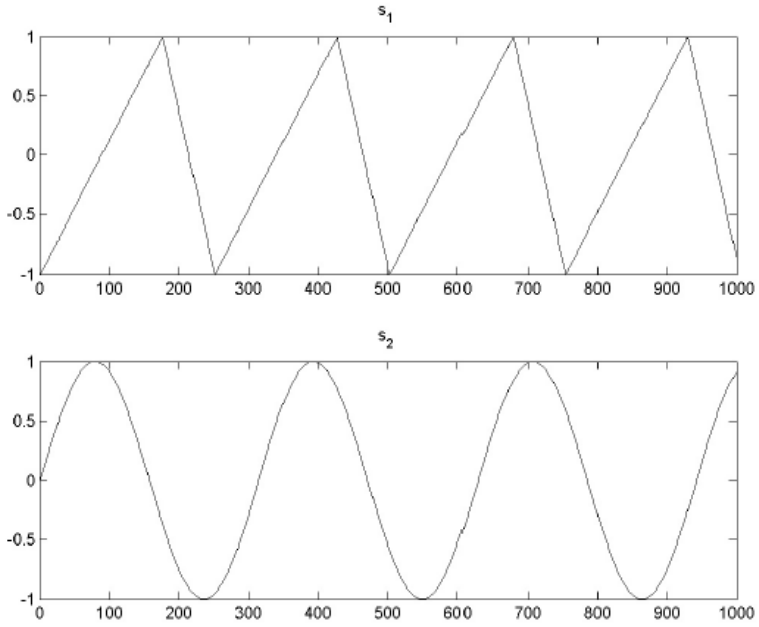

Fig. 3.3. *Independent components s1 and s2.*

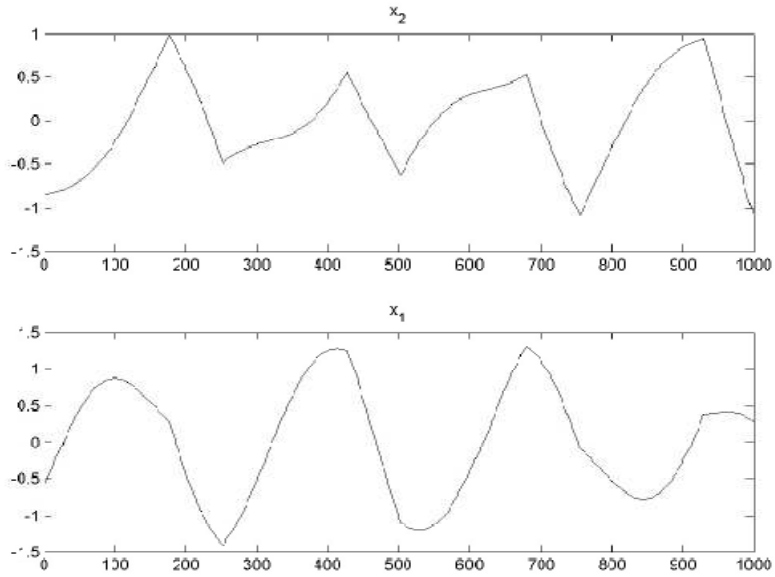

Fig. 3.4. *Observed signals, x1 and x2, from an unknown linear mixture of unknown*

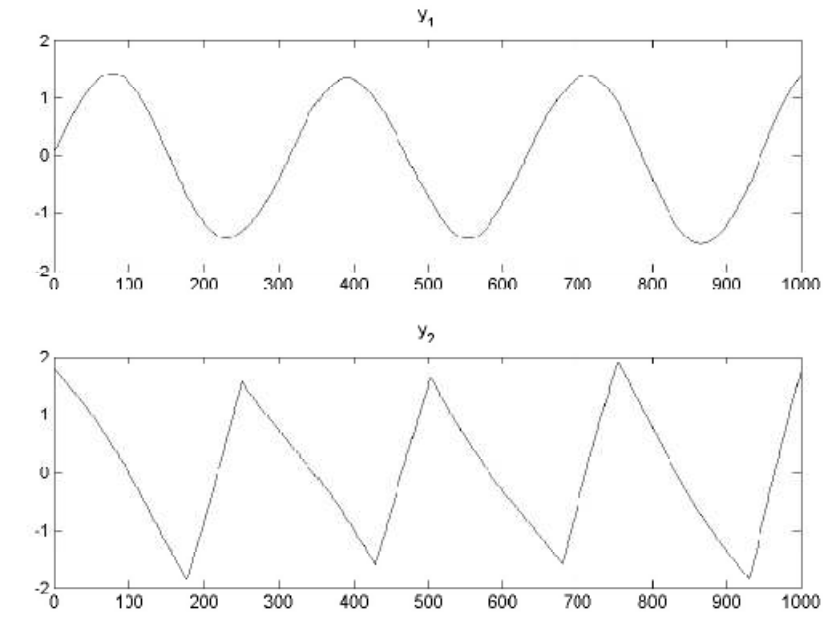

Fig. 3.5. *Estimates of independent components.*

Suppose we have the joint probability distribution for the observed signals x_1 and x_2 shown in Fig. 3.6. From the figure it is clear that the two signals are not statistically independent because, for example, if $x_1 = 0 \; or \; 3$ then x_2 is totally determined. By applying ICA, we seek to transform the data such that we obtain two independent components.

The joint distribution resulting from applying ICA to x_1 and x_2 is shown in Fig. 3.7. This is clearly the joint distribution of two independent, uniformly distributed random variables. Independence can be intuitively confirmed as each random variable is unconstrained regardless of the value of the other random variable (this is not the case for x_1 and x_2). The uniformly distributed random variables in Fig. 3.7 take values between 0 and -4, but due to the scaling ambiguity, we do not know the range of the original independent components.

The two examples in this section are simple but they illustrate both how ICA is used and the statistical underpinnings of the process. The power of ICA is that an identical approach can be used to address problems of much greater complexity.

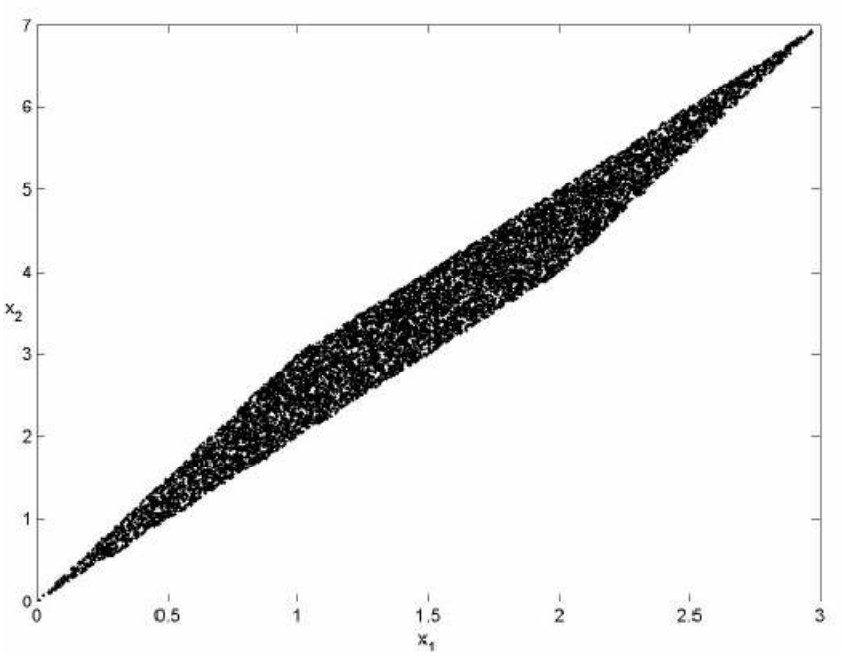

Fig. 3.6. *Joint density of observed signals x1 and x2 obtained from an unknown linear.*

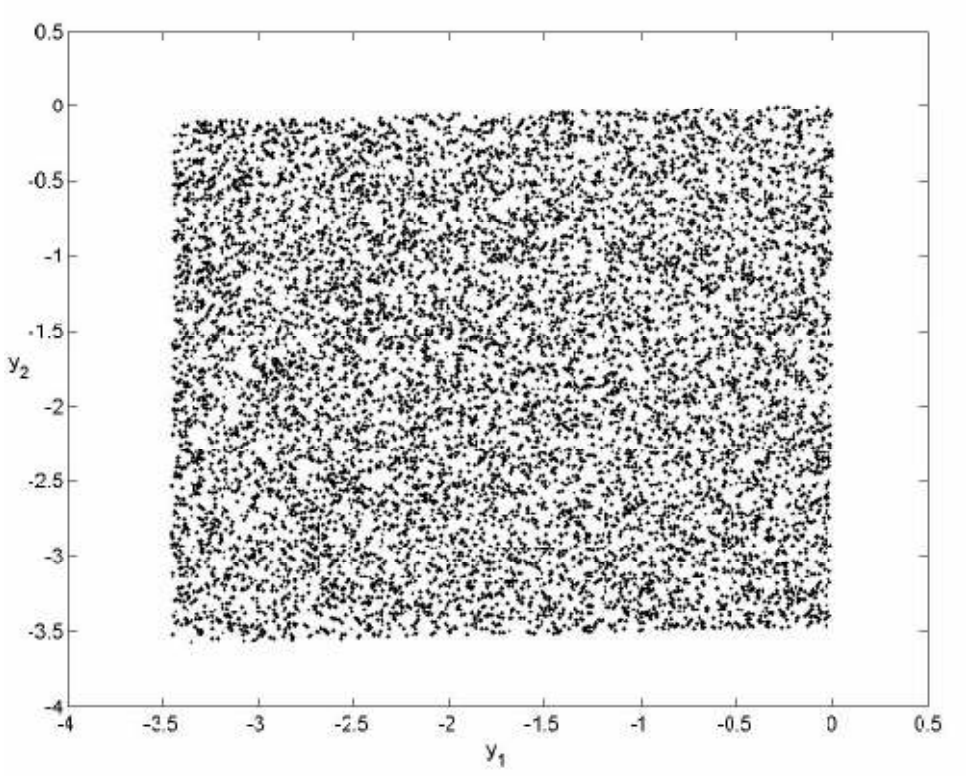

Fig. 3.7. *Joint density of estimates of independent components.*

3.8 Conclusions

Blind Source Separation (BSS) is a statistical approach to separating individual signals from an observed mixture of a group of signals. BSS relies on only very weak assumptions on the signals and the mixing process (hence the "blind" descriptor) and this blindness enables the technique to be used in a wide variety of situations. Research in the field of Blind Source Separation has resulted in the development of a family of algorithms, known as Independent Component Analysis (ICA) algorithms, which can reliably and efficiently achieve blind separation of signals.

This chapter has introduced the fundamentals of Blind Source Separation. The mathematical framework of the source mixing problem that BSS addresses was examined in some detail, as was the general approach to solving BSS. As part of this discussion, some inherent ambiguities of the BSS framework were examined as well as the two important pre-processing steps of centering and whitening. Finally, specific details of the approach to solving the mixing problem of Section 3.1 were presented and two important ICA algorithms were discussed in detail. The material covered in this chapter is important not only to understand the algorithms used in this thesis to perform BSS, but it also provides the necessary background to understand extensions to the framework of Section 3.1 in subsequent chapters.

CHAPTER 4
MIXTURE MODELLING

4.1 Introduction

In the time domain, the relation between the mixture and the sources is assumed to be linear, that is

$$\boldsymbol{y}(t) = \boldsymbol{x_1}(t) + \boldsymbol{x_2}(t), \;\; t = 1, \ldots\ldots\ldots\ldots, T \tag{4.1}$$

where, $\boldsymbol{x_i}(t)$, $i = \{1,2\}$ is the speech source signal from the i^{th} source for time $\boldsymbol{T}$.

In the time domain, since the relation between the source and observation is linear, the MMSE is easily computed and leads to a Wiener filter whose parameters are controlled by the power spectra of the sources.

A challenging topic in modern speech processing applications is to recognize, separate, or enhance a target speech signal when it is mixed with a non-stationary noise, music, or the speech signal of another speaker. In these applications, usually the log amplitude of the short-time Fourier transform is used as a primary feature for later processing. In this case, the objective is to express the log spectrum of the mixed signal in terms of the log spectra of the underlying speech signals.

Considering $\boldsymbol{x_1}, \boldsymbol{x_2}$ and $\boldsymbol{y}$ as the first speaker, second speaker and mixed real signals, respectively the related frequency spectra can be shown as follows:

$$\begin{aligned} \boldsymbol{X_1} &= [x_1(1), \ldots\ldots\ldots, x_1(k), \ldots\ldots\ldots\ldots, x_1(N_{DFT})]^T \\ \boldsymbol{X_2} &= [x_2(1), \ldots\ldots\ldots, x_2(k), \ldots\ldots\ldots\ldots, x_2(N_{DFT})]^T \\ \boldsymbol{Y} &= [y(1), \ldots\ldots\ldots, y(k), \ldots\ldots\ldots\ldots, y(N_{DFT})]^T \end{aligned} \tag{4.2}$$

where $(.)^{T}$ denotes the transpose operator, k is the frequency bin index and N_{DFT} is the number of Discrete Fourier transforms (DFT) points. The mixture value $y(k)$ is then related to $x_1(k)$ and $x_2(k)$ by the following relationship:

$$|y(k)|e^{j\varphi_y(k)} = |x_1(k)|e^{j\varphi_1(k)} + |x_2(k)|e^{j\varphi_2(k)} \quad (4.3)$$

Where $|x_i(k)|$ and $\varphi_i(k), i = 1, 2$, represent the magnitude and the phase values of the N_{DFT} point DFT corresponding to the i^{th} speaker, respectively. $\varphi_y(k)$ is mixture phase. The relation between the log spectrum of the mixed signal and log spectra of the underlying signals is expressed by

$$log_{10}(|y(k)|) = \tfrac{1}{2}\, log_{10}\{|x_1(k)|^2 + |x_2(k)|^2 + 2|x_1(k)|.|x_2(k)|.cos\,(\theta(k))\} \quad (4.4)$$

Where $k = 1,2,\ldots\ldots,K$, and $\theta(k)$, the difference between phases given as

$$\theta(k) = \varphi_1(k) - \varphi_2(k) \quad (4.5)$$

Various approximations for the mixed signal were devised to estimate the mixed signal in the frequency domain, in terms of log spectra of the signals mixed linearly in time domain. The various approximations were:-

1. **Mixture-Maximization (MIXMAX) Estimator**
2. **Quadratic Estimator**
3. **Optimum Estimator**

4.2 MIXMAX Estimator

Equation of $|y(k)|$ above implies that we need the phase information in addition to the log spectra of the underlying signals to construct the log spectrum of the mixed signal. Nonetheless, in most speech processing techniques we do not have access to the phase information of individual speakers. This is mainly because presenting a compact model for phase values is a difficult task, usually with unsatisfactory results. Nadas *et al.* [24] first encountered this problem in the context of robust speech recognition. To exclude the phase

information from further processing, Nadas *et al.* [24] visually observed that the log spectrum of the mixed signal is nearly the element-wise maximum of the log spectra of the two underlying signals. Thus, they proposed a so-called mixture-maximisation (**MIXMAX**) approximation which can be mathematically expressed as follows:

$$log_{10}(|y_{MM}(k)|) \approx max\,(log_{10}(|x_1(k)|, log_{10}(|x_2(k)|)), \quad k = 1,2,\ldots\ldots\ldots\ldots K \quad (4.6)$$

where $max(.\,,.)$ returns the larger element. This approximation has been used in many speech enhancement and separation techniques, following Nadas's paper. This approximation is in fact a nonlinear minimum mean square error (MMSE) estimator with the assumption of uniform distributions for phase information of the underlying speech signals [25].

The MIXMAX approximation can be viewed from another perspective in which we wish to obtain an estimate of $log_{10}(|y(k)|)$ in terms of $log_{10}(|x_1(k)|)$ and $log_{10}(|x_2(k)|)$, such that the mean square value of the estimator error is minimised. This estimator is given by

$$\alpha(k) = E\{log_{10}(|y(k)|) \mid log_{10}(|x_1(k)|), log_{10}(|x_2(k)|)\} \quad k = 1,2,\ldots\ldots\ldots\ldots, K \quad (4.7)$$

where $\alpha(k)$ denotes the optimum estimate of $log_{10}(|y(k)|)$ in a minimum mean square sense [26] and $E\{.\,|\,.\}$ represents the conditional expectation. Since we assume that $log_{10}(|x_1(k)|)$ and $log_{10}(|x_2(k)|)$ are given, then $log_{10}(|y(k)|)$ is only a function of the random variable $y(k)$. From the theory of random processes, we know that [26]

$$E\{g(v)\} = \int_{-\infty}^{\infty} g(v) f_v(v) dv \quad (4.8)$$

where $g(.)$ is an arbitrary function and v is a random variable. Therefore, if we denote the probability density function (PDF) of θ by $f_\theta(\theta(k))$, the estimator can be expressed by

$$\alpha(k) = \int_{-\pi}^{\pi} log_{10}|y(k)| \times f_\theta\big(\theta(k)\big) d\theta(k) \quad (4.9)$$

Then, we express $f_\theta(\theta(k))$ in terms of $f_{\varphi_1}(\varphi_1(k))$ and $f_{\varphi_2}(\varphi_2(k))$ in the following way. From Eq. (2) and recognising that $\varphi_1(k)$ and $\varphi_2(k)$ are independent, we conclude that

$$f_{\theta}(\theta(k)) = f_{\varphi_1}(\varphi_1(k)) * f_{\varphi_2}(\varphi_2(k)) \tag{4.10}$$

where $*$ denotes the convolution operator. The PDF for the phase $\varphi_i(k)$ of the short-time Fourier transform of the speech signal is usually modelled with a uniform distribution, i.e. $f_{\varphi_i}(\varphi_i(k)) = 1/2\pi$, where $\varphi_i(k) \in [-\pi, \pi]$ and $i \in \{1, 2\}$. Convolving PDFs and bearing in mind the periodic characteristics of phase values, the PDF of the phase difference $\theta(k)$ is obtained as a uniform distribution, i.e. $f_{\theta}(\theta(k)) = 1/2\pi$ where $\theta(k) \in [-\pi, \pi]$. Substituting Eq. (4.4) into Eq. (4.9) and using the obtained PDF for $\theta(k)$ we arrive at

$$\alpha(k) = \frac{1}{4\pi} \int_{-\pi}^{\pi} log_{10}(|x_1(k)|^2 + |x_2(k)|^2 + 2.|x_1(k)|.|x_2(k)|.\cos\theta(k)) d\theta(k) \tag{4.11}$$

Using an integration table, such as that of [27], gives the integration result of

$$\alpha(k) = \begin{cases} log_{10}(|x_1(k)|) & log_{10}(|x_1(k)|) > log_{10}(|x_2(k)|) \\ log_{10}(|x_2(k)|) & log_{10}(|x_2(k)|) > log_{10}(|x_1(k)|) \end{cases} \tag{4.12}$$

This result is the same as that of Eq. (4.6). In this way, we show that the MIXMAX approximation is a nonlinear MMSE estimator of the log spectrum of the mixed signal given the log spectra of the underlying speech signals with this assumption that phase values are modelled using a uniform distribution.

Fig. 4.1 *MixMax employs Logarithm/Anti-Logarithm Functions.*

4.3 Quadratic Estimator

The other estimator to be used for evaluation purposes is the Quadratic estimator, $|y_Q(k)|$ which can be formulated as,

$$\left|y_Q(k)\right| \;=\; \sqrt{|x_1(k)|^2 \;+\; |x_2(k)|^2} \tag{4.13}$$

where $|y_Q(k)|$ denotes the resulting mixture approximation obtained by Quadratic estimator which is closely related to *Ideal Ratio Mask* (IRM) as in [28].

4.4 Optimum Estimator

Although, the MixMax has previously been found well for speech enhancement application first proposed by Nadas *et al.* [24] for speech recognition, it can be observed that this estimation results in significant estimation error while employed for the SCSS framework. This is mainly due to the fact that, here in the separation scenario the final goal is to reconstruct both the underlying signal and not only to enhance a signal in presence of another speaker only by weakening or suppressing the jammer signal as is in the case of speech enhancement. In addition, as another problem, MixMax fails to separate mixtures when frequencies of speakers inter-modulate into each other, especially, when the underlying speakers have comparable magnitudes with respect to each other.

Moreover, since at each frequency band the weaker log spectral amplitude is masked by the stronger one, as a result it is not inherently possible to construct or survive both underlying signals at the same time from a given mixture which is the key idea of separation. In another phrase, the masked signal can never be correctly approximated while employing the MixMax. By considering all the aforementioned reasons and deficiencies corresponding to MixMax approximation in log-domain, that employing either Optimum or Quadratic estimators in the original domain results in a more accurate separation quality in terms of less estimation error in MSE sense for separation problem [29]. Both MixMax and Quadratic

estimators translate the feature vectors into a nonlinear space namely, log-domain and squared, respectively.

After such non-linear transformations, the estimator finds the best possible candidates for each bin of the underlying signals. Finally, the estimators translate log-amplitude features from log domain back into the original domain (i.e. STFT) and then the MSE error can be manipulated. Opposed to viewpoint that mixture in log-spectral domain is estimated in MMSE or Maximum-Likelihood sense, the sound separation problem can also be considered as the problem in which we wish to obtain an estimate of $|y(k)|$ and not $log_{10}(|y(k)|)$ in terms of $|x_1(k)|$ and $|x_2(k)|$, in minimum mean square sense (MMSE). As a consequence, the absolute estimate for mixture will be [26]:

$$|\hat{y}(k)| = E\{|y(k)| \mid |x_1(k)|, |x_2(k)|, \theta(k) \tag{4.14}$$

Considering $\boldsymbol{x_1}, \boldsymbol{x_2}$ and $\boldsymbol{y}$ as the first speaker, second speaker and mixed real signals, respectively the related frequency spectra can be shown as follows:

$$\boldsymbol{X_1} = [x_1(1), \ldots\ldots\ldots, x_1(k), \ldots\ldots\ldots\ldots, x_1(N_{DFT})]^T$$

$$\boldsymbol{X_2} = [x_2(1), \ldots\ldots\ldots, x_2(k), \ldots\ldots\ldots\ldots, x_2(N_{DFT})]^T$$

$$\boldsymbol{Y} = [y(1), \ldots\ldots\ldots, y(k), \ldots\ldots\ldots\ldots, y(N_{DFT})]^T$$

where $(.)^{\boldsymbol{T}}$ denotes the transpose operator, k is the frequency bin index and N_{DFT} is the number of Discrete Fourier transforms (DFT) points. The mixture value $y(k)$ is then related to $x_1(k)$ and $x_2(k)$ by the following relationship:

$$|y(k)|e^{j\varphi_y(k)} = |x_1(k)|e^{j\varphi_1(k)} + |x_2(k)|e^{j\varphi_2(k)} \tag{4.15}$$

Where $|x_i(k)|$ and $\varphi_i(k), i = 1, 2$, represent the magnitude and the phase values of the N_{DFT} point DFT corresponding to the i^{th} speaker, respectively. $\varphi_y(k)$ is mixture phase. The relation between the log spectrum of the mixed signal and log spectra of the underlying signals is expressed by

$$|y(k)| = \sqrt{|x_1(k)|^2 + |x_2(k)|^2 + 2|x_1(k)|.|x_2(k)|.cos\left(\theta(k)\right)} \quad (4.16)$$

Where $k = 1,2,\ldots\ldots,K$, and $\theta(k)$, the difference between phases given as

$$\theta(k) = \varphi_1(k) - \varphi_2(k)$$

Considering the uniformity assumption for phase PDF, the same assumption used by other estimators, i.e. $f_\theta(\theta(k)) = 1/2\pi$ where $\theta(k) \in [-\pi, \pi]$, the optimum estimator can be expressed as

$$\left|y_{opt}(k)\right| = \int_0^{2\pi} |y(k)| \times f_\theta\left(\theta(k)\right) d\theta(k) \quad (4.17)$$

Replacing $|y(k)|$ from Eq. (4.16) and inserting into Eq. (4.17), we have

$$\left|y_{opt}(k)\right| = \sqrt{|x_1(k)|^2 + |x_2(k)|^2}\, \frac{1}{2\pi} \int_0^{2\pi} \sqrt{1 + \alpha \cos\theta(k)}\; d\theta(k) \quad (4.18)$$

Where k $= 1,\ldots\ldots\ldots\ldots,K$ and K is the number of DFT-bins in the frequency domain, and α is defined as,

$$\alpha = \frac{2|x_1(k)||x_2(k)|}{|x_1(k)|^2 + |x_2(k)|^2} \quad (4.19)$$

As a result of Eq. (4.18), the optimum estimation of mixture, $|y_{opt}|$ consists of two terms, namely, quadratic and integral term. The quadratic term is related to DFT absolute values i.e. $|x_1(k)|$ and $|x_2(k)|$. However, in the following the integral term given in Eq. (4.18) can be calculated as a closed form of the well-known Elliptic series in terms of the underlying spectrum magnitudes related to speakers. To do so, first note since $|x_1(k)|$ and $|x_2(k)|$ are positive values, we obtain $0 < \alpha < 1$ and the second term in Eq. (4.18) can be rewritten as,

$$f(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} \sqrt{1 + \alpha \cos\theta(k)}\; d\theta(k) \quad (4.20)$$

which can numerically be computed for different values of a equally spaced between [0,1] to 1000 bins. In the following the optimum estimation given in Eq. (4.17) can be calculated as a

closed form in terms of the underlying spectrum magnitudes related to each speaker. This closed form can achieve at the optimum estimation for the given mixture with selecting a few first terms of the well-known Elliptic series which results in an acceptable accuracy.

First, considering a as defined above we may write, $c(k) = \frac{\sqrt{|x_1(k)|^2+|x_2(k)|^2}}{2\pi}$ and employing $\cos\theta(k) = 1 - 2sin^2(\frac{\theta(k)}{2})$, inserting into Eq. (6) we obtain for $|y_{opt}(k)|$ given in Eq. (4.18) as below,

$$|\hat{y}(k)| = c(k)\int_0^{2\pi}\sqrt{1+\alpha-2\alpha sin^2(\frac{\theta(k)}{2})}\ d\theta(k) \tag{4.21}$$

And substituting $\phi(k) = \theta(k)/2$ and $d\theta(k) = 2d\phi(k)$, we have

$$\left|\hat{y}_{opt}(k)\right| = 2c(k)\sqrt{1+\alpha}\int_0^{\pi}\sqrt{1-\beta^2 sin^2\phi(k)}d\phi(k)$$

$$\left|\hat{y}_{opt}(k)\right| = \frac{|x_1(k)|+|x_2(k)|}{\pi}\int_0^{\pi}\sqrt{1-\beta^2 sin^2\phi(k)}d\phi(k) \tag{4.22}$$

Where $\beta = \sqrt{\frac{2\alpha}{1+\alpha}}$ and finally optimum mixture estimation can be formulated as,

$$\left|\hat{y}_{opt}(k)\right| = \frac{|x_1(k)|+|x_2(k)|}{\pi}\ E(\beta) \tag{4.23}$$

Where $E(\beta)$ is the complete Elliptic integral of the second kind. Using table of integrals in [27] we have,

$$E(\beta) = \pi\left[1-\left(\frac{1}{2}\right)^2\beta^2-\left(\frac{1\times 3}{2\times 4}\right)^2\frac{\beta^4}{3}-\cdots\cdots\right] \tag{4.24}$$

The Elliptic series can be approximated with a finite number of terms of the following series:

$$E(\beta) = \pi\left\{1-\sum_{n=1}^{\infty}\left[\prod_{k=1}^{n}\left(\frac{2k-1}{2k}\right)^2\right]\frac{\beta^{2n}}{(2n-1)}\right\}$$

MixMax estimation deviates from the optimal estimate; this is due to the fact that it maps input *Discrete-time Fourier Transform* (DFT) amplitude to another space using a logarithm function. Next the estimation is made to find the underlying signals, and after taking inverse log function the result will be found as depicted in Fig. 4.1. Note MSE estimation is also computed in the original space. A similar approach is carried out in order to achieve at the Quadratic estimation but in the STFT domain.

CHAPTER 5
GAIN ADAPTED OPTIMUM MIXTURE ESTIMATOR

5.1 Introduction

Due to the difficult nature of the single channel speech separation problem, previous single channel speech separation techniques [30] – [45], [9] assume that the speech signals used in the modelling (training phase) and estimation (test phase) stages have the same energy level, a pre-requisite which is hardly met in practical situations. This means representing the mixture as

$$Y(t) = g_1 X_1(t) + g_2 X_2(t)$$

with the assumption that $g_1 = g_2 = 1$ and $\sum_t X_1^2(t) = \sum_t X_2^2(t)$.

In this chapter, we consider the general scenario (i.e., $g_1 \neq g_2$), where we solve the separation problem for the case in which the sources are mixed with the different energy levels, an important topic which has received little consideration in previous works. In [33], it has been depicted that g_1 and g_2 can be expressed in terms of the signal-to-signal ratio (SSR). Using this relation, an optimum estimator was derived to estimate the mixture in terms of the sources' distributions and SSRs. The results were compared with other existing estimators, namely MixMax and Quadratic estimator, and it concludes to the fact that the optimum estimator outperforms other estimators for the different SSRs. The experimental results on real speech signals, drawn from the database provided by Cooke *et al.* [46], are in well agreement with the theoretical results framed.

The problem of gain adaptation has also been studied in other speech processing areas such as robust speech recognition [47], speech enhancement [48], and music source separation [49]. However, the gain adaptation strategies in those approaches cannot be applied to the single channel speech separation for the following reasons. In speech recognition or enhancement, the gain adaptation in [47] and [48] is handled by detecting non-speech segments, whereby the noise power is estimated and updated whenever a new non-speech

segment is detected. This strategy is inappropriate in the single channel separation scenario where the power relations between the two speech signals are of interest. The technique for music source separation in [49] is also inappropriate, as it employs one covariance scaling parameter for each Gaussian component and each frame of the signal.

The problem of gain adaptation can be also considered in the feature-based CASA techniques. In these techniques, the target source is recovered more effectively in the high SSRs since the target pitch contour, which is a primary cue for separation, is well-estimated. On the other hand, the estimation of the interference pitch contour remains a challenge since the interference signal is masked by the target signal. Unfortunately, the current multi-pitch trackers fail to assign the pitch contours to the corresponding speakers when two pitch contours cross each other, or are unable to detect one of the pitch contours, if one speaker's pitch frequency lies within a multiple integer of the other.

5.2 Mathematical Description of the Mixture Estimator

The observation signal $Y(t)$ is related to the two speech signal sources by

$$Y(t) = g_1 X_1(t) + g_2 X_2(t), \qquad t = 0, \ldots\ldots\ldots\ldots, T-1 \tag{5.1}$$

where $Y(t), X_1(t)\ and\ X_2(t)$ are considered real signals. $g_1 > 0$ and $g_2 > 0$ in Eq. (5.1) represent the associated source gains, and it is assumed that the signals have equal power before scaling, such that it follows

$$G_o^2 = \frac{1}{T}\sum_{t=1}^{T} X_1^2(t) = \frac{1}{T}\sum_{t=1}^{T} X_2^2(t). \tag{5.2}$$

G_o^2 is an arbitrary, and positive, power parameter attributed to the source signals before they are scaled and mixed to produce the observed signal. Using Eq. (5.1), it can be shown that the power of the observed signal is

$$g_y^2 = \frac{1}{T}\sum_{t=1}^{T} Y^2(t)$$

$$= g_1^2 \frac{1}{T}\sum_{t=1}^{T} X_1^2(t) + g_2^2 \frac{1}{T}\sum_{t=1}^{T} X_2^2(t) + 2g_1 g_2 \frac{1}{T}\sum_{t=1}^{T} X_1(t)\, X_2(t)$$

$$g_y^2 = G_o^2(g_1^2 + g_2^2) + 2g_1 g_2 \frac{1}{T}\sum_{t=1}^{T} X_1(t)\, X_2(t) \tag{5.3}$$

As $X_1(t)$ and $X_2(t)$ are considered to be the two independent zero-mean processes and T is assumed to be large, the second term in Eq. (5.3) vanishes, because $E[X_1(t)X_2(t)] = 0$; where $E[.]$ is the expectation operator. This results in

$$g_y^2 = G_o^2(g_1^2 + g_2^2). \tag{5.4}$$

On the other hand, the Signal-to-Signal Ratio (SSR) is defined by

$$\theta = \frac{g_1^2}{g_2^2}. \tag{5.5}$$

From Eqs. (5.4) and (5.5), it is clear that

$$g_1 = \frac{g_y}{G_o\sqrt{\frac{1+\theta}{\theta}}} \; and \; g_2 = \frac{g_y}{G_o\sqrt{1+\theta}}. \tag{5.6}$$

Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{Y}$ are the first speaker, second speaker and the mixed real signals, respectively. The related frequency spectra can be shown as follows:

$$\boldsymbol{x}_1 = [x_1(1), \dots, x_1(k), \dots, x_1(N_{DFT})]^T$$

$$\boldsymbol{x}_2 = [x_2(1), \dots, x_2(k), \dots, x_2(N_{DFT})]^T$$

$$\boldsymbol{y} = [y(1), \dots, y(k), \dots, y(N_{DFT})]^T$$

where $(.)^T$ denotes the transpose operator, k is the frequency bin index and N_{DFT} is the number of Discrete Fourier Transform (DFT) points. The mixture value $y(k)$ is then related to $x_1(k)$ and $x_2(k)$ by the following relationship using Eq. (5.1)

$$|y(k)|e^{j\varphi_y(k)} = g_1|x_1(k)|e^{j\varphi_1(k)} + g_2|x_2(k)|e^{j\varphi_2(k)} \quad (5.7)$$

where $|x_i(k)|$ and $\varphi_i(k)$, $i = 1,2$, represent the magnitude and the phase values of the N_{DFT} point DFT corresponding to the i^{th} speaker, respectively, and $\varphi_y(k)$ is the mixture phase. The relation between the mix spectrum magnitude and the magnitude spectra of the underlying signals can be expressed by:

$$|y(k)| = \sqrt{g_1^2|x_1(k)|^2 + g_2^2|x_2(k)|^2 + 2g_1g_2|x_1(k)||x_2(k)|\cos\varphi(k)} \quad (5.8)$$

with $\varphi(k)$ the phase difference given as: $\varphi(k) = \varphi_1(k) - \varphi_2(k)$.

The modelling of phase values is a difficult task, hence in order to exclude the phase information from the sound separation scenario, inspiring from Radfar *et al.* [25], [43], a proof for Mixture-Maximization approximation was recently reported, which states that

$$\log_{10}(|y_{MM}(k)|) = \max(log_{10}(g_1|x_1(k)|), log_{10}(g_2|x_2(k)|)) \quad (5.9)$$

where $y_{MM}(k)$ denotes the mixture approximation given by MixMax estimator. Note that due to the monotonicity characteristic of $log(.)$ function, Eq. (5.9) can be formulated as

$$|y_{MM}(k)| = \max(g_1|x_1(k)|, g_2|x_2(k)|) \quad (5.10)$$

In Eq. (5.10), $y_{MM}(k)$ denotes mixture estimation obtained by MixMax, which simply introduces the largest value of two magnitude spectra as the estimated mixture. The other estimator used for our evaluation purposes is the Quadratic estimator, $|y_Q(k)|$, which can be formulated as

$$|y_Q(k)| = \sqrt{g_1^2|x_1(k)|^2 + g_2^2|x_2(k)|^2} \quad (5.11)$$

where, $y_Q(k)$ denotes the resulting mixture estimation obtained by the Quadratic estimator, which is closely related to *Ideal Ratio Mask* (IRM) discussed in [28]. In the following section, we derive the gain adapted optimum mixture estimator for SCSS problem. It is

demonstrated that the proposed estimator results in a lower estimation error in mean square sense for different signal-to-signal ratios.

All previous estimators are based on the uniformity assumption for phase PDF given in [29] i.e., $f\big(\varphi(k)\big) = \frac{1}{2\pi}$, where $\varphi(k) \in [0, 2\pi]$. As a result, we use the same assumption and the estimator can be expressed as follows

$$\big|y_{opt}(k)\big| = \int_0^{2\pi} |y(k)| f_{\varphi}\big(\varphi(k)\big) d\varphi(k) \tag{5.12}$$

By employing the result of $|y(k)|$ from Eq. (5.8), substituting into Eq. (5.12) and using the phase *pdf* assumption results in,

$\big|y_{opt}(k)\big| = \int_0^{2\pi} \sqrt{g_1^2|x_1(k)|^2 + g_2^2|x_2(k)|^2 + 2g_1g_2|x_1(k)||x_2(k)|\cos\varphi(k)}\ \frac{1}{2\pi}\ d\varphi(k)$,
where $k = 1, \ldots\ldots, K$.

Solving further, we obtain,

$$\big|y_{opt}(k)\big| = \frac{\sqrt{g_1^2|x_1(k)|^2 + g_2^2|x_2(k)|^2}}{2\pi} \int_0^{2\pi} \sqrt{1 + \alpha\ \cos\varphi(k)}\ d\varphi(k) \tag{5.13}$$

where $\alpha = \frac{2\, g_1g_2|x_1(k)||x_2(k)|}{g_1^2|x_1(k)|^2 + g_2^2|x_2(k)|^2}$. (5.14)

By using Eq. (5.6) and Eq. (5.14), it can be simplified as

$$\alpha = \alpha(\theta) = \frac{2\sqrt{\theta}\,|x_1(k)||x_2(k)|}{\theta\,|x_1(k)|^2 + |x_2(k)|^2} \tag{5.15}$$

Utilizing results of α and g_1, g_2 from Eq. (5.15) and Eq. (5.6), and substituting them into Eq. (5.13) results in,

$$\big|y_{opt}(k)\big| = \frac{g_y}{G_o 2\pi\sqrt{1+\theta}} \sqrt{\theta\,|x_1(k)|^2 + |x_2(k)|^2} \int_0^{2\pi} \sqrt{1 + \alpha(\theta)\cos\varphi(k)}\ d\varphi(k) \tag{5.16}.$$

Employing $\cos(\theta(k)) = 1 - 2\,sin^2\left(\frac{\theta(k)}{2}\right)$ and inserting into Eq. (5.16), we obtain

$$|y_{opt}(k)| =$$

$$\frac{g_y}{G_o 2\pi\sqrt{1+\theta}}\sqrt{\theta\,|x_1(k)|^2 + |x_2(k)|^2}\int_0^{2\pi}\sqrt{1 + \alpha(\theta) - 2\alpha(\theta)sin^2\left(\frac{\varphi(k)}{2}\right)}\,d\varphi(k).$$

$$= \frac{g_y}{G_o 2\pi\sqrt{1+\theta}}\sqrt{\theta\,|x_1(k)|^2 + |x_2(k)|^2}\sqrt{1 + \alpha(\theta)}\int_0^{2\pi}\sqrt{1 - \beta^2(\theta)sin^2\left(\frac{\varphi(k)}{2}\right)}\,d\varphi(k).$$

where, $\beta(\theta) = \sqrt{\frac{2\,\alpha(\theta)}{1+\alpha(\theta)}}$ (5.17)

Substituting $\alpha(\theta)$ from Eq. (5.15) and inserting it in Eq. (5.17) results in,

$$|y_{opt}(k)| = \frac{g_y}{G_o 2\pi\sqrt{1+\theta}}\left(\sqrt{\theta}\,|x_1(k)| + |x_2(k)|\right)\int_0^{2\pi}\sqrt{1 - \beta^2(\theta)sin^2\left(\frac{\varphi(k)}{2}\right)}\,d\varphi(k) \quad (5.18).$$

Substituting $\emptyset(k) = \frac{\varphi(k)}{2}$ and $d\varphi(k) = 2d\emptyset(k)$, we get,

$$|y_{opt}(k)| = \frac{g_y}{G_o \pi\sqrt{1+\theta}}\left(\sqrt{\theta}\,|x_1(k)| + |x_2(k)|\right)\int_0^{\pi}\sqrt{1 - \beta^2(\theta)sin^2(\emptyset(k))}\,d\emptyset(k) \quad (5.19).$$

Using the table of integrals in [27] we have,

$$\int_0^{\pi}\sqrt{1 - \beta^2(\theta)sin^2(\emptyset(k))}\,d\emptyset(k) = E(\beta(\theta)) = \pi\left\{1 - \left(\frac{1}{2}\right)^2\beta^2(\theta) - \left(\frac{1\times3}{2\times4}\right)^2\frac{\beta^4(\theta)}{3} - \ \ldots\right\}.$$

(5.20)

Using the above equation, the optimum mixture estimate finally results in,

$$\boxed{|y_{opt}(k)| = \frac{g_y}{G_o\pi\sqrt{1+\theta}}\left(\sqrt{\theta}\,|x_1(k)| + |x_2(k)|\right)E(\beta(\theta))}$$

(5.21)

Considering $\theta = 1$, we get $g_1^2 = g_2^2$, and $g_y^2 = 2G_o^2$, and the $|y_{opt}(k)|$ results in,

$$\left|y_{opt}(k)\right|_{\theta=1} = \frac{|x_1(k)| + |x_2(k)|}{\pi} E(\beta)$$

where, $\beta = \sqrt{\frac{2\alpha}{1+\alpha}}$ and $\alpha = \frac{2|x_1(k)||x_2(k)|}{|x_1(k)|^2 + |x_2(k)|^2}$. (5.22)

For $\theta = 1$, the optimum mixture estimate results in the form shown above, which is equivalent to the result in [29]. Compared to MixMax, the derived estimator results in lower estimation error in MSE sense, with the assumption of correct estimation of SSR $'\theta'$. It is demonstrated, both theoretically and experimentally, in simulation results to be presented in the following section.

5.3 Simulation Results

The simulation of the derived optimum mixture estimator with gain adaption was done in two phases. First, the optimum mixture estimator derived was verified theoretically, and compared with the other estimators, such as MixMax and Quadratic estimators. Using the results in the first phase, the derived mixture estimator is worked upon using the real speech signals. The theoretical results are discussed upon in the next section, and then the experimental results are shown.

5.3.1 Theoretical Results for Gain Adapted Optimum Mixture Estimator

In order to evaluate the performance of the derived optimum mixture estimator with other estimators in terms of their resulting Mean Square Error (MSE) for different signal-to-signal ratios (SSRs), computer simulations were conducted. The MSE for the data is given as

$$MSE = (|y(k)| - |y_{est}(k)|)^2$$

where $|y(k)|$ is the mixture signal's DFT magnitude, $|y_{est}(k)|$ is the estimated mixture, and k is the frequency bin of the DFT sequence.

Assuming the underlying speaker DFT magnitudes as $|x_1(k)| = 10$ and $|x_2(k)|$ ranging within $[0, 100]$. The following figures (Fig.5.1 – Fig.5.6) discuss the estimation error of different estimators, including the derived optimum mixture estimator, MixMax and Quadratic estimator, obtained at different SSRs in the range of 0dB to 25dB. As can be depicted from the figures, the optimum mixture estimator results in lower estimation error in comparison with the other estimators, at different SSRs. In addition, for comparable values of the underlying signals' DFT magnitudes, i.e. for $g_1|x_1(k)| \approx g_2|x_2(k)|$, the derived estimator outperforms the other estimators. For the extreme cases, i.e. $g_1|x_1(k)| \ll g_2|x_2(k)|$ and $g_1|x_1(k)| \gg g_2|x_2(k)|$, MixMax and Quadratic estimators reach asymptotically to the derived optimal estimation.

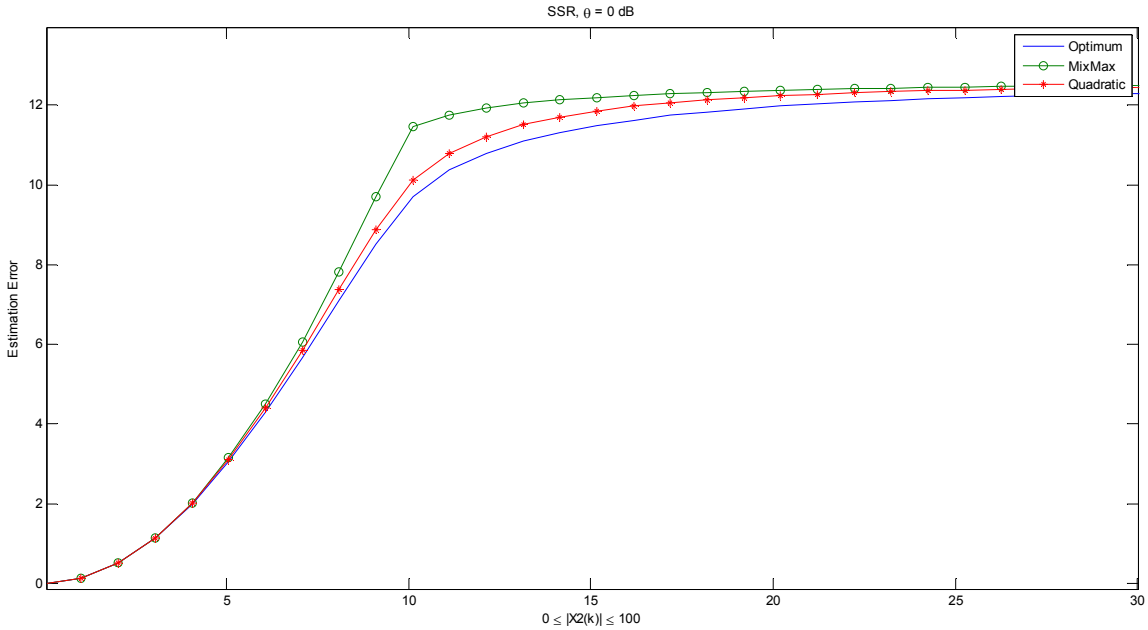

Fig. 5.1. *Estimation Error for different estimators at SSR = 0dB.*

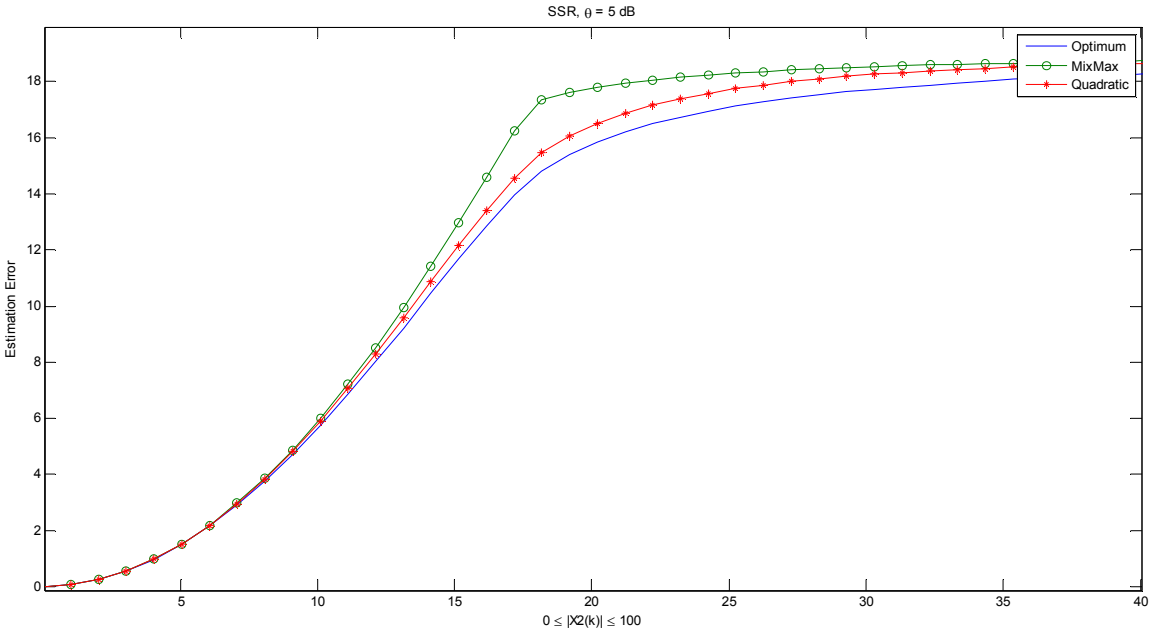

Fig. 5.2. *Estimation Error for different estimators at SSR = 5dB.*

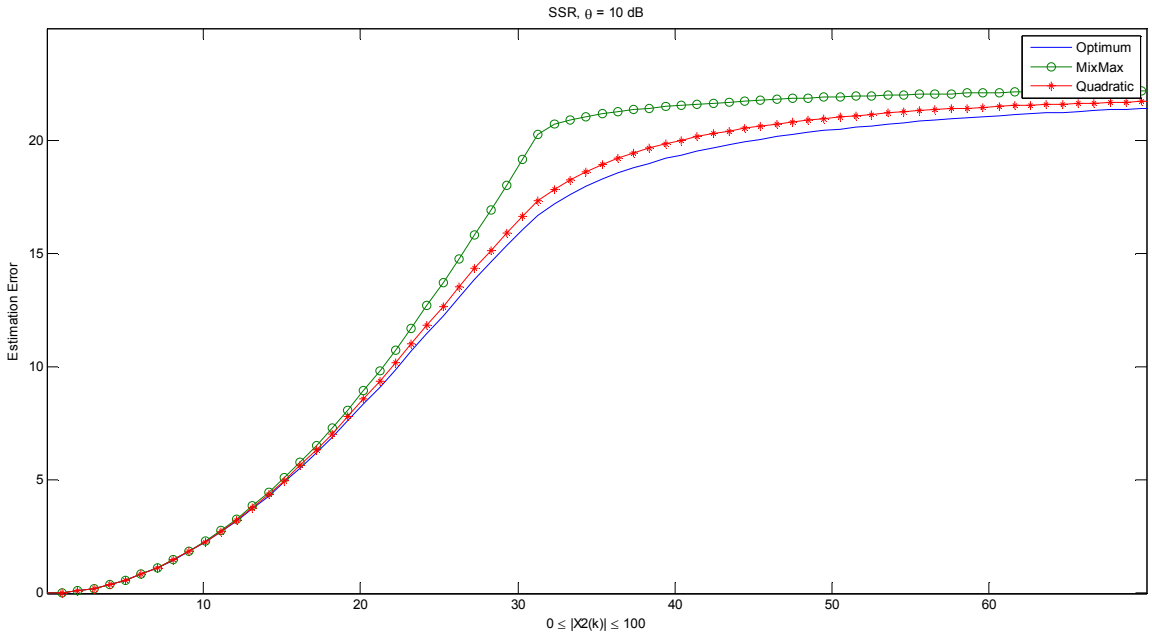

Fig. 5.3. *Estimation Error for different estimators at SSR = 10dB.*

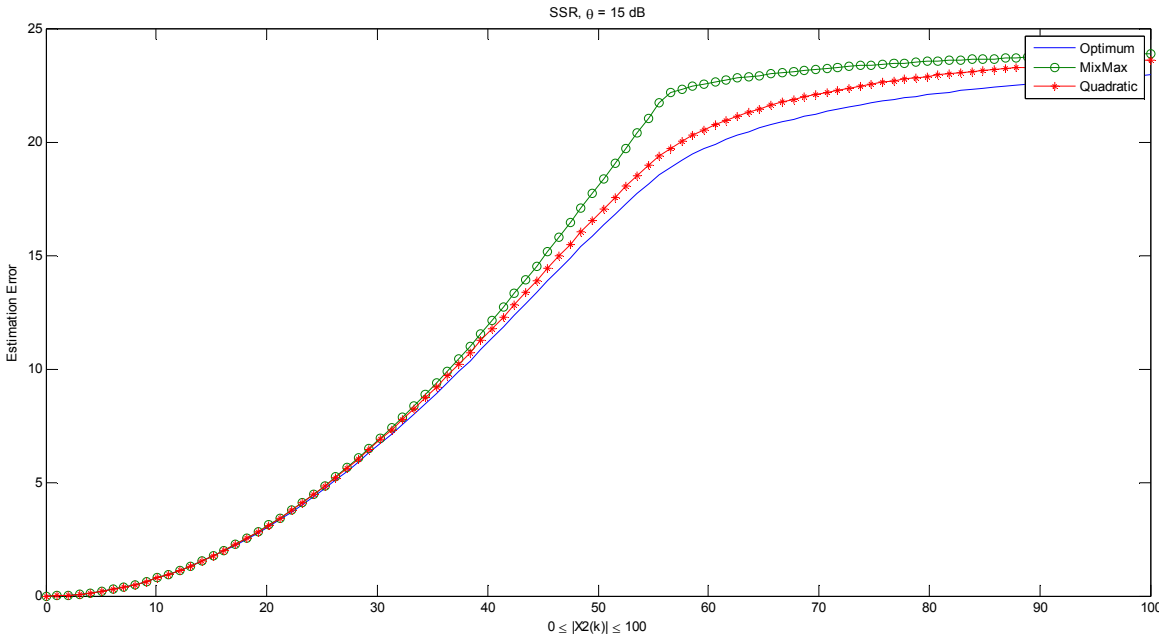

Fig. 5.4. *Estimation Error for different estimators at SSR = 15dB.*

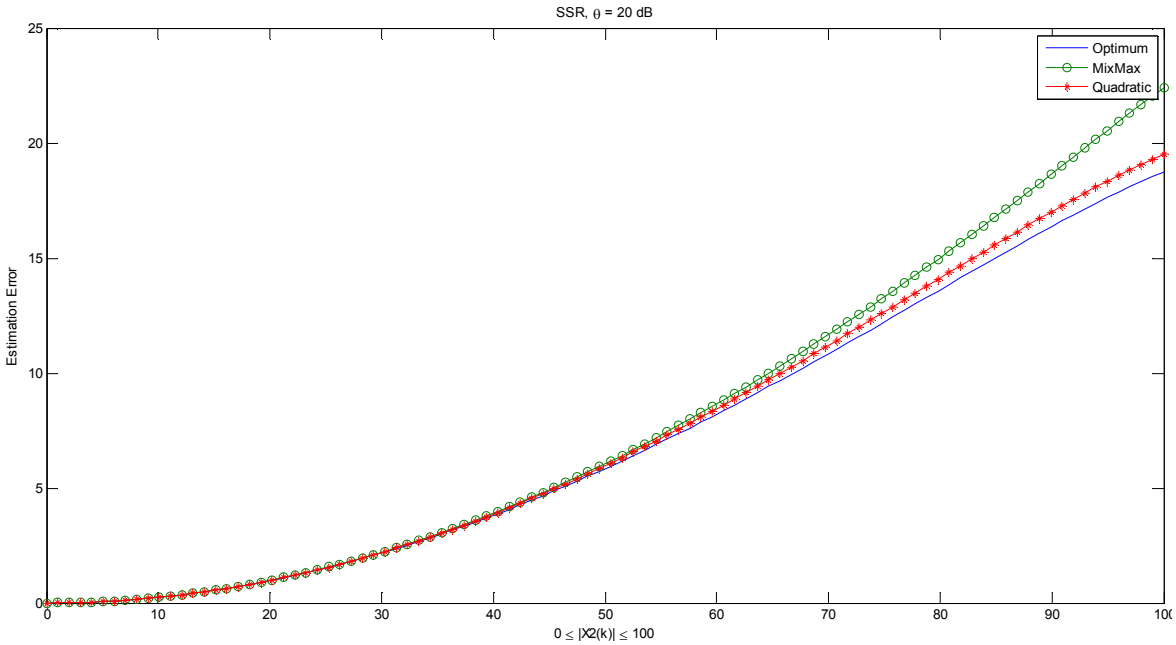

Fig. 5.5. *Estimation Error for different estimators at SSR = 20dB.*

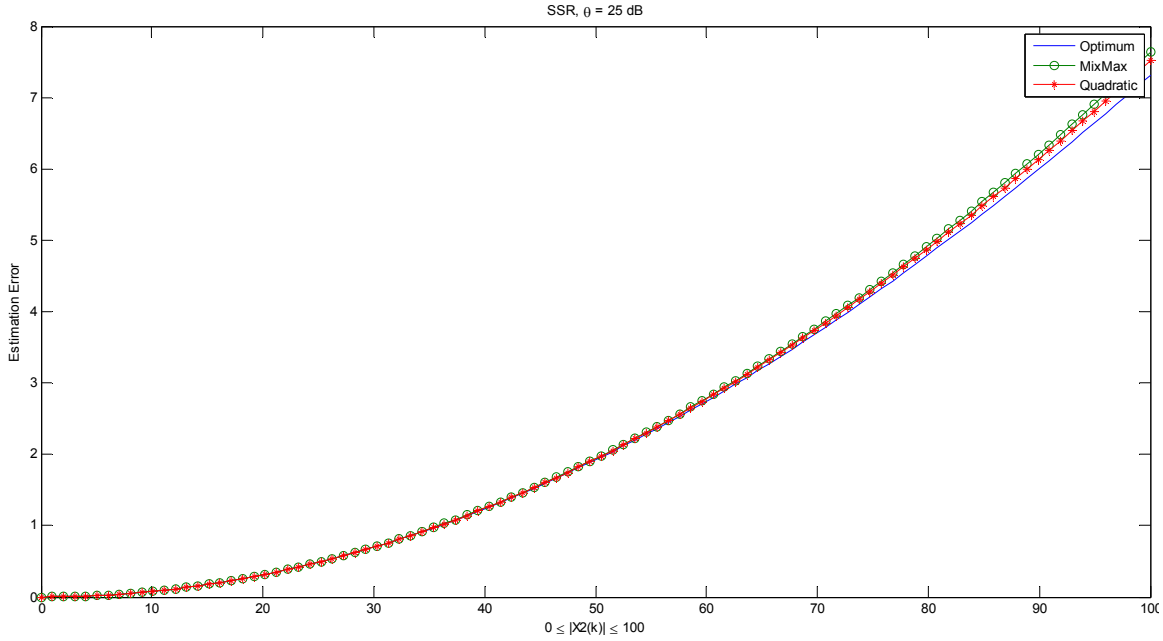

Fig. 5.6. *Estimation Error for different estimators at SSR = 25dB.*

The mixture estimates, for the estimators including Quadratic, MixMax and the derived estimator, at SSRs in the range of 0dB to 25dB, are evaluated in Fig. 5.7 - Fig. 5.12, based on the results in the Eqs. (5.10), (5.11) and (5.21) respectively. Assuming the DFT magnitude values of the speech signals as, $|x_1(k)| = 10$ and $|x_2(k)|$ ranging between [0,100].

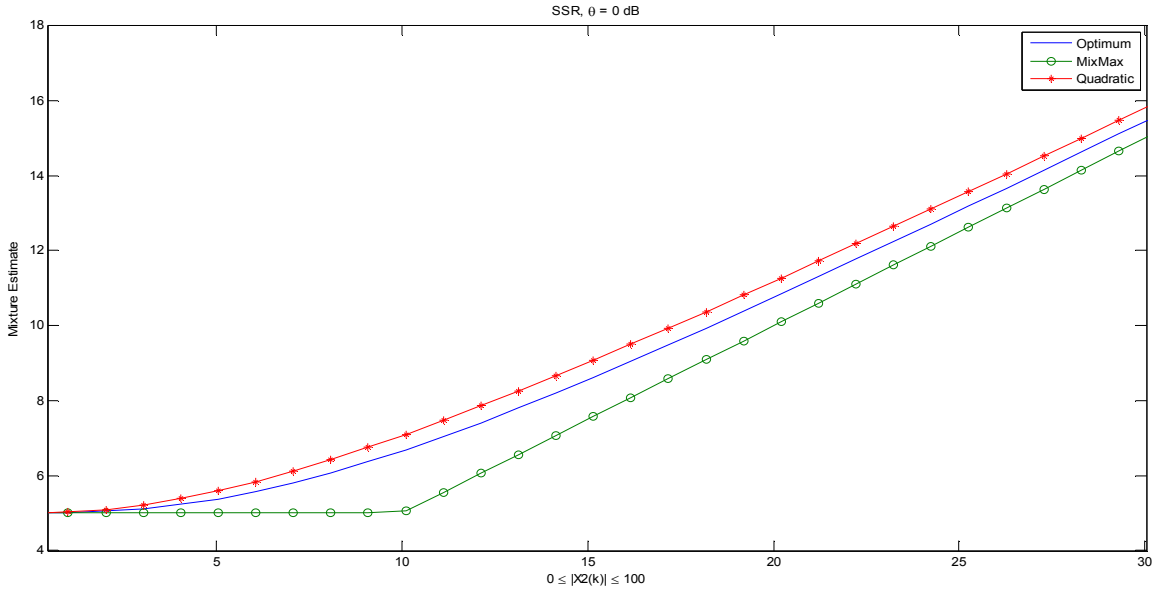

Fig. 5.7. *Comparison of mixture estimate of different estimators at SSR = 0dB.*

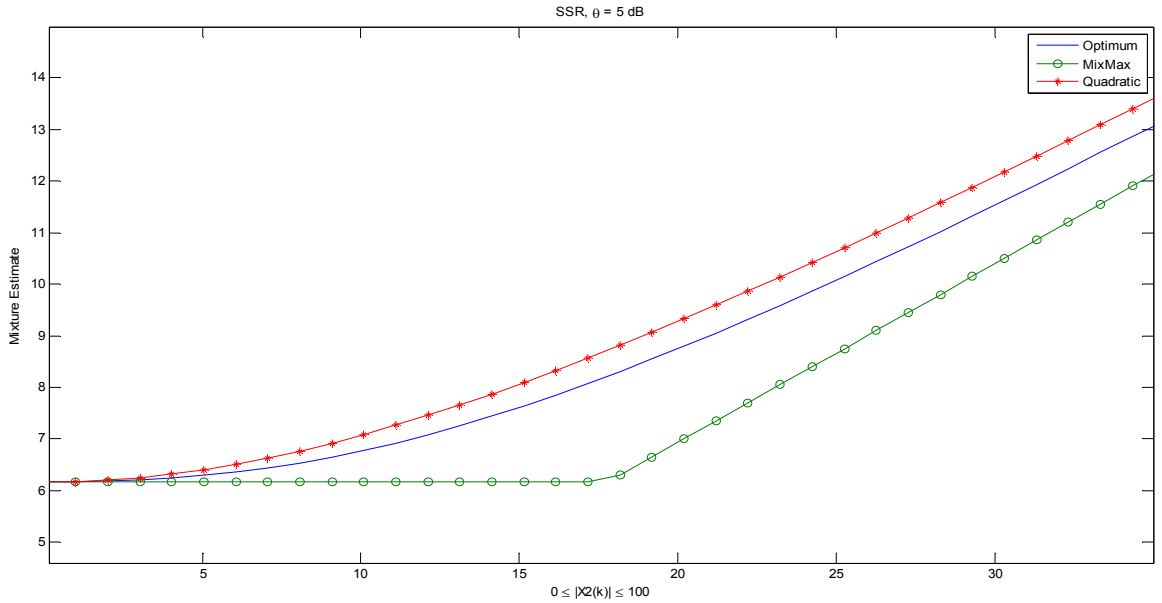

Fig. 5.8. *Comparison of mixture estimate of different estimators at SSR = 5dB.*

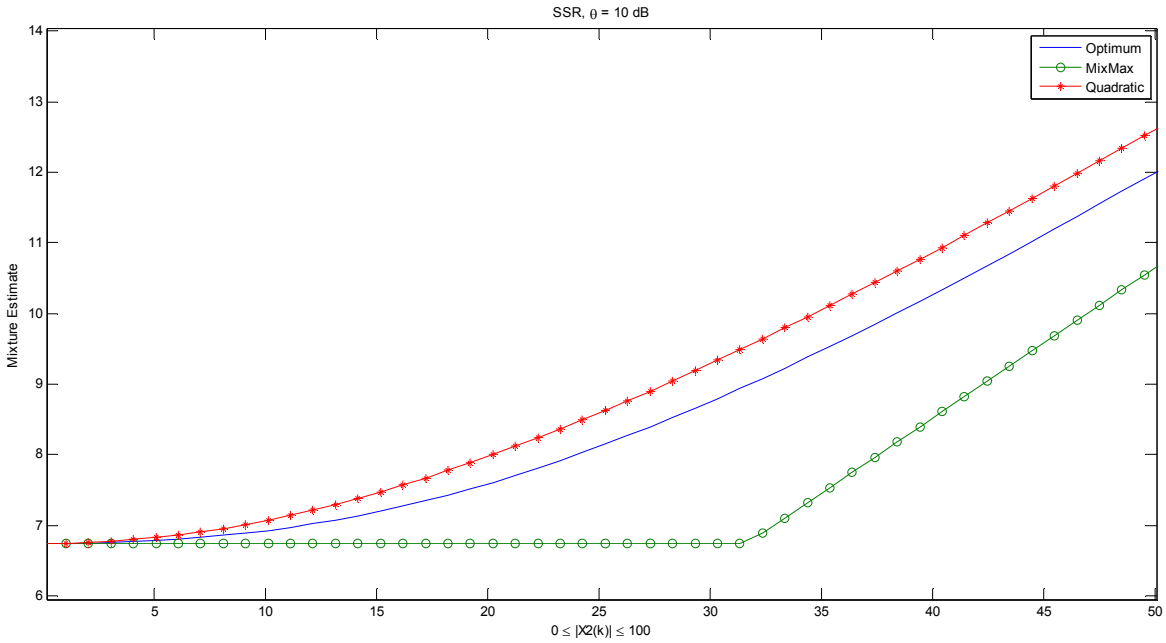

Fig. 5.9. *Comparison of mixture estimate of different estimators at SSR = 10dB.*

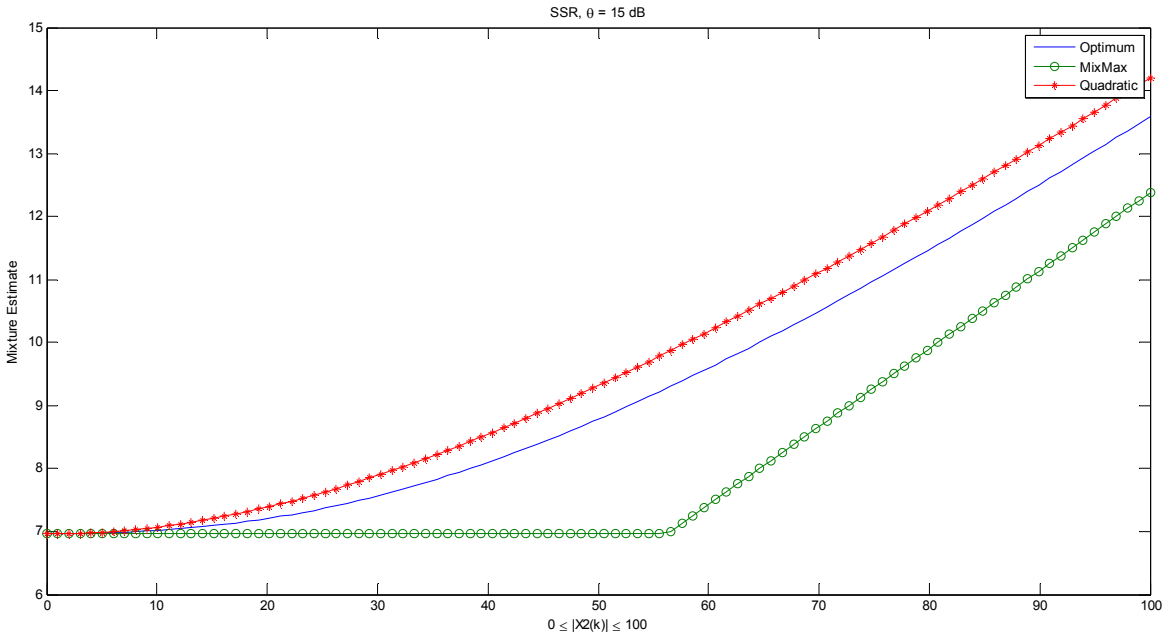

Fig. 5.10. *Comparison of mixture estimate of different estimators at SSR = 15dB.*

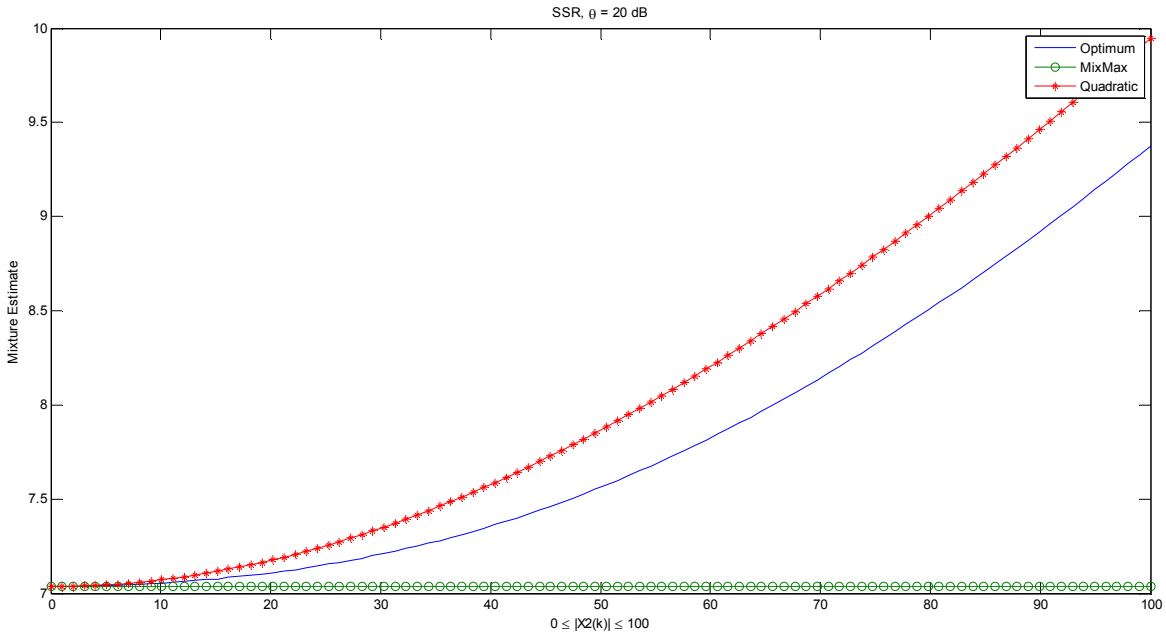

Fig. 5.11. *Comparison of mixture estimate of different estimators at SSR = 20dB.*

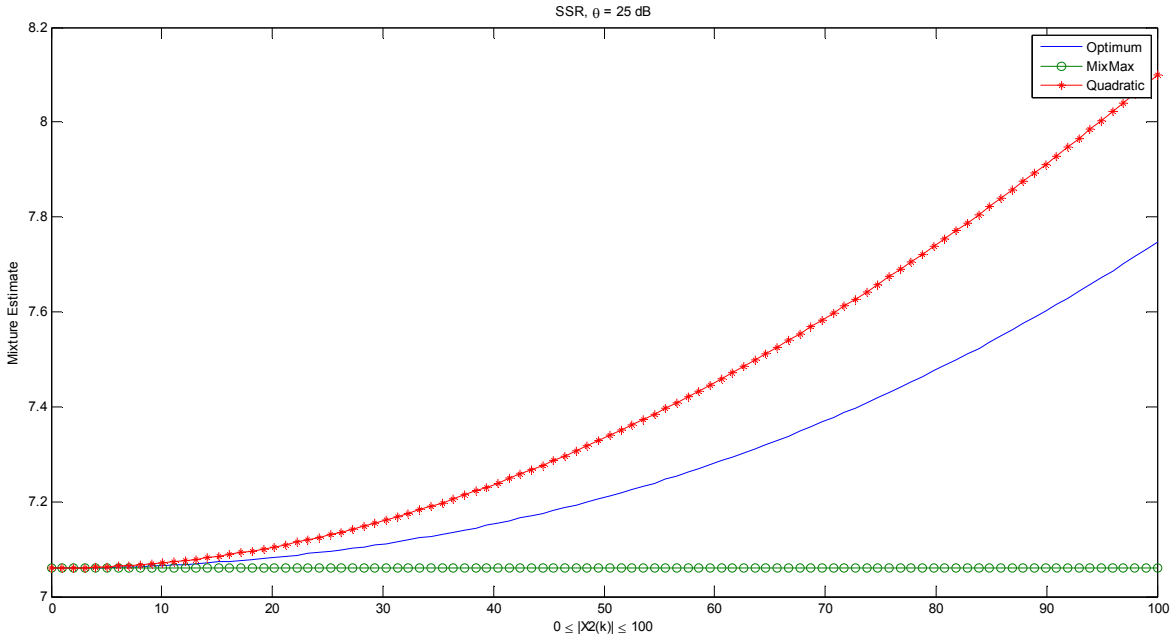

Fig. 5.12. *Comparison of mixture estimate of different estimators at SSR = 25dB.*

The estimation error for gain adapted optimum mixture estimator follows a specific pattern for each and every SSR, depending upon the magnitude values of the underlying speech signals and the SSR at which they are mixed. The estimation error is somewhat non-linear upto the point where $g_1|x_1(k)| \approx g_2|x_2(k)|$, and is somewhat linear after this point. The pattern for the estimation error is discussed in Fig. 5.13, for different SSRs in the range -6dB to +6dB.

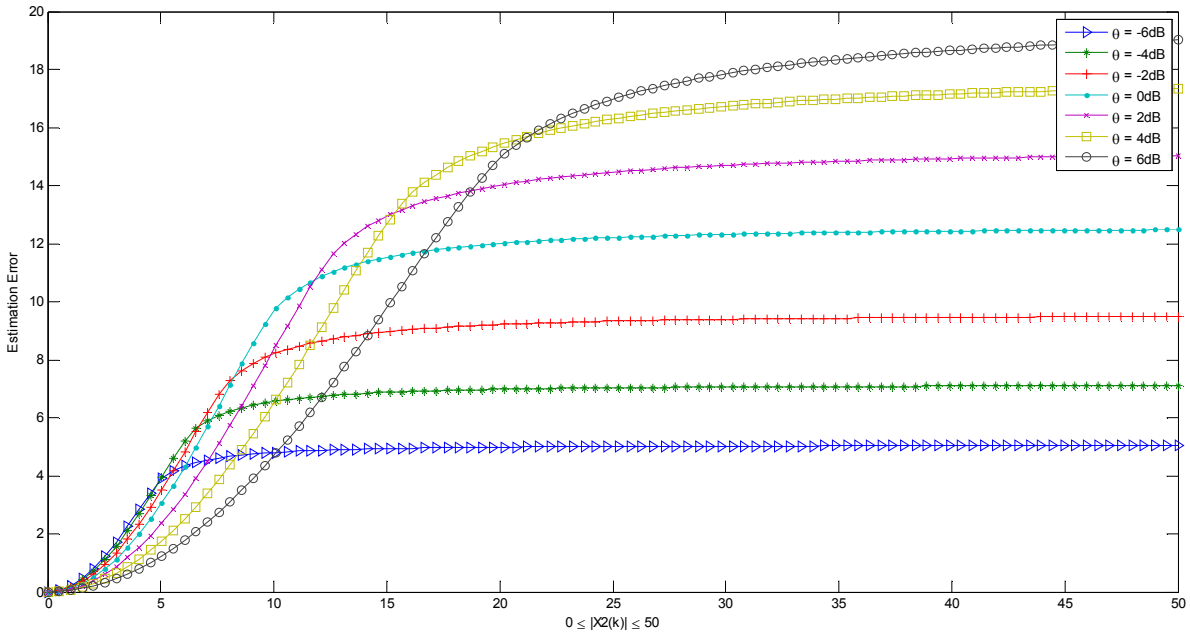

Fig. 5.13. *Estimation error for optimum mixture estimator for different SSRs.*

In order to efficiently segregate the two signals mixed with gain adaption, the issue that effects the performance of the estimator is the correct estimation of SSR $'\theta'$. The estimated SSR can be evaluated as

$$\theta_{est} = \theta + \hat{\theta}$$

where, θ_{est} is the estimated SSR, θ is the actual SSR, and $\hat{\theta}$ is the error in estimating SSR. It can also be seen as

$$E[\theta_{est}] = \theta + E[\hat{\theta}]$$

$$E[\theta_{est}] = \theta$$

where, $E[.]$ is the expectation operator.

Let actual SSR, $\theta = 0$dB, then the estimated SSR will be

$$\theta_{est} = \hat{\theta}$$

Fig. 5.14 below depicts the effect of incorrect SSR estimation, on the mixture estimate of the optimum mixture estimator. Considering the actual SSR, $\theta = 0dB$, and the estimated SSRs are in the vicinity of the actual SSR, within the range $-0.1dB$ to $+0.1dB$. It is demonstrated that the error in estimating SSR, results in deviation from the actual observation.

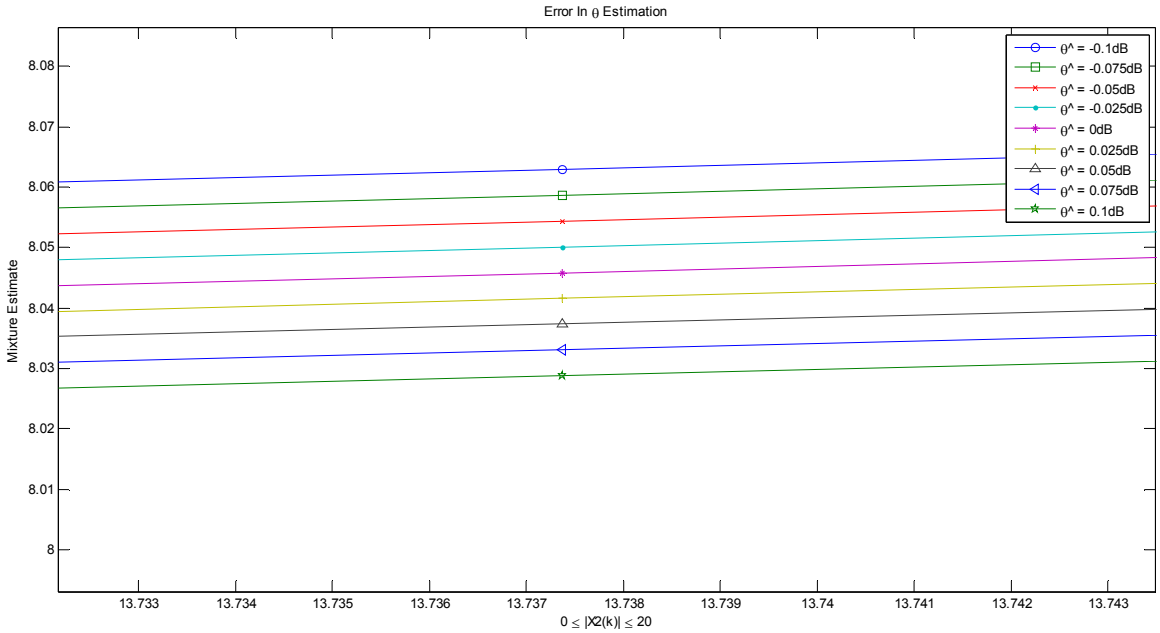

Fig. 5.14. *Error in SSR estimation.*

The estimation of the mixture of the two speech signals requires the phase information of the two speech signals, which is incorporated in the Elliptic Integral term in the derived optimum mixture estimator, given as $E(\beta(\theta))$, where θ is the SSR. The elliptic integral of second kind is given as

$$E(x) = \int_0^{\pi} \sqrt{1 - x^2 sin^2(\varphi)}\, d\varphi = \pi \left\{ 1 - \sum_{n=1}^{\infty} \left(\prod_{k=1}^{n} \left[\frac{2k-1}{2k} \right]^2 \right) \frac{x^{2n}}{2n-1} \right\}$$

The number of terms that approximate the integral is denoted as n, and it is depicted that more the terms that approximate the integral, results in lesser error in the terms of MSE. The effect of number of terms for approximation, on the estimation error for the optimum mixture estimate at SSR of 0dB, is demonstrated in Fig. 5.15 below. It can be concluded that lower estimation error results in the case of more terms to approximate the Elliptic integral series.

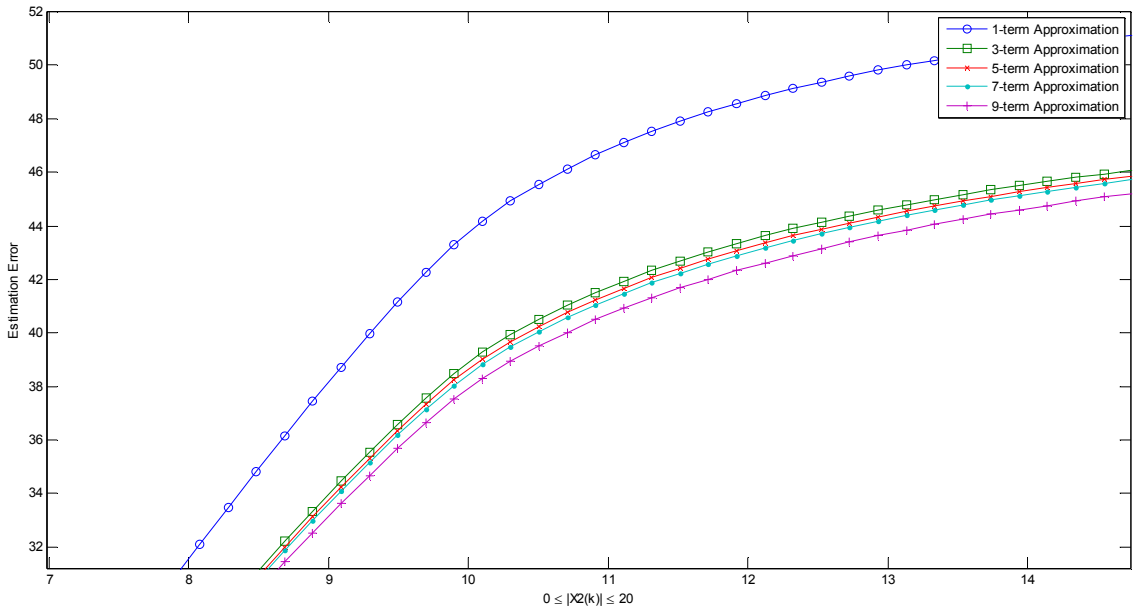

Fig. 5.15. *Estimation error for optimum mixture estimator using different term approximation of Elliptic Integral series.*

5.3.2 Experiments for Validation of Theoretical Results

In order to evaluate the performance of the derived optimum mixture estimator with other estimators, MixMax and Quadratic, we conducted the experiments which demonstrated that the derived estimator outperforms MixMax and Quadratic estimators. Cooke *et al.* [50] have provided a new database for the performance evaluation of speech separation and enhancement systems. The database consists of speech files of 34 speakers, each containing 500 utterances. The sampling rate of the speech signal is decreased to 8 kHz from the original 25 kHz in the database. We chose two speakers at random, selected the sentence

utterances of the respective speakers, and mixed the speech signals at different SSRs in the range -20dB to +20dB. The speech signals selected and worked upon are shown in Fig. 5.16 below.

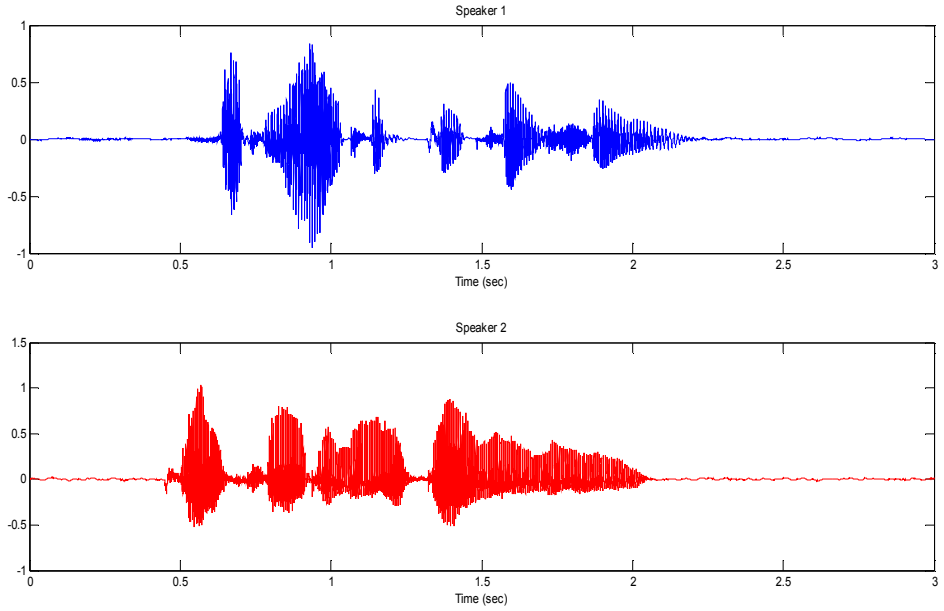

Fig. 5.16. *The speech samples of the selected speakers, used in the experiments.*

The segmented vectors are then transformed into frequency domain by taking Discrete Fourier Transform. The estimation error for each segment for a given estimator will be as,

$$e(k) = \frac{1}{N_{DFT}} \sum_{k=1}^{N_{DFT}} (|y(k)| - |y_{est}(k)|)^2$$

where $|y(k)|$ denotes the segmented original mix signal, $|y_{est}(k)|$ is the mixture estimate obtained through different estimators as given in (5.10), (5.11) and (5.21), and k is the frequency bin of the N_{DFT} =512 point DFT. The segmentation of the speech signals is done, using different including Hamming, Hanning, Bartlett (Triangular), Blackman, Kaiser windows (with rolling factor, β = 4.54, 6.76, and 8.96) [50], shown in Fig. 5.17, each of 64 ms duration corresponding to 512 samples of DFT. Therefore, the total number of segments used for simulation is 294 (here, separation between two consecutive segments is assumed to be 10ms, corresponding to 80 samples).

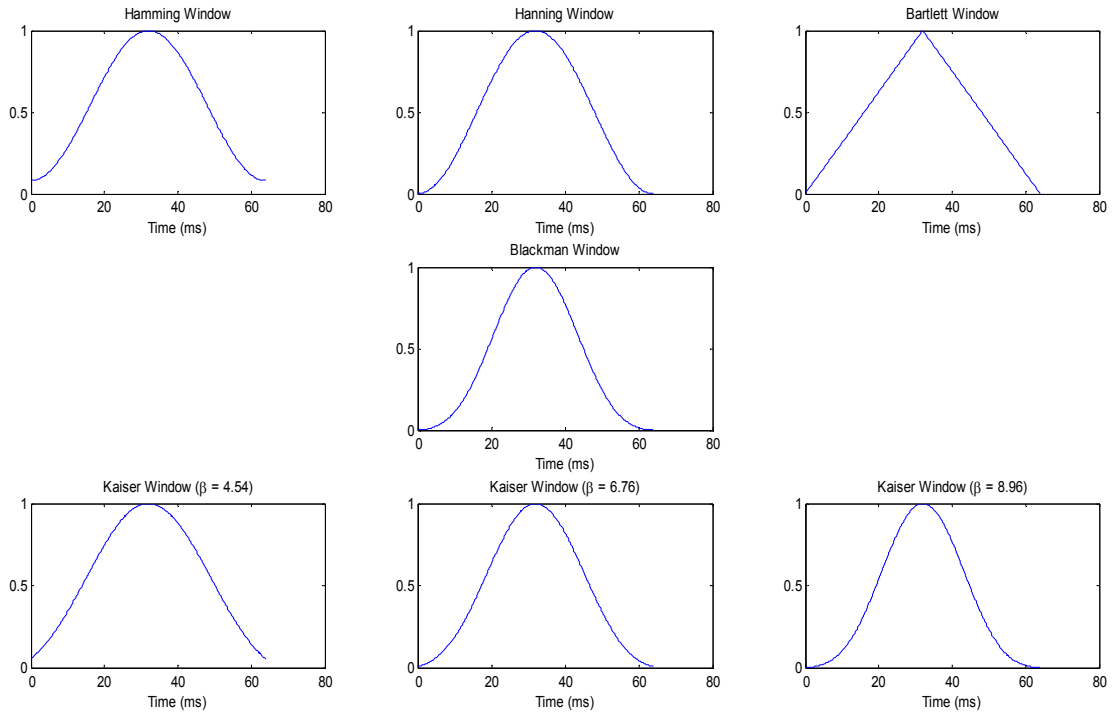

Fig. 5.17. *Different windows used for segmentation of the speech signals.*

The segmented speech samples, using the above windows are worked upon with the derived optimum mixture estimator. The speech signal consists of a set of harmonics which show themselves as peaks in the spectrum of the mixed signal. As the window size is increased, those harmonics of the two speakers that are close to each other are better resolved in the spectrum of the mixed signal. But an increase of the window duration to longer than 64 ms, causes the stationary assumption of the speech signal to no longer hold, since the duration of a voiced segment is around 60 ms [51]. Fig. 5.18 discusses the effect of different window sizes for optimum mixture estimator using different windows for segmentation of the audio speech signals. It is observed that as the window size increases, the estimation error increases, as the number of points to be estimated increases. But in order to efficiently resolve the two speech signals mixed, the window size cannot be reduced to much smaller size. Hence there is a trade-off for selection of window size for segmentation purposes of the mixed signals.

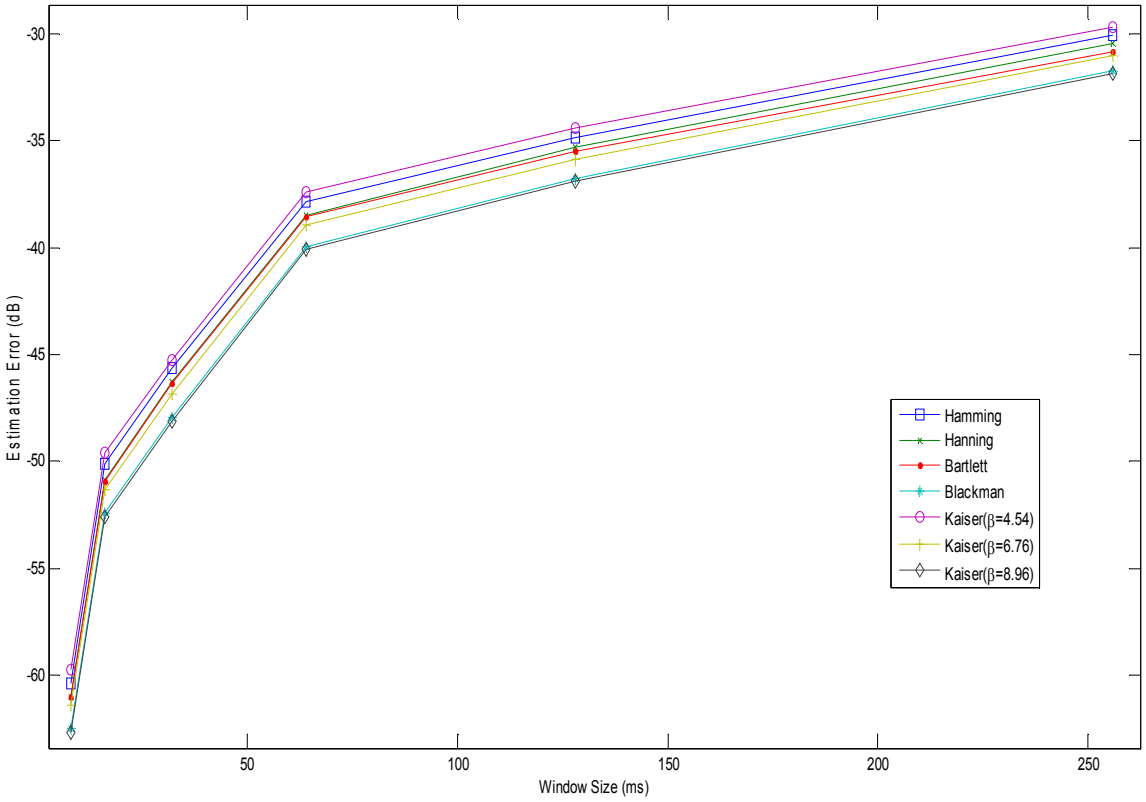

Fig. 5.18. *Comparison of different windows of different sizes.*

The optimum mixture estimator, with the other estimators, MixMax and Quadratic estimator, are worked upon the real speech signals extracted from the database. Each signal is segmented, transformed onto frequency domain using Discrete Fourier Transform, mixed at different SSRs, and then the estimates are framed using the results derived in the section above. The comparison of the three estimators on the basis of the estimation error in MSE sense is evaluated in Fig. 5.19, for different SSRs in the range of -20dB to +20dB. It can be concluded that the derived optimum mixture estimator outperforms the other two estimators, for different SSRs. The experimental verification holds good for the theoretical results derived. Hence, the derived optimum mixture estimator is the best mixture estimator for different values of the Signal-to-Signal Ratios.

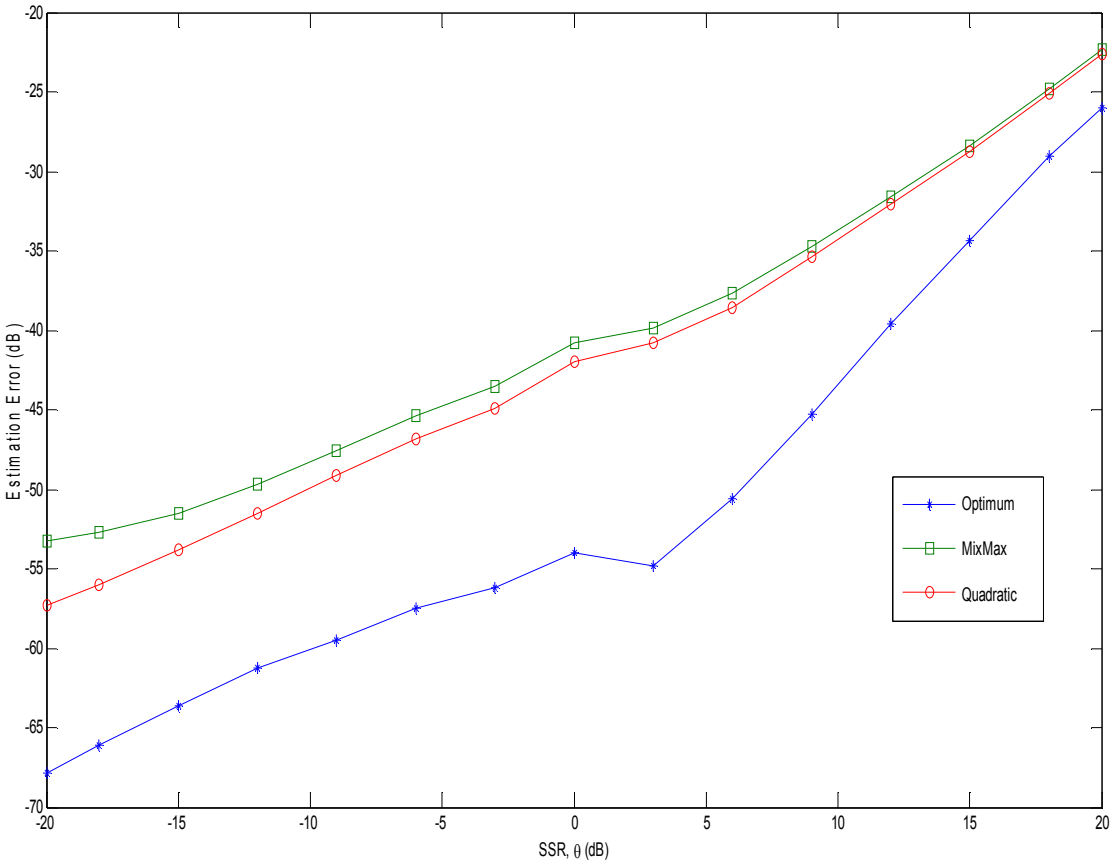

Fig. 5.19. *Comparison of different estimators for different SSRs.*

CONCLUDING REMARKS & FUTURE SCOPE

A challenging topic in separation systems is the separation of two speech signals from a single recording of their mixtures. Some common separation techniques such as BSS can hardly be adapted to separate the sources. Recently, model-based SCSS techniques, inspired from the works proposed in speech enhancement have been proposed to combat this crux problem. In these techniques, *a priori* knowledge of underlying sources is used to estimate the sources. In this thesis the probabilistic approach for single channel speech separation is presented.

The work in the thesis is an extension of the optimum mixture estimator for the Single Channel Speech Separation. It incorporates speakers' gains into the estimation process, which has not been touched upon significantly. Considering gains in the model, increases the number of unknown parameters in the separation process which in turn, complicates the separation problem. Expressing the gains as a function of the Signal-to-Signal Ratio and then deriving the MMSE based gain adapted optimum mixture estimator, which estimates the mixture of two speech signals, is presented in this thesis. In addition, the new estimator excludes the mapping process to a highly non-linear logarithm space, as proposed in MixMax, and can efficiently estimate a given mixture in an optimal manner, along with the well-known Elliptic series, for different signal to signal ratios. The experimental results proved that the proposed estimator outperforms the other estimators, MixMax and Quadratic estimator, as depicted in the theoretical results. Thus, the proposed estimator is more promising for practical applications. Moreover, this technique can be applied to the scenario with equal gains but signals having different levels of energy. We have demonstrated a satisfactory solution to the model based two speaker separation problems, and though this method in principle can be applied to more than two speakers, effective separation of more than two speakers from one mixture remains to be explored and demonstrated.

A probabilistic approach for the single channel speech separation is presented in this thesis. As our future work, we aim to present a model-based separation framework which employs the gain adapted optimum mixture estimator proposed in this thesis, as the central part for the

separation of the underlying speech signals. Our experience in this work indicates that the separation using the estimator derived in this thesis will outperform other model-based separation systems which use MixMax estimator as the core part of the separation process [52].

REFERENCES

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

[2] C. Jutten and J. Herault, "Space or time adaptive signal processing by neural network models," in *Proc. AIP Conf.*, 1986, pp. 206–11.

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

[4] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, pp. 1–10, 1991.

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

[6] S. I. Amari and J. F. Cardoso, "Blind source separation-semiparametric statistical approach," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2692–2700, Nov. 1997.

[7] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[8] M. S. Lewicki and T. J. Sejnowski, "Learning nonlinear overcomplete representations for efficient coding," in *Advances in Neural Information Processing Systems*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. Cambridge, MA: MIT Press, 1998.

[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Interspeech'06, Int. Conf. Spoken Lang. Process. (ICSLP'06)*, Pittsburgh, PA, Sep. 2006, pp. 2614–2617.

[10] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Proc. Int. Comput. Music Conf.*, 2003, pp. 231–234.

[11] G. J. Jang and T. W. Lee, "A probabilistic approach to single channel source separation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 1173–1180.

[12] K. Diamantaras and S. Kung, *Principal Component Neural Networks: Theory and Applications*, Wiley, 1996.

[13] D. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, 1992, pp. 771–4.

[14] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information processing,* vol. 8, 1996, pp. 757–63.

[15] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 6, pp. 1483–92, 1997.

[16] E. Kandel, J. Schwartz, and T. Jessell, *Principles of neural science*, Elsevier, New York, 1991.

[17] J. Pickles, *An introduction to the physiology of hearing*, Academic Press, London, 1988.

[18] P. Divenyi, Ed., *"Speech Separation by Humans and Machines"*, 1st ed., New York: Springer, 2004.

[19] D. Wang and G. Brown, *Computational auditory scene analysis: Principles, algorithms and applications*, Wiley-IEEE Press, New York, 2006.

[20] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., 2001.

[21] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed. Addison Wesley Longman, 1994.

[22] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 6, pp. 1483–92, 1997.

[23] S. Gazor and W. Zhang, "Speech probability distribution," *Signal Processing Letters*, vol. 10, no. 7, pp. 204–7, 2003.

[24] A. Nadas, D. Nahamoo and M.A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 10, pp. 1495–1503, 1989.

[25] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau & A. Sayadiyan, "A non-linear minimum mean square error estimator for the mixture-maximization approximation," *Electronic Letters,* vol. 42, no. 12, pp. 75–76, June 2006.

[26] A. Papoulis, *Probability, random variables, and stochastic processes*, 3rd ed., New York: McGraw-Hill, 1991.

[27] M. R. Spiegel, *Schaum's mathematical handbook of formulas and tables,* 2nd ed., New York: McGraw-Hill, June 1998.

[28] Y. Li and D. Wang, "On the Optimality of Ideal Binary Time-Frequency Masks," in *Proc. of ICASSP*, pp. 3501-3504, 2008.

[29] P. Mowalee, A. Sayadiyan & M. Sheikham, “Optimum Mixture Estimator for Single-Channel Speech Separation,” *International Symposium on Telecommunications,* 2008.

[30] S. T. Roweis, “One microphone source separation,” in *Proc. Neural Inf. Process. Syst.,* pp. 793–799, 2000.

[31] S. T. Rowies, “Factorial models and refiltering for speech separation and denoising,” *EUROSPEECH–03,* Vol. 7, pp. 1009–1012, May 2003.

[32] M. H. Radfar & R. M. Dansereau, “Single channel speech separation using soft mask filtering,” *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, no. 8, 2299–2310, Nov 2007.

[33] M. H. Radfar and R. M. Dansereau, "Long-term gain estimation in model-based single channel speech separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA2007, New Paltz, NY)*, pp. 143-146, October 21-24, 2007.

[34] M. H. Radfar and R. M. Dansereau, "Single channel speech separation using minimum mean square error estimation of sources' log spectra," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing* (MLSP'2007 Thessaloniki, Greece), pp. 128-132, Aug. 27-29, 2007.

[35] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Performance Evaluation of Three Features for Model-Based Single Channel Speech Separation Problem," in *Proc. of International Conference on Spoken Language Processing (Interspeech-- ICSLP)*, Pittsburgh, Pennsylvania, USA, pp. 2610-2613, Sept. 17-21 , 2006.

[36] M. H. Radfar and R. M. Dansereau, "Single Channel Speech Separation Using Maximum a Posteriori Estimation," in *Proc. of Interspeech'2007, Intern. Conf. on Spoken Language Processing (ICSLP'2007 Antwerp, Belgium)*, pp. 958-961, Aug. 27-31, 2007.

[37] M. N. Schmidt & R. K. Olsson, “Linear regression on sparse features for single-channel speech separation,” in *Proc. of IEEE workshop on applications of signal processing to audio and acoustics (WASPAA2007),* pp. 26–29, New Paltz, New York, October 2007.

[38] A. M. Reddy & B. Raj, “Soft mask methods for single channel speaker separation,” *IEEE Transactions on Audio, Speech and Language Processing,* vol. 1*5,* no. 6, pp. 1766–1776, Aug. 2007.

[39] R. Weiss & D. Ellis, “Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking,” in *Proc. of workshop on Statistical and Perceptual Audition SAPA-06,* pp. 31–36, October 2006.

[40] T. Beierholm, B. D. Pedersen & O. Winther, "Low complexity Bayesian single channel source separation," in *Proc. of ICASSP–04,* Vol. 5, pp. 529–532, May 2004.

[41] T. Kristjansson, T. H. Attias & J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. of ICASSP–04,* pp. 817–820, May 2004.

[42] M. H. Radfar, R. M. Dansereau & A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. I, 2007, doi:10.1155/2007/84186.

[43] M. H. Radfar, R. M. Dansereau & A. Sayadiyan, "Monaural speech segregation based on fusion of source driven with model-driven techniques," *Speech Communication,* vol. 49, no. 6, pp. 464–476, June 2007.

[44] M. J. Reyes-Gomez, D. Ellis & N, Jojic, "Multiband audio modeling for single channel acoustic source separation," in *Proc. ICASSP–04,* Vol. 5, pp. 641–644, May 2004.

[45] A. M. Reddy & B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *INTERSPEECH–2004,* pp. 2445–2448, October 2004.

[46] M. P. Cooke, J. Barker, S. P. Cunningham & X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Proc. of JASA*, Nov. 2005.

[47] Y. Ephraim, "Gain-adapted hidden markov models for recognition of clean andnoisy speech," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 40, no. 6, pp. 1303–1316, June 1992.

[48] D. Y. Zhao & W. B. Kleijn, "HMM-based gain modelling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, no. 3, pp. 882–892, March 2007.

[49] L. Benaroya, F. Bimbot & R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Speech Audio Processing*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[50] J. G. Proakis, D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 3rd ed., New York: Prentice-Hall, 1996.

[51] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "On the choice of window size in model-based single channel speech separation," *Proc. of the IEEE Canadian Conf. on Elec. and Comp. Eng. (CCECE'2006 Ottawa)*, pp. 981-984, May 7-10, 2006.

[52] M. H. Radfar, R. M. Dansereau, and W.Y. Chan, "Monaural speech separation based on gain adapted minimum mean square error estimation," *Journal of Signal Processing Systems*, vol. 10, no. 1 ,pp. 61:21-37, October 2008.