Blind Convolutive Speech Separation and Dereverberation

Tariqullah Jan

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

February 2012

Abstract

Extraction of a target speech signal from the convolutive mixture of multiple sources observed in a cocktail party environment is a challenging task, especially when the room acoustic effects and background noise are present in the environment. Such acoustic distortions may further degrade the separation performance of many existing source separation algorithms. Algorithmic solutions to this problem are likely to have strong impact on many applications including automatic speech recognition, hearing aids and cochlear implants, and human-machine interaction. In such applications, to extract the target speech, it is usually required to deal with not only the interfering sound, but also the room reverberations and background noise.

To address this problem, several methods are developed in this thesis. For the blind separation of a target speech signal from the convolutive mixture, a multistage algorithm is proposed in which a convolutive independent component analysis (ICA) algorithm is applied to the mixture, followed by the estimation of an ideal binary mask (IBM) from the separated sources obtained with the convolutive ICA algorithm. In the last step, the errors introduced due to estimation of the IBM are reduced by cepstral smoothing.

The separation performance of the above algorithm, however, deteriorates with the increase in surface reflections and background noise within the room environment. Two different methods are therefore developed to reduce such effects. In the first method which is also a multistage method, acoustic effects and background noise are treated together using an empirical-mode-decomposition (EMD) based algorithm. The noisy reverberant speech is decomposed adaptively into oscillatory components called intrinsic mode functions (IMFs) via an EMD algorithm. Denoising is then applied to selected high frequency IMFs using an EMD-based minimum mean squared error (MMSE) filter, followed by spectral subtraction of the resulting denoised high and low-frequency IMFs. The second method is a two-stage dereverberation algorithm in which the smoothed spectral subtraction mask based on a frequency dependent model is derived and then applied to the reverberant speech to reduce the effects of late reverberations. Wiener filtering is then applied such that the early reverberations are attenuated.

Finally, an algorithm is developed for joint blind separation and blind dereverberation. The proposed method consists of a step for the blind estimation of reverberation time (RT). The method is employed in three different ways. Firstly, the available mixture signals are used to estimate blindly the RT, followed by the dereverberation of the mixture signals. Then, the separation algorithm is applied to these resultant mixtures. Secondly, the separation algorithm is applied first to the mixtures, followed by the blind dereverberation of the segregated speech signals. In the third scheme, the separation algorithm is split such that the convolutive ICA is first applied to the mixtures, followed by the blind dereverberation of the signals obtained from convolutive ICA. Then, the T-F representation of the dereverberated signals is used to estimate the IBM followed by cepstral smoothing.

Key words: Independent component analysis (ICA), convolutive mixtures, ideal binary mask (IBM), estimated binary mask, cepstral smoothing, musical noise, empirical mode decomposition (EMD), spectral subtraction, speech dereverberation, speech enhancement, reverberation time (RT), blind estimation of RT

Email: t.jan@surrey.ac.uk

WWW: http://www.eps.surrey.ac.uk/

Acknowledgements

I would like to express my deepest appreciation to my supervisor Dr. Wenwu Wang who has given me an excellent opportunity to work with him; provided continual support, advice and guidance throughout my research. I would also like to thank my co-supervisor Professor Josef Kittler for his kind behaviour and support throughout my PhD studies. In addition, I would like to thank Professor DeLiang Wang for his kind support at all times and advice during this period. I would also like to extend my gratitude to the University of Engineering and Technology Peshawar, Pakistan for providing me scholarship during my candidature. I am grateful to my friends and colleagues in the centre for their invaluable help. Special thanks to our ex-centre administrator James Field who was always very helpful and cooperative.

To my wonderful family, words cannot describe how much love and their support means to me. I thank my lovely mother and my kind father who were encouraging and supporting me emotionally throughout my studies. To all my sisters and brothers, who were supporting me emotionally. Also special thanks to my very close friends Sana, Affan, Aftab, Ibrahim, Hamza, Majid, and Humayun for their support which enabled me to focus on my thesis.

Last but never least, all this would never have been possible without my faith in ALLAH, who I believe gave me the strength, power and knowledge I needed. Thank you.

List of Figures

List of Tables

Contents

Acronyms and Mathematical Symbols

List of Acronyms

Acronym/Abbreviation	Meaning
AIR	Acoustic Impulse Response
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
BD	Blind Dereverberation
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CPP	Cocktail Party Problem
DFT	Discrete Fourier Transform
EMD	Empirical Mode Decomposition
IBM	Ideal Binary Mask
ICA	Independent Component Analysis
IMF	Intrinsic Mode Function
ISTFT	Inverse Short Time Fourier Transform
LP	Linear Prediction
ML	Maximum Likelihood
MMSE	Minimum Mean Squared Error
MOS	Mean Opinion Score
NMF	Non-negative Matrix Factorization
NTF	Non-negative Tensor Factorization
PDF	Probability Density Function
PEL	Percentage of Energy Loss
PNR	Percentage of Noise Residue
PSD	Power Spectral Density
RIR	Room Impulse Response
RT	Reverberation Time
SDR	Signal to Distortion Ratio
SegSRR	Segmental Signal to Reverberation Ratio
SNR	Signal to Noise Ratio
SRR	Signal to Reverberation Ratio
SS	Spectral Subtraction
STFT	Short Time Fourier Transform

List of Symbols

M_1	Ideal binary mask in time-frequency domain
S_1	Target speech signal in time-frequency domain
S_2	Interference speech signal in time-frequency domain
x_j	The jth mixture signal in the time domain
s_i	The ith source signal in the time domain
h_{ji}	Room impulse response from source s_i to microphone x_j
$\mathbf{X}$	$= [X_1, ..., X_M]^T$, time-frequency representations of the microphone signals, and $[\cdot]^T$ denotes vector transpose
n	Discrete time index
$\mathbf{S}$	$= [S_1, ..., S_N]^T$, time-frequency representations of the source signals
$\mathbf{H}$	Mixing matrix
$\mathbf{W}$	Denoted as $[[W_{11}, W_{12}]^T, [W_{21}, W_{22}]^T]^T$ is an unmixing filter
$\mathbf{Y}$	$= [Y_1, Y_2]^T$, estimated source signals in the time-frequency domain after convolutive ICA
k	Frequency bin index
m	Time frame index
$\mathbf{y}$	$= [y_1, y_2]^T$, estimated source signals in the time domain after convolutive ICA
$\tilde{y}_i$	The ith estimated normalized source signal in the time domain after convolutive ICA
$\tilde{Y}_i$	The ith estimated normalized source signal in the time-frequency domain after convolutive ICA

M_1^f	Estimated binary mask for $\tilde{Y}_1$
M_2^f	Estimated binary mask for $\tilde{Y}_2$
$y_i^{'}$	The ith estimated source signal in the time domain after binary masking
M_i^c	The ith estimated binary mask in the cepstrum domain
$\overline{M}_i^s$	The ith estimated smoothed binary mask in the cepstrum domain
l	Quefrency bin index for the cepstrum domain
l_{env}	Quefrency bin index that represents the spectral envelope of the mask $\mathbf{M}^f$ defined as $[M_1^f, M_2^f]^T$
l_{pitch}	Quefrency bin index showing the structure of the pitch harmonics in $\mathbf{M}^f$ defined as $[M_1^f, M_2^f]^T$
λ	Parameter for controlling the smoothing level
Y^c	Cepstrum domain representation of the segregated speech signal $y^{'}$
$\overline{M}_i^f$	The ith estimated smoothed binary mask in the time-frequency domain
$\overline{Y}_i^f$	The ith estimated smoothed source signal in the time-frequency domain
$x(n)$	Noisy reverberant speech in time domain
$\tilde{z}_j(n)$	The jth IMF component in time domain
$u(n)$	Upper envelope in Time Domain
$l(n)$	Lower envelope in Time Domain
$r_C(n)$	Residue signal in time domain
$\mid B_j \mid$	The magnitude spectrum of the jth IMF component

$\mid \hat{B}_j \mid^2$	The estimated noise power of the jth IMF component
$\tilde{z}_j(k;m)$	Spectrum of the jth noisy IMF component, where k is discrete frequency index and m is the time frame index
$\hat{z}_j(k;m)$	Spectrum of the jth estimated IMF component
SNR_{prio}	Prior signal to noise ratio
SNR_{inst}	Instantaneous signal to noise ratio
α	Weighting factor
$\|S_{l_j}(k;m)\|^2$	The short term power spectrum of the late reverberations in the jth IMF component
γ	Scaling factor
$\omega(m)$	Smoothing window
ρ	The relative delay of the late reverberations
$\|\tilde{s}_j(k;m)\|^2$	The power spectrum of the jth IMF component of the estimated version of the original speech
$\tilde{s}_j(n)$	The inverse FFT of $\tilde{s}_j(k;m)$
$\hat{s}(n)$	The final enhanced signal in the time domain
$h(n)$	Room impulse response in the time domain
$\beta(n)$	Sequence of zero-mean mutually independent and identically distributed Gaussian random variables
α_1	Decay rate
f_s	Sampling frequency
$T_{60}(k)$	Frequency dependent reverberation time
$\sigma^2_{x_{late}}$	The spectral variance of the late reverberant speech

$\sigma_x^2(m,k)$	The variance of the reverberant speech
w	The analysis window of length N
R	The hop size
τ	Forgetting factor
φ	*posteriori* signal-to-distortion ratio
ρ_1	Power ratio
E_s	Moving average window
ψ	Scaling factor for determining the level of smoothing
F_s	Smoothing filter
G_{late}	Smoothed mask
coh	Coherence between signals
Υ	Power spectral density
L	Likelihood function
σ^2	Variance of Gaussian distribution
$ln(L)$	Log likelihood
B	Number of samples in a frame
l_{sub}	Sub frame index
$\tau_{l_{sub}}$	Weighting factor
ϱ	Variance of Laplace distribution
$\mathcal{L}(\theta,\varrho)$	Laplace distribution with mean θ and variance ϱ

List of Publications

PUBLICATIONS

Journal Articles

T. Jan, W. Wang, and D. L. Wang, "A multistage approach to blind separation of convolutive speech mixtures." *Speech Communication*, Volume 53, Pages 524–539, April 2011.

T. Jan and W. Wang, "Blind dereverberation using blind estimation of reverberation time based on statistical model." *IEEE Transaction on Audio, Speech, and Language Processing, to be submitted.*

T. Jan and W. Wang, "Empirical mode decomposition for joint denoising and dereverberation." *Digital Signal Processing (Elsevier), submitted.*

Conference Papers

T. Jan, W. Wang, and D. L. Wang, "A multistage approach for blind separation of convolutive speech mixtures." *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1713–1716, Taiwan, April 2009.

T. Jan and W. Wang, "Empirical mode decomposition for joint denoising and dereverberation." *19th European Signal Processing Conference (EUSIPCO 2011)*, Pages 206–210, Barcelona Spain, August 2011.

T. Jan, W. Wang, and D. L. Wang, "Binaural Speech Separation Based on Convolutive ICA and Ideal Binary Mask Coupled with Cepstral Smoothing." *Eighth IMA International Conference on Mathematics in Signal Processing*, Cirencester UK, December 2008.

T. Jan and W. Wang, "Frequency Dependent Statistical Model for the Suppression of Late Reverberations." In *IEEE International Workshop on Statistical Signal Processing, submitted.*

T. Jan and W. Wang, "Blind reverberation time estimation based on Laplacian distribution." In *IEEE International Workshop on Statistical Signal Processing, submitted.*

Technical Workshop

T. Jan, W. Wang, and D. L. Wang, "A multistage approach for blind separation of convolutive speech mixtures." *Poster Presentation* in *One-day meeting for young speech researchers, UK Speech*, 16 July 2008, University of Surrey, Guildford.

Book Chapters

T. Jan and W. Wang, *Machine Audition: Principles, Algorithms and Systems.*, chapter "Cocktail party problem: Source separation issues and computational methods.", IGI Global, USA, 2010.

Chapter 1

Introduction

1.1 Problem Description and Motivation

The extraction of a target speech signal from a mixture of multiple signals is classically referred to as the cocktail party problem (CPP), the concept of which was introduced for the first time by Cherry in 1953 [30]. It can be also formulated as: "How do we recognize what one person is saying when others are speaking at the same time", which has turned out to be a highly complex problem when background noise and acoustic disturbance are taken into consideration. Although it poses big challenges in many signal processing applications, human listeners with normal hearing are generally very skilful in separating the target speech within a complex auditory scene [172]. It has been observed that people with perceptive hearing loss suffer from insufficient speech intelligibility [40, 86]. It is difficult for them to pick up the target speech, in particular, when there exist some interfering sounds and background noise nearby. However, amplification of the signal is not sufficient to increase the intelligibility of the target speech as all the signals (both target and interference) are amplified. For this application scenario, it is highly desirable to produce a machine that can offer clean target speech to these hearing impaired people.

Despite being studied for decades, the CPP remains a scientific challenge that demands further research efforts [172]. Computational modelling and algorithmic solutions to this problem are likely to have strong impact on several applications including hearing

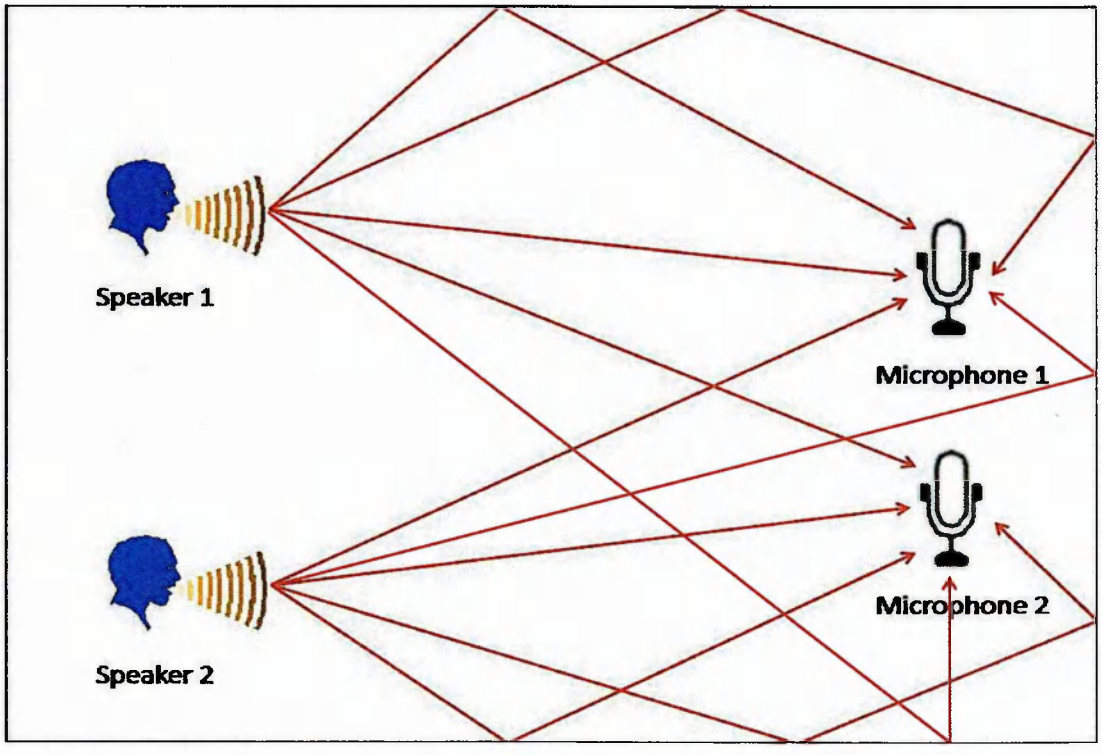


Figure 1.1: A simplified scenario of the cocktail party problem with two speakers and two listeners (microphones).

aids and cochlear implants, human-machine interaction and robust speech recognition in uncontrolled natural environments. Figure 1.1 illustrates the cocktail party effect using a simplified scenario with two simultaneous conversations in the room environment.

The key challenge is to recover the target speech from the mixture of speech signals recorded in a cocktail party environment such that the interference of the competing speech signals is suppressed. One promising technique to address this problem is under the framework of blind source separation (BSS) where the mixing process is generally described as a linear convolutive model, and independent component analysis (ICA) [73, 97] can then be applied to separate the convolutive mixtures either in the time domain [32, 45, 46], in the transform domain [2, 8, 64, 68, 102, 121, 136, 139, 178, 189], or their hybrid [90, 98], assuming the source signals are statistically independent [8, 44, 102, 107, 120, 121]. Although the convolutive BSS problem, i.e. separating unknown sources from their convolutive mixtures, has been studied extensively, the separation performance of many developed algorithms is still limited, and leaves much room for

further improvement. This is especially true when dealing with reverberated and noisy mixtures.

Another technique proposed to tackle this problem is under the framework of computational auditory scene analysis (CASA). It is the study of auditory scene analysis (ASA) by computational means. ASA is the process by which the human auditory system performs sound localization and recognition in order to pick up the target signal from the cocktail party environment. Recently in CASA, a technique called ideal binary mask (IBM), has shown promising properties in suppressing interference and improving intelligibility of target speech. IBM is obtained by comparing the T-F representations of the target speech and background interference, with one/unity assigned to a T-F unit where the target energy is stronger than the interference energy and zero otherwise [168]. The target speech can then be obtained by applying the IBM to the T-F representation of the mixture, together with an inverse transform. The IBM technique was originally proposed as a computational goal or performance benchmark of a CASA system [168, 172]. Recent studies reveal that by suppressing the interference signals from the mixtures, the IBM technique can significantly improve the intelligibility of the target speech [173]. This simple yet effective approach offers great potential for improving speech separation performance of ICA algorithms. Different from many ICA approaches with linear models [101], signals estimated in the T-F plane have mostly non-overlapping supports for different speaker signals and thus one can use IBM to extract the target speech from their mixture signal. The IBM is obtained by assuming both the target speech and interfering signal are known *a priori*. However, in practice, only mixtures are available, and the IBM must be estimated from the mixtures, which is a major computational challenge.

To overcome these limitations a computationally very efficient algorithm is developed in this thesis to estimate the IBM from intermediate separation results that are obtained by applying an ICA algorithm to the mixtures. The limitation of the aforementioned CASA methods, i.e., having to estimate the IBM directly from the mixtures, is mitigated as the IBM can now be estimated from the coarsely separated source signals obtained by ICA algorithms. The estimated IBM can be further used to enhance the separation quality of the coarsely separated source signals. To deal with the estimation

errors of the binary mask, a cepstrum based processing method was employed.

Another major challenge in addressing the CPP is the presence of acoustic effects in an enclosed cocktail party environment that can degrade the quality of the extracted target speech signal. As the listeners (or microphones) are not always located near the desired (target) speech signal and hence the signals received at the listeners (or microphones) are typically degraded by not only the interfering sound source nearby, but also the reverberations introduced by the multi-path propagation from the target source due to surface reflections within the room. Reverberation effects in speech can be described as sounding distant with noticeable colouration and echo. These detrimental perceptual effects generally increase with distance between the speaker and the listener (or microphone). Furthermore, with the spread in the time of arrival of reflections at the microphone, reverberation causes blurring of speech phonemes. These detrimental effects seriously degrade the intelligibility of the target speech and the performance of the speech separation algorithms. Therefore extraction of a target speech signal from a mixture is not sufficient to mitigate the CPP but there is a need to develop methods that can reduce the effects caused by the reverberations.

One more challenge is the ambient noise which is also the source of interference that degrades the quality of target speech while addressing the CPP. It is well known that background noise reduces the intelligibility of speech and that the greater the level of background noise the greater the reduction in intelligibility. Human listeners with normal hearing are able to understand speech in a moderately noisy environment because speech is a highly redundant signal and thus even if part of the speech signal is masked by noise, other parts of the speech signal will convey sufficient information to make the speech intelligible, or at least sufficiently intelligible to allow for effective speech communication. There is less redundancy in the speech signal for a person with hearing loss since part of the speech is either not audible or is severely distorted because of the hearing loss. Background noise that masks even a small portion of the remaining, impoverished speech signal will degrade intelligibility significantly because there is less redundancy available to compensate for the masking effects of the noise. As a consequence, people with hearing loss have much greater difficulty than people with normal hearing in understanding speech in noise. Therefore, it is necessary to develop methods

that can reduce the ambient noise in order to improve the intelligibility of the target speech extracted from the mixture in the cocktail party environment.

The separation performance of the algorithm developed in this thesis for the blind separation of target speech from convolutive mixtures has been restrained due to acoustic effects and ambient noise. Hence an algorithm is developed which can reduce the effects of reverberations and background noise resulting in improved speech intelligibility. The developed method is using empirical-mode-decomposition (EMD) based subband processing. Noisy reverberant speech is decomposed adaptively into oscillatory components called intrinsic mode functions (IMFs) via an EMD algorithm, followed by, denoising the selected IMFs using EMD-based minimum-mean squared error (MMSE) filter. Then spectral subtraction is applied to the resulting denoised high-frequency IMFs and low-frequency IMFs. Finally, the enhanced speech signal is reconstructed from the processed IMFs.

Another method is proposed to deal with the room reverberation separately. It is a two stage method, in the first stage a frequency dependent statistical model of the decay rate of the late reverberations (details about late reverberation are given in Chapter 2) is used to estimate the spectral variance of late reverberation, followed by estimation of the spectral mask containing the gain functions. Then, the smoothing filter is applied to the spectral mask to reduce the artifacts, and finally the smoothed gain function is applied to the reverberant signal to suppress the late reverberations. In the second stage, a dual-channel Wiener filter is used to deal with the early reverberations (details about early reverberation are given in Chapter 2).

Finally, a joint blind dereverberation and separation algorithm is proposed. The developed method has been employed in three different ways. Firstly, the available mixture signals are used to estimate blindly the reverberation time (RT) based on a maximum-likelihood (ML) method and statistical modelling of the sound decay rate of the reverberant speech, followed by the dereverberation of the mixture signals using the method based on the frequency dependent statistical model. Then, the separation algorithm is applied to these resultant mixtures so that the source (target) speech signals can be obtained. Secondly, the separation algorithm is applied primarily to the mixtures to

segregate the speech signals, followed by the blind estimation of RT from the separated speech signal. Then, dereverberation is employed to the segregated (target) speech signals. In the third scheme, the separation algorithm is split such that the convolutive ICA is first applied to the mixtures to obtain the estimated source signals. Then, the signal obtained from the convolutive ICA is used to estimate the RT followed by the blind dereverberation of the signals obtained from convolutive ICA. Then, the T-F representation of dereverberant signals are used to estimate the IBM followed by cepstral smoothing to enhance the target speech signal.

This thesis is organized as follows: in Chapter 2, some background has been provided along with the literature review of the key techniques employed to address the CPP. The proposed algorithm based on convolutive ICA and IBM followed by the cepstral processing, for the blind separation of convolutive speech mixtures, with systematic evaluation and experimental results for both simulated and real data is described in Chapter 3. In Chapter 4, a novel algorithm is presented for the enhancement of noisy reverberant speech, using EMD based subband processing. It is shown in this chapter that the developed algorithm offers considerable performance improvement for both simulated and real data. Chapter 5 describes a new method for the reduction of room reverberations using the frequency dependent statistical model. The comparison of the algorithm with a related recent approach is given in this chapter based on experimental results for both simulated and real recorded data. In Chapter 6, a new algorithm is presented for blind estimation of RT which is then incorporated into the algorithms developed in Chapter 3 and 5 for performing blind dereverberation and separation from the speech mixtures. Experimental evaluation results are also provided in this chapter. Chapter 7 concludes the thesis with recommendations for future research.

1.2 Contributions

The major contributions of this thesis are summarized as follows:

1) An efficient algorithm is proposed for the blind separation of convolutive speech mixtures. The proposed algorithm is a multistage algorithm with novel combinations of three steps, including the convolutive source separation algorithm adopted in the first

step followed by the estimation of IBM from the separated sources obtained with the convolutive ICA algorithm in the second step, and the cepstral smoothing technique is employed in the third step for reducing the musical noise caused by estimation of IBM. Extensive evaluations have been performed on the proposed algorithm by comparison with related recent approaches in terms of both objective performance indices and subjective listening tests. Results show that the multistage algorithm improves significantly the separation performance over these methods. Moreover, the proposed algorithm is a computationally more efficient one as compared to the recent approach. Pitch frequency is calculated in the proposed multistage algorithm from the segregated speech signal which is different from the method used previously utilizing the estimated mask for the pitch estimation.

2) A novel algorithm is developed to deal with the late reverberations and noise jointly using EMD based subband processing. The results show that this novel method leads to an improved enhancement performance in comparison to a related recent approach.

3) A new method is developed to suppress the room reverberations using the frequency dependent statistical model. In this algorithm, the spectral variance of the late reverberations is estimated based on a frequency dependent statistical model of the decay rate of the late reverberations. For early reflections, a dual-channel Wiener filter is used to reduce their effects. The results indicate that this method performs considerably better in comparison with the most recent methods.

4) An algorithm is proposed for the blind dereverberation and separation together for the convolutive speech mixtures. The proposed algorithm consists of a new method for blind estimation of RT from the reverberant speech signal (i.e., mixtures). A Laplacian distribution based decay model is proposed in which an efficient procedure for locating free decay segments from reverberant speech is also incorporated.

Chapter 2

Background and Literature Survey

2.1 Cocktail Party Problem

This section is focusing on the discussion of one of the most challenging problems within the audio community called CPP [30]. It was proposed to address the phenomenon associated with the human auditory system that, in a cocktail party environment, humans have the ability to focus their listening attention on a single speaker when multiple conversations, background interferences and noise are present simultaneously. The main distortions need to be tackled in CPP are classified as, (1) distortion due to interfering sound, (2) distortion due to room reverberations, and (3) distortion due to background noise. Many researchers and scientists from a variety of research areas attempt to tackle this problem [10,21,23,49]. Despite all these works, CPP remains an open problem and demands further research effort.

As the solution to the CPP offers many practical applications, engineers and scientists have spent their efforts in understanding the mechanism of the human auditory system, and hoping to design a machine which can work similarly to the human auditory system. However, there are no machines produced so far that can perform as humans in a real cocktail party environment. Based on the three different types of distortions that need

to be handled, background and literature review on related methods are provided in this chapter. However, the main contributions of this thesis focus on the first two types of distortions.

2.1.1 Audio sources in a cocktail party environment

Audio sources are usually classified as speech, music, or natural sounds. Each of the categories has its own specific characteristics which can be exploited during its processing. Speech sounds are basically composed of discrete phonetic units called phonemes [39,124]. Due to the co-articulation of successive phonemes, each signal that corresponds to a specific phoneme exhibits time varying properties. The resultant signal is composed of periodic harmonic pulses which are produced due to the periodic vibration of the vocal folds, a noise part which is generated because of the air passing via lips and teeth, or a transient part due to the release of pressure behind the lips or teeth. Harmonics within the generated signal have periodic frequency components which are multiples of a fundamental frequency component. In real speech signals the fundamental frequency component of the periodic phonemes varies due to the articulation, but typically for male speech is 140 Hz, and 200 Hz for female speech with variation of 40 Hz for each.

Music sources [63] generally constitute of sequences of notes or tones produced by musical instruments, singers and synthetic instruments. Each note is composed of a signal which further can be made of a periodic part containing harmonic sinusoids produced by blowing into a pipe, bowing a string, a transient part generated due to hitting a drum, plucking a string, or a wideband noise produced by blowing into wind instruments. For example, in western music the periodic frequencies of the notes generated typically remain constant or varying slowly. Musical instruments usually produce musical phrases which are composed of successive notes without any silence between the notes. Unlike monophonic music, polyphonic sounds are composed of several simultaneous notes that are generated by multiple musical instruments.

The third source comes from the environment, called natural sounds [59]. Their characteristic varies depending on the origin of the natural sound. Similar to the speech

and music signals it can also be classified as periodic, transient and noise. For example, a car horn produces the natural periodic sound signal, a hammer thrashing the hardwood generates the transient signal and raining results in a wideband noise signal. The discrete structure of natural sound is simpler as compared with the organization of notes and phonemes. In this work, the focus will be on the first type of audio source signal i.e. speech signals.

2.2 Distortion Due to Interfering Sound

In order to deal with the distortions generated due to interfering sound in the vicinity, a variety of methods have been proposed. For example, the computational auditory scene analysis (CASA) approach attempts to simulate the human auditory system via mathematical modeling using computational means [142,168,172]. BSS is also used to address this problem by many researchers. [102,121,147,178]. BSS approaches are based on the ICA technique assuming that the source signals coming from different speakers are statistically independent [73, 97]. Non-negative matrix factorization (NMF) and its extension non-negative tensor factorization (NTF) have also been applied to speech and music separation problems [151, 155, 166, 176]. Another interesting approach is the sparse representation of the sources in which the source signals are assumed to be sparse and hence only one of the source signals in the mixture is active while others are relatively insignificant for a given time instant [16, 128, 191]. Some model based approaches have also been employed to address this problem [134, 163]. The following sections provide a detailed review of these techniques.

2.2.1 Computational auditory scene analysis

CASA is the study of ASA by computational means [172]. It is believed that the human ability to function well in everyday acoustic environments is due to a process termed ASA, which produces a perceptual representation of different sources in an acoustic mixture [21]. In other words, listeners organize the mixture into streams that correspond to different sound sources in the mixture. The concept of ASA was

coined by Bregman in 1990 [21]. According to Bregman, organization in ASA takes place in two main steps: segmentation and grouping. In segmentation, the acoustic input (mixture) is decomposed into sensory elements or segments, each of which should primarily originate from a single source. In grouping, the segments that are likely to arise from the same source are grouped together. Segmentation and grouping are guided by ASA cues that characterize intrinsic sound properties, including harmonicity, onset and offset, and location, as well as prior knowledge of specific sounds.

A typical CASA system is shown in Figure 2.1, which has four stages: external processing, feature extraction, segmentation, and grouping and reconstruction. External processing processes the input signal using an auditory peripheral model, resulting in a *cochleagram* which is a two-dimensional time-frequency (T-F) representation. A cochleagram is composed of T-F units, each of which corresponds to the response of a specific auditory filter within a time frame. The second stage extracts auditory features, producing a number of feature representations. In the segmentation stage, the system generates a collection of segments or contiguous regions in a cochleagram. On the basis of extracted features and segments, the grouping and reconstruction stage produces streams corresponding to individual sound sources. The grouping and reconstruction stage includes *simultaneous grouping* which organizes segments overlapping in time into *simultaneous streams*, and *sequential grouping* which organizes segments or simultaneous streams across time into complete streams [34, 35, 168, 172].

In general, there are two types of approaches for the separation of the target signal in the cocktail party environment in the context of CASA. The first one is called signal-driven approach which is used for the segregation of the auditory scene into the different components belonging to the different sound streams [21]. The second one called knowledge-driven approach uses the prior knowledge of the unknown speech sources, so that the target signal can be separated from the interference. In 1994, Brown and Cooke investigated some of the key issues related to the early CASA methods [24]. Specifically they avoid the assumptions made about the type and number of sources. They proposed to model the human auditory system into separate parts. The key parts are ear filtering, cochlear filtering and central processing (combination of different auditory maps which show onset, offset, periodicities and frequency transitions). Wang

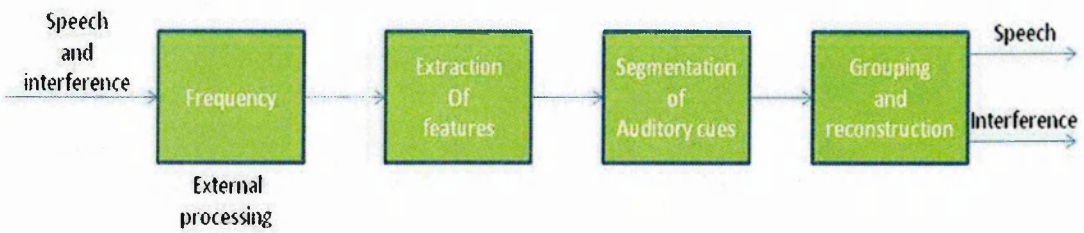


Figure 2.1: Schematic diagram of a typical CASA system

and Brown (1999) [170] extended the work of Brown and Cooke by replacing the central processing with a double layer oscillator network and applied simple computational methods for auditory feature extraction.

A technique called ideal binary masking has been recently used in CASA to segregate the target signal from the interference [172]. Consider a microphone signal recorded in a cocktail party: $x(n) = s_1(n) + s_2(n)$, where $s_1(n)$ is the target speech signal and $s_2(n)$ is the interference speech signal and n is the discrete time instant. Denote X, S_1 and S_2 as the time-frequency (T-F) representation of $x(n)$, $s_1(n)$ and $s_2(n)$ obtained from some T-F transformation respectively. Then, the ideal binary mask (IBM) for $s_1(n)$ with respect to $s_2(n)$ is defined as follows,

$$M_1(m,k) = \begin{cases} 1 & \text{if } \mid S_1(m,k) \mid > \mid S_2(m,k) \mid, \\ 0 & \text{otherwise .} \end{cases} \tag{2.1}$$

where m, k are the discrete time frame and frequency bin indices respectively. The target speech $s_1(n)$ can then be extracted by applying the IBM to X, followed by an inverse T-F transform. The decision is binary, and hence the intelligibility of the segregated speech signal is high. But on the other hand the resultant mask M_1 entirely depends on the availability of the target and interference speech signals. In practice, the target and interference signals are usually unknown, and the mask has to be estimated from the mixtures.

Recently, some methods have been developed in which the limitation of the CASA methods, i.e., having to estimate the IBM directly from the mixtures, is mitigated,

(see for example, [129,145]). In these methods a separation algorithm is applied to the available mixtures to estimate the source signals followed by the estimation of the IBM from these estimated source signals.

Similarly, for the estimation of IBM, spatial localization cues, i.e., interaural time difference (concerning humans, it is the difference in arrival time of a sound between two ears) and interaural level difference (sound from the right side has a higher level at the right ear than at the left ear, because the head shadows the left ear, such difference is called interaural level difference), have also been considered recently (see for example, [65,143]).

2.2.2 Blind source separation

Another technique to address the problem of speech separation is BSS, where the mixing process is usually described as a linear convolutive model and convolutive ICA algorithms can then be applied to segregate the source signals from their mixtures assuming the sources are statistically independent [8,102,107,120,121,129]. BSS is an approach used for the estimation of the source signals having only the information of the mixed signals observed at each input channel, without prior information about sources and the mixing channels. Its potential applications include speech segregation in the cocktail party environment, teleconferences and hearing aids. In such applications, the mixture signals are reverberant, due to the surface reflections of the rooms. ICA is a major statistical tool for the BSS problem, for which the statistical independence between the sources is assumed [73,97]. The mathematical model [1] used to describe ICA is given as,

$$\begin{aligned} x_1(n) &= a_{11}s_1(n) + a_{12}s_2(n) + ...a_{1N}s_N(n) \\ &\cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \\ x_M(n) &= a_{M1}s_1(n) + a_{M2}s_2(n) + ...a_{MN}s_N(n) \end{aligned}$$

where $s_1(n), .., s_N(n)$ represent unknown source signals in the cocktail party environment, $x_1(n), .., x_M(n)$ denote the mixture signals (e.g. microphone recordings). If the coefficients a_{ij} $(i = 1, .., M,\ j = 1, .., N)$ are scalars, the resultant mixtures are referred

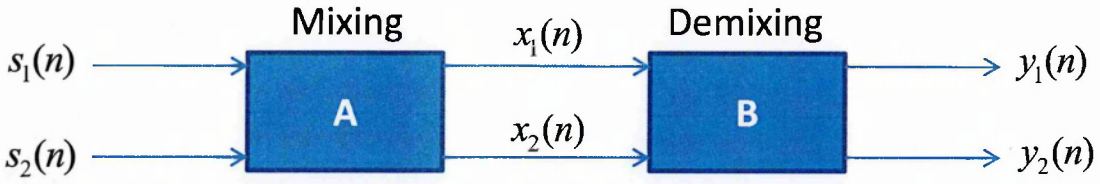


Figure 2.2: Schematic diagram for a typical BSS system with two sources and two mixtures. Unknown source signals: s, observed signals: x, estimated signals: y

to as instantaneous mixtures, and if they are filters, the mixtures are referred to as convolutive mixtures. If N=M, i.e., the number of sources equals to the number of mixtures, it is called exactly determined BSS problem. If $N > M$, it is the under-determined case, and $N < M$ the over-determined BSS problem. A schematic diagram of a typical two input two output BSS system is given in Figure 2.2, in which **A** represents the unknown mixing system and **B** is the demixing system used for the estimation of the unknown source signals.

For separating convolutive mixtures, the BSS approach using ICA can be applied either in the time domain [32, 45, 129] or in the frequency domain [8, 102, 121, 136, 178] or their hybrid [90, 98], assuming that the source signals are statistically independent. The time-domain approaches attempt to extend the instantaneous ICA model to the convolutive case. They can achieve good separation performance once the algorithms converge, as the independence of segregated signals is measured accurately [102]. However the computational cost for the estimation of the filter coefficients in the convolutive operation can be very demanding, especially when dealing with reverberant mixtures using long time delay filters [5, 25, 44, 46, 104].

To improve the computational efficiency, the frequency domain BSS approaches transform the mixtures into the frequency domain, and then apply an instantaneous but complex valued ICA algorithm to each frequency bin [8, 111, 126, 147, 152, 178, 189]. In [8] the authors discussed why the separation performance of frequency domain BSS is poor when there is long reverberation. First, they have shown that it is not good to be constrained by the condition that the frame size of the FFT should be greater than the length of a room impulse response. This is because the lack of data causes the collapse of the assumption of independence between the source signals in each frequency bin

when the data length is short, or when a longer frame size is used. On the other hand, they have shown that a short frame also results in a poor performance, because long reverberation can not be covered by a short frame. Therefore, there is an optimum frame size determined by a trade-off between maintaining the assumption of independence and covering the whole reverberation in frequency domain BSS. Similarly a new type of non-linear function has been suggested in [147] for an ICA approach in order to process the complex numbers. The function has been derived from the probability density function of the signals in the T-F domain with the assumption of phase independence between these signals. The new non-linear function is obtained as a result, based on the polar coordinates of a complex number. The effect of this new function has also been analysed in [147] for separating speech signals in the convolutive environment. Another very interesting approach employed for frequency-domain BSS is adaptive and based on second order statistics [152]. The advantage of this method is that no parameter tuning is required for separating the signals. As a result, many complex valued and instantaneous ICA algorithms that have already been developed can be directly applied to the frequency domain BSS. However, an important issue associated with this approach is the permutation problem, i.e., the permutation in each frequency bin may not be consistent with each other so that the separated speech signal in the time domain contains the frequency components from the other sources. Different methods have been developed to solve this problem. By reducing the length of the filter in the time domain [25, 126] the permutation problem can be overcome to some extent. A source localization approach has also been employed to mitigate the permutation inconsistency [148, 159]. Another technique for the alignment of the permutations across the frequency bands is based on correlation between the separated source components at each frequency bin using the envelope similarity between the neighboring frequencies [112]. Some other recently used methods are based on the physical behaviour of the acoustic environment [118] or coherent source spectral estimation [119], the method for modeling frequency bins using the generalized Gaussian distribution [105].

The third approach is the combination of both time and frequency domain approaches. In some methods [12, 98], the coefficients of the FIR filter are updated in the frequency domain and the non-linear functions are employed in the time domain for evaluating the

independence of the source signals. Hence no permutation problem exists any more, as the independence of the source signals is evaluated in the time domain. Nevertheless, the limitation of this hybrid approach is the frequent switch between two different domains at each step and thereby consuming extra time on these inverse transformation operations.

The separation performance of many developed algorithms is however still limited, and there is much room for improvement. This is especially true when dealing with reverberant and noisy mixtures. For example in the frequency-domain BSS framework, if the frame length of the DFT is long and the number of samples in each frequency bin is small, the independence assumption may not be satisfied. Similarly, if the short length DFT frame is used, the long reverberations cannot be covered and hence the segregation performance is limited [8].

Apart from the above discussed methods, some authors consider the assumption of W-disjoint orthogonality for speech signals in order to separate the source signals from the observed data. For example in [80], for a given windowing function $W(n)$, two sources, $s_i(n)$ and $s_j(n)$ are called W-disjoint orthogonal if the supports of the short-time Fourier transform of $s_i(n)$ and $s_j(n)$ are disjoint [80]. The windowed Fourier transform of $s_i(n)$ is defined as,

$$s_i^W(m,k) = \sum_{n=0}^{N-1} W(n-m)s_i(n)e^{-i2\pi kn/N} \tag{2.2}$$

The W-disjoint orthogonality assumption can be expressed as below [80].

$$s_i^W(m,k)s_j^W(m,k) = 0,\ \forall i \neq j,\ \forall k,m \tag{2.3}$$

where k and m are the frequency index and time frame index respectively. This equation implies that either of the sources is zero for any k and m as long as two sources do not come from the same source. If $W(n) = 1$, then $s_i^W(m,k)$ can be interpreted as the Fourier transform of $s_i(n)$, which can then be referred to as $s_i(k)$. Therefore, W-disjoint orthogonality can be written as,

$$s_i(k)s_j(k) = 0,\ \forall i \neq j,\ \forall k \tag{2.4}$$

which represents the property of disjoint orthogonality [80].

Another challenging problem is to separate moving sources rather than stationary in a cocktail party environment. A recent work [114] is devoted to the blind separation of moving sources. Here a multimodal approach is proposed for the segregation of moving speech sources. The key issue in blind estimation of moving sources is the time varying nature of the mixing and unmixing filters, which is hard to track in the real world. In this work the authors applied the visual modality for the separation of moving sources as well as stationary sources. The 3-D tracker based on particle filtering is used to detect the movement of the sources. This method performs well for the blind separation of moving sources in a low reverberant environment.

So far, two important techniques for convolutive speech separation were discussed in detail. It is interesting to make a comparison between these two techniques. In the case of BSS, the unknown sources are assumed to be statistically independent. However, no such assumption is required for CASA. On the other hand, the IBM technique used in the CASA domain needs to estimate the binary mask from the target and interference signals which should be obtained from the mixture in practice. Another difference is in the way how the echoes within the mixture are dealt with by these two techniques. In BSS algorithms [8, 102, 121, 178], this is modeled as a convolutive process. On the other hand CASA approaches deal with echoes based on some intrinsic properties of audio signals, such as, pitch, which are usually preserved (with distortions) under reverberant conditions. However, the human auditory system has a remarkable ability of concentrating on one speaker by ignoring others in a cocktail party environment. Some of the CASA approaches [171] work in a similar manner i.e. extracting a target signal by treating other signals as background sound. In contrast, BSS approaches attempt to separate every source signal simultaneously from the mixture. Motivated by the complementary advantages of the CASA and BSS approaches, a multistage approach is developed in [76, 77] where a convolutive BSS algorithm is combined with the IBM technique followed by cepstral smoothing. The details of this method will be discussed later in Chapter 3.

2.2.3 Model based approaches

Another method to address the speech separation problem is based on the statistical modeling of signals and the parameters of the model are estimated from the training data, e.g., [74,134,135,163]. In [163], a Gaussian mixture model (GMM) is employed for modeling of the joint probability density functions (pdf) of the sources by exploiting the non-Gaussianity and/or non-stationarity of the sources and hence the statistical properties of the sources can vary from signal to signal.

In [134] the model-based approach is used for single channel speech separation. The authors considered the problem as a speech enhancement problem in which both the target and interference signals are non-stationary sources with the same characteristics in terms of pdf. Firstly, in the training phase, the patterns of the sources are obtained using Gaussian composite source modeling. Then the patterns representing the same sources are selected. Finally, the estimation of the sources can be achieved using these selected patterns. Alternatively, a filter can be built on the basis of these patterns and then applied to the observed signals in order to estimate the sources.

Source separation in the wavelet domain by model-based approaches has been considered in [74]. This method consists of a Bayesian estimation framework for the BSS problem where different models for the wavelet coefficients have been presented. However there are some limitations with the model based approach. The trained model can only be used for the segregation process of the speech signals with the same probability distribution, i.e., the pdf of the trained model must be similar to that of the observation data. In addition, the model based algorithms may perform well only for a limited number of speech signals.

2.2.4 Non-negative matrix/tensor factorization

Non-negative matrix factorization (NMF) was proposed by Lee and Seung in 1999. Using the constraint of non-negativity, NMF decomposes a non-negative matrix $\mathbf{V}$ into the product of two non-negative matrices $\mathbf{W}$ and $\mathbf{H}$, given as:

$$\mathbf{V}_{m\times n} = \mathbf{W}_{m\times r}\mathbf{H}_{r\times n} \tag{2.5}$$

where $(n+m)r < mn$. Unlike other matrix factorizations, NMF allows only additive operations i.e. no subtractions [92, 95, 96]. As NMF does not depend on the mutual statistical independence of the source components, it has a potential to segregate the correlated sources. NMF has been applied to a variety of signals including image, speech or music audio. In [33] the authors attempted to separate the general form of signals from the observed data i.e. both positive and negative signals using the constraints of sparsity and smoothness. For machine audition of audio scenes, NMF has also found some applications. For example, it has been applied to music transcription [157, 167] and audio source separation [51, 52, 127, 150, 155, 156, 166, 167, 174, 176, 177]. In these applications, the audio data are usually transformed to non-negative parameters, such as spectrogram, which are then used as the input to the algorithms. The application of the NMF technique to speech separation is still an emerging area which attracts increasing interests in the research community.

2.2.5 Sparse representation and compressed sensing

Separation of signals blindly from their under-determined mixtures has attracted a great deal of attention over the past few years. It is a challenging source separation problem. One of the most common methods adopted for this problem is based on the sparse representation of signals [37, 50, 191, 192]. Closely related to sparse representation, there is an emerging technique called compressed sensing, which suggests that a signal can be perfectly recovered based on information rate, instead of the Nyquist rate, and random sampling, instead of uniform sampling, under certain conditions. It has been observed that compressed sensing exploits two important properties [26–28, 41]. The first one is sparsity, which means that many natural signals can be represented in some proper basis in sparse (compressible) form. The second property is incoherence, i.e. the signal which is represented in some proper basis in sparse form should be dense as compared to the original representation of the signal. It is basically the extension of duality property between time and frequency domain.

There are similarities between the compressed sensing and source separation and their connections have been explored by [15], and further investigated by [184, 185]. It was

found that the compressed sensing based signal recovery methods can be applied to the source reconstructions provided that the unmixing matrix is available or has been estimated [15, 37, 50, 191, 192].

2.3 Distortion Due to Room Reverberation

2.3.1 Characteristics of reverberation

Reverberation is caused by the multi-path propagation of an acoustic signal from its source to the microphone. Room reverberation is introduced due to surface reflections within a room, as illustrated in the Figure 1.1. Both the speakers produce wavefronts propagating outward, with some reaching the microphones directly and some others reflecting off the walls and superimposing at the microphones. The energy and phase of the reflections reaching the microphones are different from those of the direct signals due to the differences in the length of the propagation paths. As a result, delayed and attenuated copies of the source signal are present in the microphone signals, described as reverberation [61, 93, 115].

The signal received at the microphone is generally composed of a direct sound coming from the source to the microphone, reflections that arrive shortly after the direct sound (also called *early reflections*), and reflections that arrive after early reverberation (commonly known as *late reverberation*). The combination of direct sound and early reflections are sometimes named as *early sound component. Early reverberation* is not perceived as a separate sound to the direct sound as long as the delay of the reflections does not exceed a limit of approximately 80-100 msec with respect to the arrival time of the direct sound, however it can be perceived to reinforce the direct sound and is therefore considered useful with regard to speech intelligibility. This phenomenon is often referred to as the precedence effect. *Early reverberation* mainly causes spectral distortion due to non-flat frequency response called colouration. *Late reverberation* which arrives at the microphone with longer delays is perceived as separate echoes or as reverberation and impairs speech intelligibility. This is due to the two masking effects introduced by the late reverberations, namely self masking where the speech

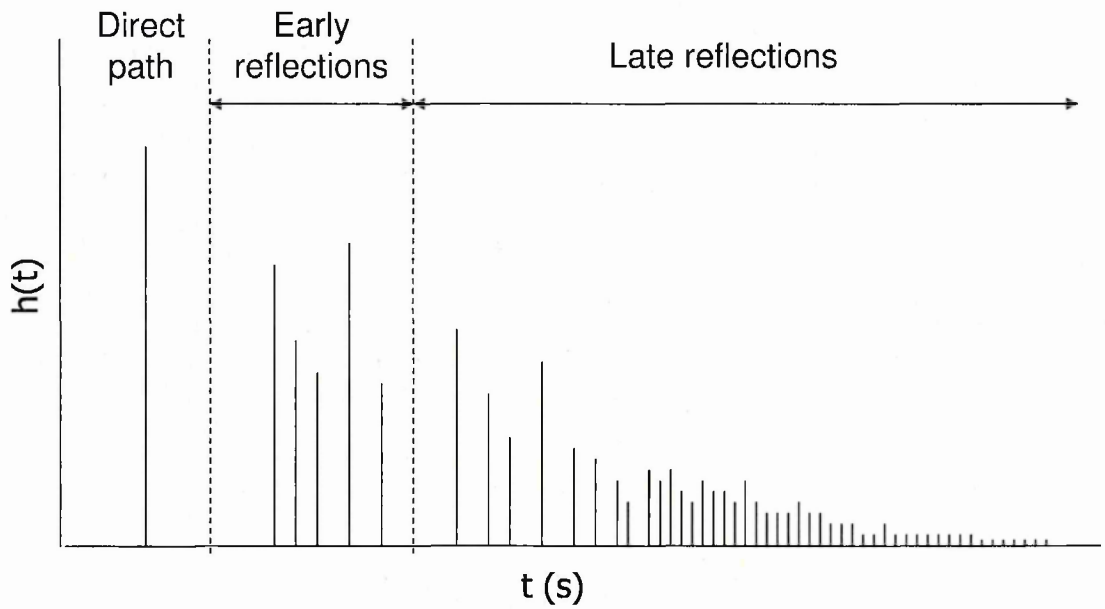


Figure 2.3: Schematic diagram for room impulse responses.

spectrum is smeared by the late reverberations, and overlap masking where the energy of the preceding phoneme overlaps with that of the subsequent phonemes. It can have severe effects on the performance of automatic speech recognition (ASR) systems. Also it is one of the main factor in performance degradation of the source separation algorithms [61, 93, 115].

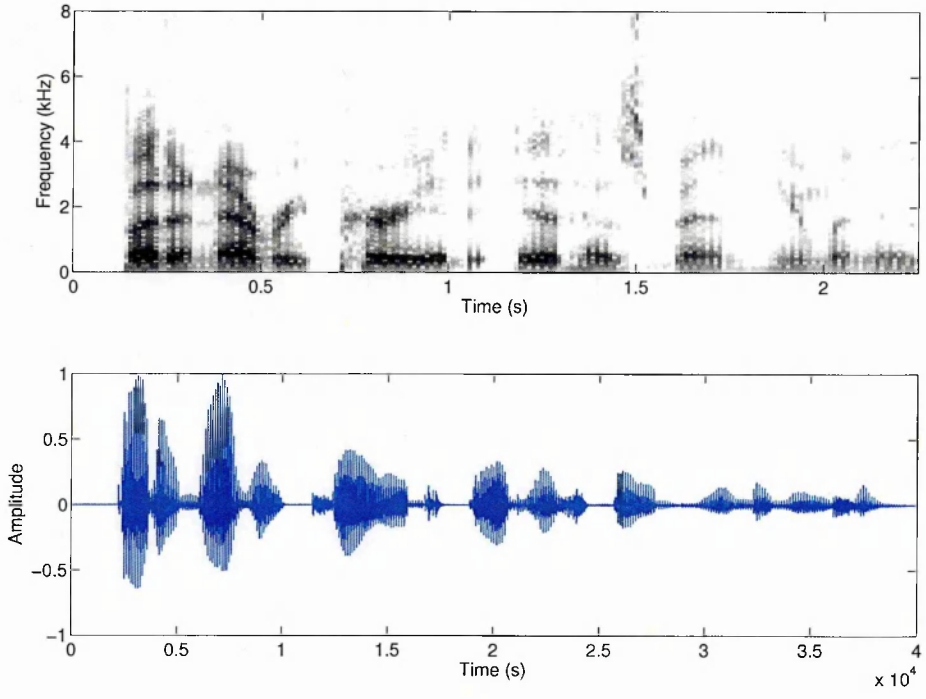
The behaviour of the acoustic channel between the source and microphone can be characterized by a room impulse response (RIR). It represents the signal recorded at the microphone in response to a source that generates a sound impulse. As shown in Figure 2.3, the RIR can be split into three main sections, the direct path, the early reflections and late reflections. The direct sound, early reverberations and late reverberations are the convolution of these segments with the desired signal. Additionally, it is also observed that the energy of the reflections decays at an exponential rate. This exponential decay property of the RIR gives rise to the concept of reverberation time (RT). It is defined as the time required for the average sound-energy at a given frequency to reduce to one-millionth of its initial steady-state value after the sound source has been switched off and this corresponds to a decrease of 60 decibels (dB).

Now to explain the effects of reverberation on speech perception, an example is given

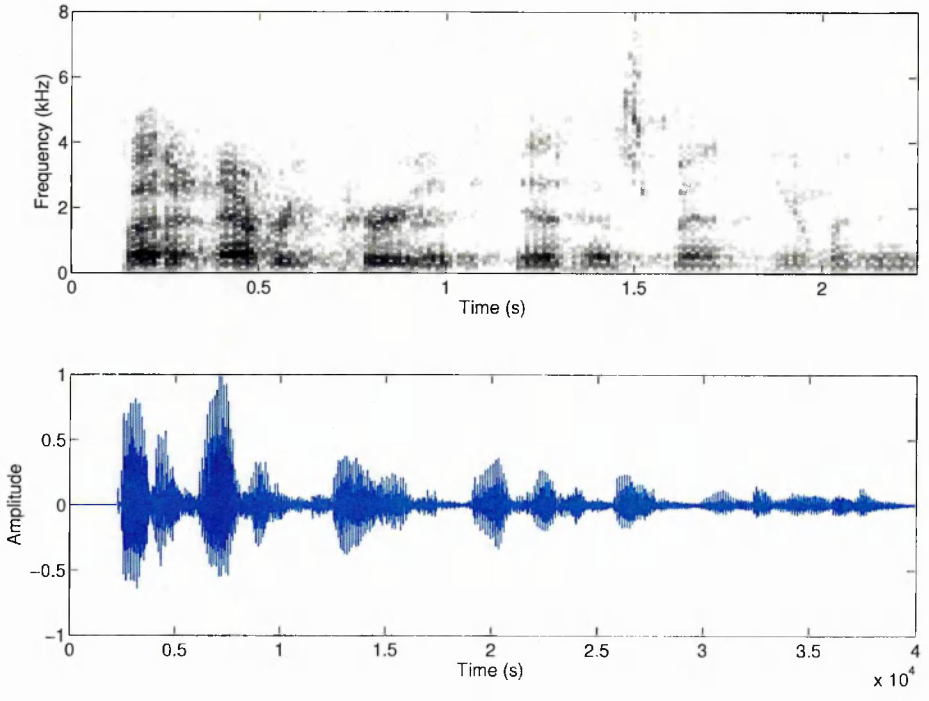
in Figure 2.4. The effects of reverberation are clearly visible and audible in the spectrogram and waveform of a speech signal. The Figure 2.4(a) shows the spectrogram and waveform for an anechoic speech signal taken from the TIMIT database sampled at 16 kHz. The speech formants (resonance frequencies affiliated with the vocal tract [72]) are clearly visible in the spectrogram in this figure. Similarly, phonemes are differentiable in the waveform. The simulated room model [4] is used to generate the reverberant signal from the anechoic speech signal at RT = 0.5 sec with a source-microphone distance of 1 m. The spectrogram and waveform of the reverberant speech signal are shown in Figure 2.4(b). The distortion caused by the acoustic channel is visible in both the spectrogram and the waveform. In the spectrogram a blurring effect is visible, while in the waveform smearing of the phonemes can be seen. These distortions result in an audible difference between the anechoic and the reverberant speech, and hence degraded speech intelligibility. Hence methods should be developed to reduce such detrimental effects of reverberation on the speech signal. Therefore, in this thesis two algorithms are developed to deal with the reverberations. The details of both the developed methods will be discussed in Chapters 4 and 5.

2.3.2 Approaches for reverberation suppression

In the literature many methods have been proposed to deal with the effects of room reverberation, including for example, the dereverberation algorithms based on inverse filtering [38, 58, 85, 108, 109, 117, 160, 188], cepstral filtering [13, 123, 164], temporal envelop filtering [11, 91, 110], information using source excitation [186, 187], and methods based on spectral processing [3, 53, 94, 125, 179]. These methods can be broadly classified into three categories, *spectral processing methods* such as spectral subtraction assisted methods, *temporal processing methods* such as inverse filtering, cepstral filtering, temporal envelop filtering, and methods based on excitation source information, and *spectral-temporal methods* such as methods based on the combination of temporal and spectral processings.



(a) Spectrogram (top) and waveform (bottom) of an anechoic speech signal



(b) Spectrogram (top) and waveform (bottom) of the reverberant speech signal

Figure 2.4: Spectrograms and waveforms of (a) an anechoic speech signal taken from the TIMIT speech database, and (b) the reverberant version of this measured at a distance of 1 m, with a reverberation time of 0.5 sec using a simulated room model [4].

Temporal processing methods

Oppenhiem *et al.* [123] proposed a dereverberation method based on a low time cepstral liftering technique for a single microphone. Cepstral liftering in low time is equivalent to low-pass filtering in the time-frequency domain. The idea is based on the observation that the clean speech cepstrum is mainly concentrated in the low time, i.e., close to the origin unlike the acoustic channel impulse response which is located far away from the origin. However it is practically difficult to find the proper cutoff time for low time liftering [13, 164].

Another technique used commonly to reduce the reverberation is based on inverse filtering. The key idea is to recover the original signal by passing the reverberant signal through a filter that inverts the reverberant signal [38, 58, 85, 108, 109, 117, 160, 188]. Inverse filter can help in successful dereverberation if the room impulse response is known, or blindly estimated. This is known to be a difficult task. Recently, Kinoshita *et al.* [85] proposed a dereverberation algorithm that estimates the energy of late reverberant components based on the concept of inverse filtering, named as long-term multiple step linear prediction. Firstly, they used long-term multiple step linear prediction to estimate the energy of late reverberations in the time domain. Then they convert the late reverberant signal into the frequency domain and subtract its power spectrum from that of the observed signal.

Temporal envelope filtering based algorithms were proposed in [11]. The main theme of this method is that the clean speech signal is produced inside an enclosure (envelope) having fine details of time-intensity distribution. Reverberations added to such a clean speech signal have a blurring effect on its envelope, because of the reflections of different intensities and delays added to the clean speech. Hence the average envelope modulation spectrum of the clean speech can be recovered from the reverberant speech by filtering the time trajectories of spectral bands in reverberant speech [11, 91, 110].

Yegnanarayana and Murthy developed a reverberant speech enhancement method by manipulating the excitation source information that is contained in the linear prediction (LP) residual signal, based on the characteristics of the LP residual of reverberant speech [186]. The processing method involves identifying and manipulating the resid-

ual signal in different regions of the reverberant speech, namely, regions which is high signal-to-reverberation ratio (SRR), low SRR, and only reverberant. A weight function is derived at gross and fine levels to modify the LP residual signal. In [187], Yegnanarayana *et al.* proposed a multichannel reverberant speech enhancement technique by exploiting the features of the excitation source in speech production. The authors use time-aligned Hilbert envelopes to represent the strength of the peaks in the LP residual. The Hilbert envelopes are then summed and used as a weight function which is applied to the LP residual of one of the microphones. In most of the LP residual-based methods, it is assumed that room reverberation would introduce only zeros into the microphone signals and, as a result, would primarily affect only the nature of the speech excitation sequence, having little impact on the all-pole filter [14]. Therefore, speech dereverberation can be accomplished by processing only the speech excitation signal, leaving the LP coefficients untouched.

Spectral processing methods

Spectral based processing of reverberant speech is another common approach used in the literature [3, 53]. In [94], Lebart *et al.* introduced a single channel speech dereverberation method based on spectral subtraction to reduce the reverberation effect. The reverberation suppression method based on spectral subtraction is not sensitive to fluctuations in the impulse response. The method estimates the power spectrum of the reverberation based on a statistical model of late reverberation and then subtracts it from the power spectrum of the reverberant speech. The authors assumed that the reverberation time is frequency independent and the energy related to the direct sound could be ignored. The authors also assume that the SRR of the observed signal is smaller than 0 dB which limits the use of the proposed solution to situations in which the source-microphone distance is smaller than the critical distance (The distance between source and microphone at which the direct path energy is equal to the combined energy of the early and late reflections).

Wu and Wang [179] proposed a two-stage model to enhance reverberant speech. In the first stage, an inverse filter of the room impulse response is estimated, to increase the

SRR by maximizing the kurtosis of the LP residual to reduce the early reflections. In the second stage, late reverberation effects are removed by a spectral subtraction approach. The maximum kurtosis part [58] employed in [179] requires at least 500 iterations to obtain the inverse filtered speech. However, as mentioned in [179], if the inverse filter is not precisely estimated, inverse filtering may even degrade the reverberant speech rather than improve it. In [56] a similar two-stage approach is proposed using multichannel blind deconvolution with spectral subtraction for the enhancement of reverberant speech.

Spectral-temporal methods

In [57], the authors proposed a reverberant speech enhancement algorithm using spatio-temporal and spectral processing. The speech signals are first spatially averaged followed by temporal larynx cycle averaging of the LP residual of the voiced speech to primarily attenuate the early reverberation. This is followed by spectral subtraction to attenuate the late reverberation. This method takes the advantage of a multi-microphone system for spatial averaging. A similar two-stage single-microphone system is also developed in [60]. In the first stage, the spectral processing technique proposed in [61] is used to suppress late reverberation. In the second stage, the early reflections are suppressed by the LP residual processing in a similar way as in [57]. The basis is that the waveform of the LP residual between adjacent larynx-cycles varies slowly, so that each such cycle can be replaced by an average of itself and its nearest neighboring cycles. The averaging results in the suppression of spurious peaks in the LP residual caused by room reverberation. The dynamic programming projected phase-slope algorithm (DYPSA) algorithm [116] is employed for automatic estimation of glottal closure instants in voiced speech. However, no attempt is made to eliminate spurious instants detected in the unvoiced and silence regions by the DYPSA algorithm. Therefore, a high and low SRR region detector needs to be incorporated in [57] and [60] to eliminate spurious instants.

Recently, an algorithm has been proposed in [88] for the enhancement of reverberant speech based on the combination of temporal and spectral processing. In this method,

spectral processing is performed first, and in the second step the spectrally-processed speech signal is then subjected to temporal processing. The main reason behind this spectro-temporal processing is the identification of high SRR regions, primarily when the RT is high. Due to the convolutive nature of reverberant speech, low SRR and reverberation-only regions (late reverberant regions) also look like speech signals that makes it difficult to separate low and high SRR regions. Therefore, spectral processing is first performed in [88] to eliminate the late reverberant regions and then temporal processing is performed.

Another technique presented in [67] by Hazrati *et al.* proposed a multi-stage subband-based blind dereverberation algorithm suitable for reverberant speech enhancement. The proposed algorithm operates by first splitting the reverberant inputs into different subbands. In the second stage, the inverse filters are estimated using the blind deconvolution multiple input-output inverse-filtering theorem based approach, while in the third-stage power spectrum of the late impulse components are subtracted from the power spectrum of the inverse filtered speech in order to suppress the late reverberant energy.

Lebart *et al.* [93] proposed a statistical model for late reverberations. With this model, the spectral variance of the late reverberations can be estimated from the reverberant speech [93]. This work has been carried out further by Jeub *et al.* for the suppression of late reverberations [78]. This original model was developed as frequency independent where a fixed reverberation time (RT) was used for all the frequency channels in the estimation of the decay rate of room reverberations. However, it was suggested by Habets *et al.* [62] that the spectral variance of the late reverberations can be more accurately estimated if a frequency dependent statistical model is adopted. Such an idea will be explored in Chapter 5.

2.4 Distortion Due to Background Noise

Background noise is another form of interference affecting the speech quality and intelligibility. Although, this thesis is not focussing on the distortions caused by the background noise, a novel algorithm is developed in this thesis to enhance the noisy

reverberant speech (will be discussed in detail in Chapter 4), based on the EMD technique. Therefore, it is necessary in this thesis to provide background and literature review of interference by background noise, with focus on the EMD technique.

2.4.1 Conventional methods for noise reduction

Before describing the EMD based denoising techniques, a brief overview of the classical methods for the enhancement of noisy speech is provided here. Different noise reduction methods have been proposed in the literarure, particularly in the case of additive white Gaussian noise [42,43,47,132,149,158]. When noise estimation is available, then filtering gives accurate results. Linear methods such as Wiener filtering [132], and the method based on MMSE filtering [47] are also used because linear filters are easy to implement and design. These linear methods are not so effective for signals presenting sharp edges or impulses of short duration. Furthermore, real signals are often nonstationary. In order to overcome these shortcomings, nonlinear methods have been proposed and especially those based on wavelets thresholding [42,43]. The idea of wavelet thresholding relies on the assumption that signal magnitudes dominate the magnitudes of noise in a wavelet representation so that wavelet coefficients can be set to zero if their magnitudes are less than a predetermined threshold [42]. A limitation of the wavelet approach is that basis functions are fixed, and thus do not necessarily match all real signals.

2.5 EMD for data analysis

EMD has been proposed recently as one of the versatile methods for the analysis of non-stationary and nonlinear data. The idea was given by Huang *et al.* [71] for analyzing non-stationary and nonlinear processes. The major benefit of the EMD is that basis functions are derived adaptively from the data itself unlike the traditional methods where basis functions are fixed. EMD extracts, sequentially and intrinsically, the energy associated with various intrinsic time scales in the signal. The output components after this extraction are named as intrinsic mode functions (IMF), starting from high frequency to lower ones. As the phenomena occurring naturally are non-stationary

and nonlinear, EMD can be a useful tool for their analysis. In the literature many applications of EMD can be found towards the analysis of climate and speech data, as both of them are complicated and contain rich properties [55,71,181]. In the context of speech, literature shows that EMD plays an important role in the algorithms employed for the enhancement of noisy speech signals [18–20,54,83,141,180].

Historically, Fourier analysis has dominated the data analysis efforts since it has been introduced and still used for different kinds of data. Although Fourier analysis can be used for the data under very general conditions, it imposes some very important restrictions on the system under observation: the system must be linear and the data must follow a periodic pattern or must be stationary [71,181]. Besides Fourier analysis, other non-stationary methods were used by the research community for the analysis of data. For example wavelet analysis, smoothing by moving averaging, the spectrogram and least squares estimation of the trend. Further details can be found in many fundamental data processing books, (see, for example, [22]).

2.5.1 EMD for noise reduction

Several works have explored the use of EMD for noise reduction and noisy speech enhancement. Rilling *et al.* in [141] examined the usefulness of the EMD technique towards the analysis of a more general form of white Gaussian noise, i.e., fractional Gaussian noise. The estimation of the scaling exponents has also been studied and explored. Similarly, Flandrin *et al.* in [55] investigated the advantages of EMD in the analysis of fractional Gaussian noise. They found that EMD behaves like a dyadic filter bank. Recently, a method is proposed in [29] for the enhancement of a noisy speech signal using adaptive EMD. The main idea is to combine adaptive noise cancellation with the EMD technique in order to improve the performance in terms of enhancement. The noisy signal is decomposed into its IMFs and adaptive noise cancellation is applied on an IMF level.

In [83] the authors proposed a method for the enhancement of noisy speech signals based on the idea of thresholding the IMFs obtained from noisy speech using hard or soft shrinkage. They proposed two strategies for the noise reduction named as EMD-

shrinkage in which EMD is incorporated with hard shrinkage, and EMD-MMSE in which EMD has been combined with minimum mean squared error (MMSE) filter. The enhanced signal is reconstructed from the processed IMFs. The method based on an EMD-MMSE filter in [83] will be explored in the method proposed in Chapter 4 of this thesis. Similarly in [20] an algorithm has been developed for the noisy speech enhancement based on EMD. The Savitzky-Golay filter and soft thresholding are investigated in this method.

Another recent technique investigated in [82] explores the performance of EMD for the enhancement of noisy speech signals. The adaptive centre weighted average filter which works in the time domain is combined with EMD. The authors claimed that in the context of noise reduction, an adaptive weighted average filter works better on IMF components rather than the full-band noisy speech signal. Similarly, in [81] an algorithm was proposed for the denoising of the voiced speech based on EMD associated with an appropriate sifting process. The noisy speech signal is decomposed into its corresponding IMFs. As the noise is mainly occupying the lower order IMFs (high frequency components), whereas the speech signal energy is focussed into the low frequency IMF components. Hence an adaptive weighting average filter has been used for the high frequency IMFs only rather than all the derived IMF components. In this thesis, the interesting IMFs properties are exploited, and an algorithm is developed for dealing with both additive noise and late reverberations, as explained in Chapter 4.

2.6 Summary

In this chapter a general review has been provided for the issues related to CPP and the different solutions proposed. Firstly, classification of audio source in a cocktail party has been discussed. Then, different types of distortions present in a cocktail party environment have been analysed. The distortions generated due to interfering sound in the vicinity and the different methods proposed to deal with such distortions have been discussed, i.e., CASA approaches, methods under the framework of BSS, NMF/NTF based methods, sparse representation and compressed sensing, and model based approaches. Similarly, room reverberations also caused distortions and as a result

affected the speech quality and intelligibility. Therefore, in this chapter characteristics of the room reverberations have been discussed in detail followed by the different methods proposed in the literature for the treatment of such reverberations. Another source of distortion is the background noise and hence different methods used for the reduction of such noise have been reviewed, with a particular emphasis on the EMD based denoising methods. In subsequent chapters, contributions will be presented for dealing with each of the above three types of distortions.

Chapter 3

A Multistage Approach to Blind Separation of Convolutive Speech Mixtures

This chapter addresses the problem of separating convolutive speech mixtures using the two-microphone recordings, based on the combination of independent component analysis (ICA) and ideal binary mask (IBM), together with a post-filtering process in the cepstral domain. The proposed algorithm consists of three steps. First, a convolutive ICA algorithm is applied to separate the source signals from two-microphone recordings. In the second step, an IBM is estimated by comparing the energy of the corresponding time-frequency (T-F) units from the separated sources obtained with the convolutive ICA algorithm. The last step is to reduce musical noise caused by T-F masking using cepstral smoothing. The performance of the proposed approach is evaluated using both reverberant mixtures generated using a simulated room model and real recordings in terms of both objective measurements and subjective listening tests. The proposed algorithm offers considerably higher efficiency and improved speech quality while producing similar separation performance compared with a recent approach.

3.1 Introduction

As discussed in Chapter 2, both ICA and IBM techniques can be used to address the problem of the separation of source signals from their convolutive mixtures. However the separation performance of many developed algorithms based on ICA is still limited, and leaves much room for further improvement, especially when dealing with reverberant and noisy mixtures. Similarly, the separation algorithms developed for the convolutive speech mixtures based on IBM technique required *prior* knowledge of both the target speech and interfering signal. However, in practice, only mixtures are available, and therefore only the IBM estimated from the mixtures can be used, which itself is a major computational challenge. To overcome the limitations of both the ICA and IBM techniques, an effective algorithm is developed in this chapter in which both the methods are combined such that the IBM can be estimated from the intermediate separation results that are obtained by applying an ICA algorithm to the mixtures. The errors generated due to estimation of the IBM are mitigated by cepstrum based processing method.

The proposed approach in this chapter is essentially motivated by Pedersen *et al.* [129] who proposed a method for the blind separation of source signals in which the IBM has been estimated from intermediate separation results that are obtained by applying an ICA algorithm to the mixtures. The limitation of the CASA methods as mentioned in Chapter 2, i.e., having to estimate the IBM directly from the mixtures, is mitigated as the IBM can now be estimated from the coarsely separated source signals obtained by ICA algorithms. The estimated IBM can be further used to enhance the separation quality of the coarsely separated source signals. Such a combination was shown to achieve good separation performance. However, both the mixing model and separation algorithm considered in [129] are instantaneous, which in practice may not be sufficient for real recordings. In this chapter, combination of ICA and IBM techniques is explored for the separation of convolutive speech mixtures by using a convolutive mixing model and a convolutive separation algorithm. Another related work was proposed in [145] where the target speech is extracted from the mixture using ICA and time-frequency masking. However, a common problem with T-F masking is the errors introduced in

the estimation of the binary mask which has not been well addressed. To deal with the estimation errors of the binary mask, a cepstrum based processing method is employed here.

In the algorithm proposed in this chapter, first a convolutive ICA method is applied [178] to the microphone recordings. As is common with many other existing ICA algorithms, the separated target speech from this step still contains a considerable amount of interference from other sources. The performance steadily degrades with an increase of reverberation time. In order to reduce the interference within the target speech, the IBM is estimated by comparing the energy of the corresponding T-F units from the outputs of the convolutive ICA algorithm, and then applied to the original mixtures to obtain the target speech and interfering sources. As will be confirmed in the experiments, this process considerably improves the separation performance by reducing the interference to a much lower level. However, a typical problem with the binary T-F masking is the introduction of errors in the estimation of the masks. The errors may result in some isolated T-F units, causing fluctuating musical noise [7,101].

The estimated IBM is further processed using cepstral smoothing [101]. More specifically, the binary mask is transformed into the cepstral domain, and the transformed mask is smoothed over time frames using the overlap-and-add technique. In the cepstrum domain, it is easier to distinguish between the unwanted isolated random peaks and mask patterns resulting from the spectral structure of the segregated speech. Therefore, different levels of smoothing can be applied to the binary T-F mask in different frequency ranges. The smoothed mask, after being transformed back into the T-F plane, is then applied to the outputs of the previous step in order to reduce the musical noise.

The proposed approach is essentially a multistage algorithm, as depicted by a block diagram in Figure 3.1 for two microphone mixtures. In the first stage, convolutive speech mixtures $x_1(n)$ and $x_2(n)$ are processed by the convolutive ICA algorithm in [178], where n represents the discrete time index. The resultant estimated source signals of this stage are denoted as $y_1(n)$ and $y_2(n)$. In the second stage, the T-F representations of $y_1(n)$ and $y_2(n)$ are used to estimate the IBM, and the resultant

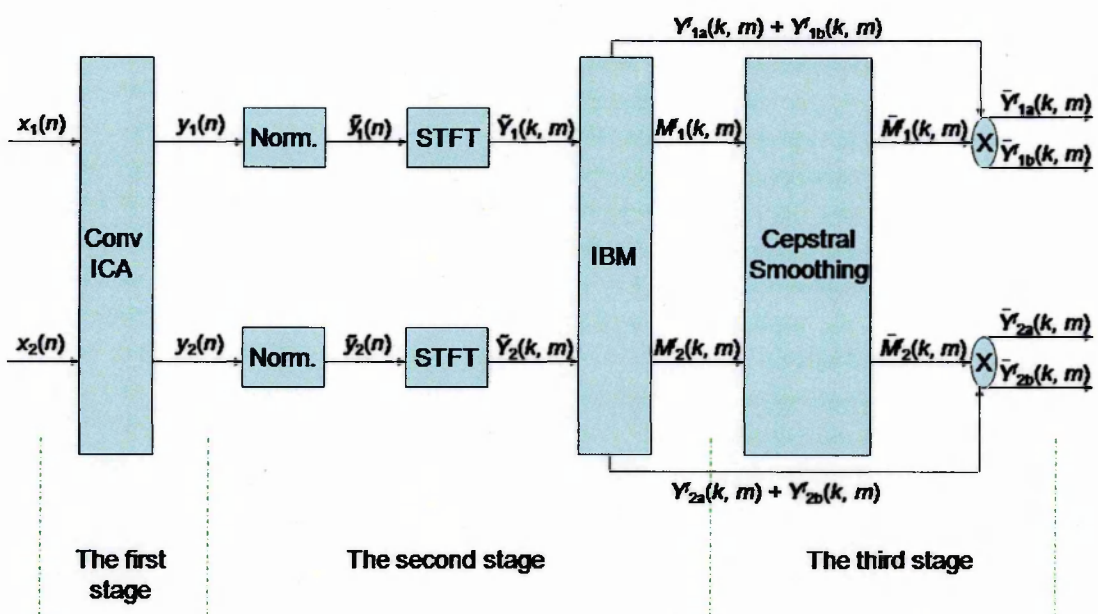


Figure 3.1: Block diagram of the proposed multistage approach. In the first stage, a convolutive ICA algorithm (denoted as "Conv ICA") is applied to the mixture signals $x_j(n)$ $(j = 1, 2)$ to obtain the coarsely separated signals $y_i(n)$ $(i = 1, 2)$. In the second stage, $y_i(n)$ is first normalised (denoted as "Norm") to obtain $\tilde{y}_i(n)$, which is then transformed to $\tilde{Y}_i(k, m)$ using the STFT followed by the estimation of the binary masks $M_i^f(k, m)$. In the third stage, cepstral smoothing is applied to the estimated masks $M_i^f(k, m)$ and the smoothed masks $\overline{M}_i^f(k, m)$ are then used to enhance the separated speech signals obtained from the second stage.

masks are denoted by $M_1^f(k,m)$ and $M_2^f(k,m)$, where k represents the frequency index, and m is the time frame index. The final stage is to perform smoothing of the estimated IBM in the cepstral domain to reduce the musical noise. The smoothed version of the estimated IBM is denoted by $\overline{M}_1^f(k,m)$ and $\overline{M}_2^f(k,m)$, as shown in Figure 1. Finally, the smoothed masks (after being converted back to the spectral domain) are applied to the outputs of the previous step, followed by an inverse T-F transform to obtain the estimated source signals in the time domain.

The remainder of the chapter is organised as follows. The convolutive ICA approach and its utilization in the first stage of the proposed method is presented in Section 3.2. Section 3.3 describes in detail the second stage of the algorithm, i.e., how to estimate the IBM from the outputs of the convolutive ICA algorithm. Musical noise reduction using cepstral smoothing, i.e., the final stage of the proposed algorithm, is explained in Section 3.4. Section 3.5 thoroughly evaluates the proposed method and compares it with two related methods [129] and [178]. Further discussions about the results and some conclusions are given in Section 3.6.

3.2 BSS of Convolutive Mixtures in the Frequency Domain

In a cocktail party environment, N speech signals are recorded by M microphones, which can be described mathematically by a linear convolutive model

$$x_j(n) = \sum_{i=1}^{N}\sum_{p=1}^{P} h_{ji}(p)s_i(n-p+1) \qquad (j=1,...,M) \tag{3.1}$$

where s_i and x_j are the source and mixture signals respectively, h_{ji} is a P-point room impulse response [4] from source s_i to microphone x_j. The BSS problem for convolutive mixtures in the time domain is converted to multiple instantaneous problems in the frequency domain by applying the short time Fourier transform (STFT) to equation (3.1), see e.g. [2,8,64,68,126,136,139,146,148,154,178,189], and using matrix notations, as follows

$$\mathbf{X}(k,m) = \mathbf{H}(k)\mathbf{S}(k,m) \tag{3.2}$$

where $\mathbf{X}(k,m) = [X_1(k,m), ..., X_M(k,m)]^T$ with its elements $X_j(k,m)$ being the T-F representations of the microphone signals $x_j(n)$, $\mathbf{S}(k,m) = [S_1(k,m), ..., S_N(k,m)]^T$ whose elements $S_i(k,m)$ are the T-F representations of the source signals $s_i(n)$, and $[\cdot]^T$ denotes vector transpose. The mixing matrix $\mathbf{H}(k)$ is assumed to be invertible and time invariant. In this study a two-input two-output system has been considered, i.e., $N = M = 2$.

To find the sources, an unmixing filter $\mathbf{W}(k)$ can be applied to the mixtures, also shown in Figure 3.2

$$\mathbf{Y}(k,m) = \mathbf{W}(k)\mathbf{X}(k,m) \tag{3.3}$$

where $\mathbf{Y}(k,m) = [Y_1(k,m), Y_2(k,m)]^T$ represents the estimated source signals in the T-F domain and $\mathbf{W}(k)$ is denoted as $[[W_{11}(k), W_{12}(k)]^T; [W_{21}(k), W_{22}(k)]^T]^T$, which can be estimated based on the assumption of independence. Many algorithms have been developed for this purpose [6, 8, 9, 32, 126, 146]. In this work a convolutive ICA approach in [178] is used for the estimation of $\mathbf{W}(k)$. Applying an inverse STFT (ISTFT), $\mathbf{Y}(k,m)$ can be converted back to the time domain denoted as

$$\mathbf{y}(n) = \text{ISTFT}(\mathbf{Y}(k,m)) \tag{3.4}$$

where $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$ denotes the estimated source signals in time domain. This inverse transform is for the purpose of applying a scaling operation to the estimated sources, as explained in the next section. Similar to many existing ICA approaches, e.g., [126], however, the separation performance of [178], especially the quality of the separated speech, is still limited due to the existence of a certain amount of interference within the separated speech. The performance further degrades with an increase of the reverberation time (RT). Such degradation is caused partly by the tradeoff between the filter length used in the convolutive model and the frame length of the STFT within the frequency-domain algorithms. For a high reverberation condition, an unmixing filter with long time delays is usually preferred for covering sufficiently the late reflections. On the other hand, the frequency domain operation usually requires the frame length of the STFT to be significantly greater than the length of the unmixing filter, in order to keep the permutation ambiguities across the frequency bands to a minimum. The filter length constraint may be relaxed when other techniques, such as beamforming

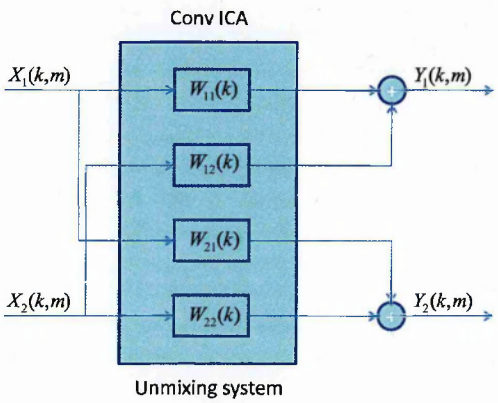


Figure 3.2: Block diagram showing the first stage of the proposed approach. The mixture signals in T-F domain, i.e., $X_j(k,m)$ are the input to a frequency-domain BSS algorithm. The unmixing filter $W_{ij}(k)$ $(i,j=1,2)$ is then estimated in the frequency domain, and $Y_i(k,m)$ is the T-F representation of the separated signals.

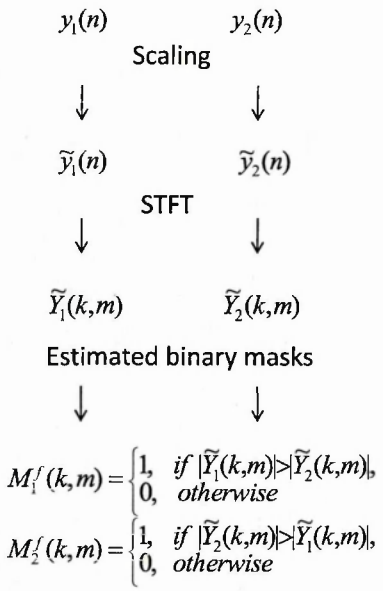


Figure 3.3: Flow chart showing the second stage of the proposed method. The separated signals from the first stage i.e., $y_i(n)$ $(i=1,2)$ are scaled to $\tilde{y}_i(n)$, which are transformed to the T-F domain $\tilde{Y}_i(k,m)$ using the STFT. The final step is to estimate the binary masks $M_i^f(k,m)$ from $\tilde{Y}_i(k,m)$.

and source envelope correlations [112, 148, 159], are used for solving the permutation problem; however the performance of such techniques deteriorates considerably for highly reverberant acoustic conditions. To improve the quality of the separated speech signals, it is considered to further apply the IBM technique, as detailed in the next section.

3.3 Combining Convolutive ICA and Binary Masking

In order to explain the connection of this stage with the previous stage, a flow chart is shown in Figure 3.3. The two outputs $y_1(n)$ and $y_2(n)$ obtained from the first stage

are used here to estimate the binary masks. Since these outputs are arbitrarily scaled, it is necessary to reduce the scaling ambiguity using normalisation, given as follows

$$\tilde{y}_i(n) = \frac{y_i(n)}{max(\mathbf{y}_i)} \qquad i = 1, 2 \tag{3.5}$$

where max denotes the maximum element of its vector argument $\mathbf{y}_i = [y_i(1), ..., y_i(L)]^T$, and L is the length of the signal. After this, the two normalized outputs are transformed into the T-F domain using the STFT as

$$\tilde{Y}_i(k, m) = \text{STFT}(\tilde{y}_i(n)) \qquad i = 1, 2 \tag{3.6}$$

Without the scaling operation, the processing by (3.4), (3.5) and (3.6) can be omited within the algorithm. By comparing the energy of each T-F unit of the above two spectrograms, the two binary masks are estimated as [169]

$$M_1^f(k, m) = \begin{cases} 1 & \text{if } \mid \tilde{Y}_1(k, m) \mid > \tau \mid \tilde{Y}_2(k, m) \mid, \\ 0 & \text{otherwise} \qquad \forall k, m. \end{cases} \tag{3.7}$$

$$M_2^f(k, m) = \begin{cases} 1 & \text{if } \mid \tilde{Y}_2(k, m) \mid > \tau \mid \tilde{Y}_1(k, m) \mid, \\ 0 & \text{otherwise} \qquad \forall k, m. \end{cases} \tag{3.8}$$

where τ is a threshold for controlling the sparseness of the mask, and $\tau = 1$ has been used in the experiment. For example if τ >1, then the two estimated masks will be having fewer unity/one values in comparison to the two estimated masks obtained above for τ =1, and hence become more sparse. The masks are then applied to the T-F representation of the original two-microphone recordings in order to recover the source signals, as follows

$$Y_i^f(k, m) = M_i^f(k, m) X_i(k, m) \qquad i = 1, 2 \tag{3.9}$$

The source signals in the time domain are recovered for the purpose of pitch estimation in the next section, using the inverse STFT (ISTFT).

$$y_i'(n) = \text{ISTFT}(Y_i^f(k, m)) \qquad i = 1, 2 \tag{3.10}$$

As observed in the experiments, the estimated IBM considerably improves the separation performance by reducing the interference to a much lower level, leading to the separated speech signals with improved quality over the outputs obtained in Section 3.2. However, a typical problem with the binary T-F masking is the introduction of errors in the estimation of the masks causing fluctuating musical noise [7, 101]. To mitigate this problem, a cepstral smoothing technique is employed [101] as detailed in the next section.

3.4 Cepstral Smoothing of the Binary Mask

The basic idea is to apply different levels of smoothing to the estimated binary mask across different frequency bands. Essentially, the levels of smoothing are determined based on the speech production mechanism. To this end, the estimated IBM is first transformed into the cepstral domain, and the different smoothing levels are then applied to the transformed mask. The smoothed mask is further converted back to the spectral domain. Through this method, the musical artifacts within the signals can be reduced, and at the same time, the broadband structure and pitch information of the speech signal are well preserved [101,122], without being noticeably affected by the smoothing operation. Representing the binary masks of (3.7) and (3.8) in the cepstrum domain given as

$$M_i^c(l,m) = DFT^{-1}\{\ln(M_i^f(k,m))\mid_{k=0,..,K-1}\} \tag{3.11}$$

where l and k are the quefrency bin index and the frequency bin index respectively [101], DFT represents the discrete Fourier transform, ln denotes the natural logarithm operator and K is the length of the DFT. To avoid the infinity error due to ln, a lower bound is applied to $M_i^f(k,m)$ in (3.11). After applying smoothing, the resultant smoothed mask is given as

$$\overline{M}_i^s(l,m) = \lambda_l \overline{M}_i^s(l,m-1) + (1-\lambda_l) M_i^c(l,m) \quad i = 1,2 \tag{3.12}$$

where λ_l is a parameter for controlling the smoothing level, and is selected according to different values of l

$$\lambda_l = \begin{cases} \lambda_{env} & \text{if } l \in \{0, ..., l_{env}\}, \\ \lambda_{pitch} & \text{if } l = l_{pitch}, \\ \lambda_{peak} & \text{if } l \in \{(l_{env}+1), ..., K\} \setminus l_{pitch} \end{cases} \tag{3.13}$$

where $0 \leq \lambda_{env} < \lambda_{pitch} < \lambda_{peak} \leq 1$, l_{env} is the quefrency bin index that represents the spectral envelope of the mask $\mathbf{M}^f(k,m)$ defined as $[M_1^f(k,m), M_2^f(k,m)]^T$, and l_{pitch} is the quefrency bin index showing the structure of the pitch harmonics in $\mathbf{M}^f(k,m)$. The principle employed for this range of λ_l is illustrated as follows. $\mathbf{M}^c(l,m)$=$[M_1^c(l,m), M_2^c(l,m)]^T$, $l \in \{0, .., l_{env}\}$, basically represents the spectral envelope of the mask $\mathbf{M}^f(k,m)$. In this region the value selected for λ_l is relatively low to avoid distortion in the envelope. Similarly, low smoothing is applied if l is equal to l_{pitch}, so that the harmonic structure of the signal is maintained. The symbol "\" is used to exclude l_{pitch} from the quefrency range $(l_{env}+1), ..., K$. High smoothing is applied in this last range in order to reduce the artifacts without harming the pitch information and structure of the spectral envelope. Different from [101], the pitch frequency is calculated in this work by using the segregated speech signal obtained in Section 3.3. Specifically, pitch frequency can be computed as

$$l_{pitch} = \text{argmax}_l\{Y^c(l,m) \mid l_{low} \leq l \leq l_{high}\}, \tag{3.14}$$

where $Y^c(l,m)$ is the cepstrum domain representation of the segregated speech signal $y'(n)$ obtained in (3.10). Note that the subscript i in symbols λ_l, l and $Y^c(l,m)$ within (3.13) and (3.14) have been omitted for notational convenience. The range l_{low}, l_{high} is chosen so that it can accommodate pitch frequencies of human speech in the range of 50 to 500 Hz. The final smoothed version of the spectral mask is given as

$$\overline{M}_i^f(k,m) = exp(DFT\{\overline{M}_i^s(l,m) \mid_{l=0,...,K-1}\}), \tag{3.15}$$

This smoothed mask is then applied to the segregated speech signals of Section 3.3, as follows

$$\overline{Y}_i^f(k,m) = \overline{M}_i^f(k,m) Y_i^f(k,m) \qquad i = 1,2 \tag{3.16}$$

Table 3.1: The proposed multistage algorithm

1) Initialize the parameters, such as M, N, overlapfactor, and read the speech mixtures into $x(n)$.
2) Convert $x(n)$ to the T-F representation $\mathbf{X}(k,m)$ using STFT, and apply the convolutive ICA algorithm in [178] to the mixture $\mathbf{X}(k,m)$ for estimating $\mathbf{W}(k)$. Obtain $\mathbf{Y}(k,m)$ according to (3.3).
3) Use (3.4), (3.5) and (3.6) to calculate $\tilde{Y}_i(k,m)$.
4) Estimate $M_i^f(k,m)$ according to (3.7) and (3.8), where $i=1,2$.
5) Compute $Y_i^f(k,m)$ based on (3.9) and $y_i^t(n)$ using (3.10). Compute the cepstrum domain representation of $y_i'(n)$, i.e., $Y^c(l,m)$.
6) Calculate $M_i^c(l,m)$ using (3.11).
7) Use (3.12) to calculate $\overline{M}_i^s(l,m)$, where λ_l is chosen according to (3.13), and $l=l_{pitch}$ is determined by (3.14).
8) Compute $\overline{M}_i^f(k,m)$ based on (3.15), and $\overline{Y}_i^f(k,m)$ according to (3.16).
9) Apply the ISTFT to $\overline{Y}_i^f(k,m)$ to obtain the separated signals in the time domain.

By further applying the ISTFT to $\overline{Y}_i^f(k,m)$, the separated source signals can then be obtained in time domain. According to the explanation in the above sections, the algorithm presented in this chapter is summarized in Table 3.1.

3.5 Results and Comparisons

In this section, the performance of the proposed method is evaluated using simulations. The algorithm is applied to both artificially mixed signals and real room recordings.

3.5.1 Experimental setup and evaluation metrics

A pool of 12 different speech signals from the TIMIT database has been used in the experiments. These speech signals were uttered by six male and six female speakers with 11 different languages [129]. All the signals have the same loudness level. The Hamming window is used with an overlap factor set to 0.75. The duration of the speech signal is 5 seconds with a sampling rate of 10 KHz. The rest of the parameters are set as: l_{env}=8, l_{low}=16, l_{high}=120, λ_{env}=0, λ_{pitch}=0.4, and λ_{peak}=0.8. Performance indices used in evaluation include signal to noise ratio (SNR), the percentage of energy loss (PEL) and the percentage of noise residue (PNR) [70, 129]. The expressions of PEL and PNR are given below

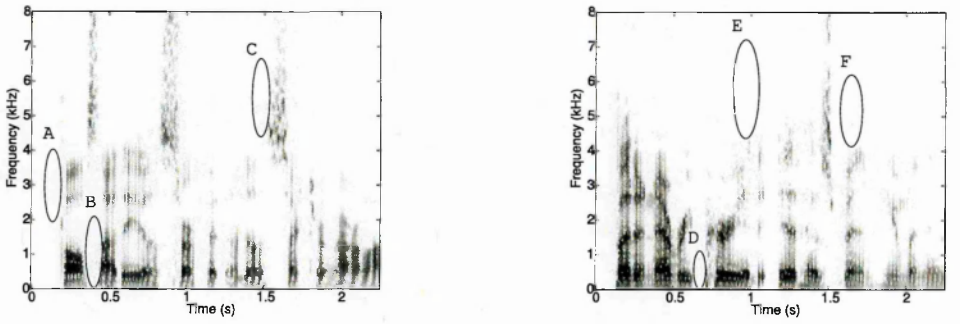
$$PEL = \frac{\sum_n (e_1(n))^2}{\sum_n (I(n))^2} \tag{3.17}$$

$$PNR = \frac{\sum_n (e_2(n))^2}{\sum_n (\overline{y}(n))^2} \tag{3.18}$$

where $\overline{y}(n)$ and $I(n)$ represent the estimated signal and the signal resynthesized after applying the ideal binary mask [129]. $e_1(n)$ stands for the signal present in $I(n)$ but absent in $\overline{y}(n)$ while $e_2(n)$ shows the signal present in $\overline{y}(n)$ but absent in $I(n)$. SNR_i is the ratio of the desired signal to the interfering signal taken from the mixture, where i refers to the input. SNR_o is the ratio of the desired signal resynthesized from the ideal binary mask to the difference of the desired resynthesized signal and the estimated signal, where o refers to the output [129]. Notations mSNR_i, mSNR_o and ΔSNR are also used in the evaluation where mSNR_i and mSNR_o are the average results for fifty random tests and $\Delta\text{SNR}=\text{mSNR}_o-\text{mSNR}_i$. All the SNR measurements are given in decibels (dB) in the subsequent experiments.

3.5.2 A separation example

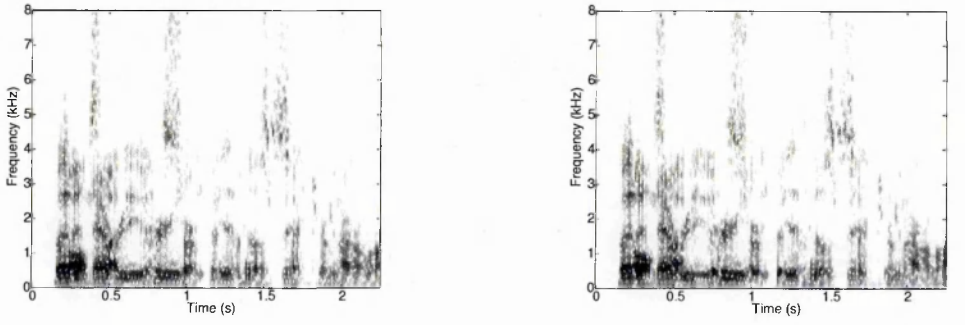
To show the performance of the proposed method for interference suppression, an example is given here when applying the algorithm to the separation of two speech mixtures obtained by mixing two sources from the pool described in the above section using the simulated room model [4], with RT set to 100 msec. The spectrograms of the two source signals are shown in Figure 3.4(a) and (b), and the two mixture signals in Figure 3.5(a) and (b). For the computation of the spectrograms, the FFT frame length was set to 2048 (i.e., 204.8 msec), and the window length (or frame shift) was fixed to 512 giving, 75% overlap between neighboring windows. Other parameters were the same as those specified in the above section. Figure 3.6(a) and (b) show the spectrograms of the output signals obtained from the first stage of the proposed algorithm. The results obtained from the second stage of the proposed algorithm are shown in Figure 3.7(a) and (b), and from the third stage in Figure 3.8(a) and (b). For the convenience of comparison, some T-F regions within the spectrograms are highlighted to show the performance improvement for interference suppression at each stage. In particular, three regions are shown in one of the two source signals, which are marked as A, B and C for the original one (i.e. the source signal before the mixing operation) and as A_i, B_i and C_i for the separated one (i.e. the source signals estimated from the mixtures),



(a) (b)

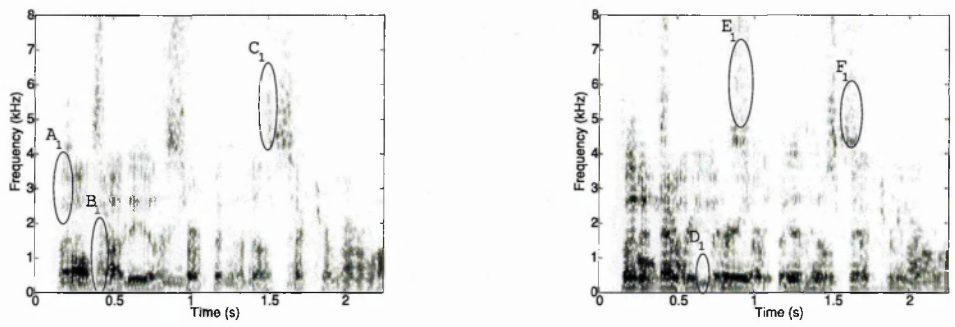
Figure 3.4: Spectrograms of the two original speech signals used in the separation example. Three areas in each are highlighted for purposes of comparison with Figures 3.5-3.8 .

where $i = 1, 2, 3$ is the stage index. Similarly three regions in the other source are marked as D, E and F for the original one and as D_i, E_i and F_i for the separated one after each stage of the algorithm. From the highlighted regions, it can be observed that the interference within one source that comes from the other is reduced gradually after the processing of each stage. Compared with the output of the first stage, the interference within the estimated sources from the output of the third stage has been reduced significantly.



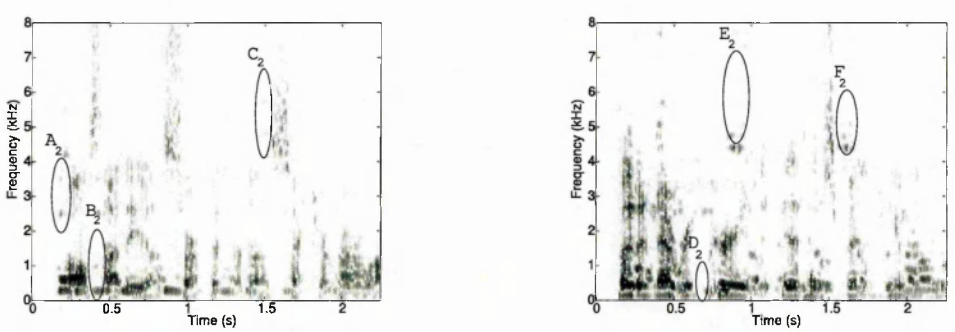
(a) (b)

Figure 3.5: Spectrograms of the mixture signals that were generated by using the simulated room model with RT set to 100 msec. Both signals in (a) and (b) are the mixtures of two speech sources but with different attenuation and time delays.



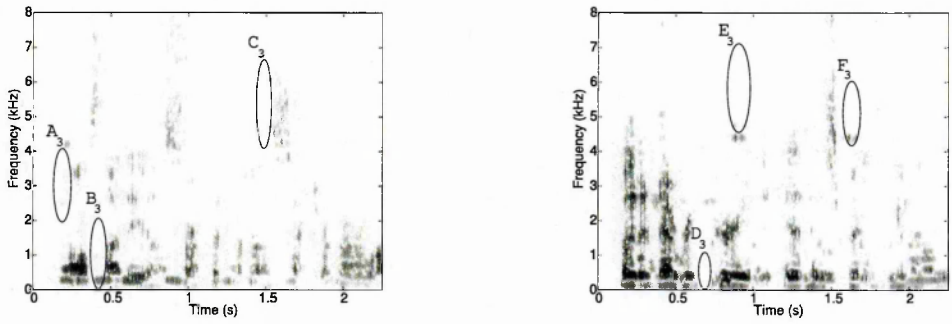
(a) (b)

Figure 3.6: Spectrograms of the separated speech sources obtained from the output of the first stage of the proposed algorithm, i.e., by applying the convolutive ICA algorithm. It can be observed that a considerable amount of interference from the other source still exists in the highlighted regions.



(a) (b)

Figure 3.7: Spectrograms of the separated speech sources obtained from the output of the second stage of the proposed algorithm, i.e., by applying the estimated IBM. The interferences in the highlighted regions have been considerably reduced as compared with those in Figure 3.6.



(a) (b)

Figure 3.8: Spectrograms of the separated speech sources obtained from the output of the third stage of the proposed algorithm, i.e., by applying cepstral smoothing to the estimated IBM. The interferences in the highlighted regions have been further reduced as compared with those in Figures 3.6 and 3.7.

3.5.3 Objective evaluation

First, the performance of the proposed algorithm has been evaluated for the separation of convolutive mixtures that were generated artificially by using the simulated room model [4], for which the RT can be specified explicitly and flexibly. The robustness of the proposed algorithm has been assessed to the changes of the key parameters used in the algorithm, such as the window length and the FFT frame length, as well as to evaluate the performance variations against different conditions for generating the mixtures, such as the reverberation time and the noise level. In each of the subsequent experiments, change is made only to one parameter, i.e., the one that has to be tested, but keep all the other parameters fixed (as those already specified in Section 3.5.1). For each of these evaluations, the results obtained were the averaged performance of the results for 50 different convolutive mixtures, with each consisting of two speech sources randomly picked up from a pool of 12 speech signals [129]. In the experiments, it has been observed that ΔSNR measured from the output of the third stage is slightly lower (hence negligible) than that measured from the output of the second stage of the proposed algorithm, although subjective listening tests suggest that the quality of the separated speech has been improved (as shown in Section 3.5.4). For this reason, the

results of $mSNR_o$ shown in this section are measured from the output of the second stage (as shown in our preliminary work [76]). However, more comprehensive results for $mSNR_o$ measured at each stage of the proposed algorithm are given in Section 3.5.5. Analysis of variance (ANOVA) based statistical significance evaluation ([69], chapter 11) of the performance difference between the second and third stage of the algorithm is also given in Section 3.5.5.

In the first experiment, the window length was varied from 256 to 2048 samples, while the other parameters were set identical to those in Section 3.5.1 and 3.5.2. The results are given in Table 3.2. It can be seen that the highest ΔSNR is obtained for the window length of 512. Therefore, the window length equal to 512 samples was used in the following experiments.

In the second experiment, the FFT frame length was changed from 512 to 2048. The average results for different FFT frame lengths are given in Table 3.3. It can be seen that by increasing the FFT frame length from 512 to 2048 samples, the performance of the proposed algorithm in terms of SNR, PEL and PNR is all improved. The best performance is obtained at 2048. Hence, the FFT frame length used for the subsequent experiments was fixed to 2048 samples.

In the third experiment, the reverberation time of the simulated room has been changed when generating the mixtures. The average results in terms of PEL, PNR and ΔSNR for the various RTs are summarized in Table 3.4, where the unit for RT is msec. A noticeable trend in this table is that the performance degrades gradually with an increase of RT, which is not unexpected due to the increasing sound reflections for higher room reverberations.

In the fourth experiment, different levels of microphone noise is considered by adding white noise to the mixtures, where the noise level was calculated with respect to the level of the mixtures, with a weaker noise corresponding to a smaller number [129]. The average ΔSNR values for different noise levels are given in Table 3.5. It can be observed that the performance of the algorithm decreases as the noise level is increased, and similar to [129], the algorithm can tolerate the noise levels up to -20 dB.

Lastly, the performance of the proposed algorithm is evaluated (without considering

Table 3.2: Separation results for different window lengths

Window Length	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
256	9.10	15.30	1.10	7.11	6.01
512	8.60	14.48	1.10	7.44	6.34
1024	9.30	14.70	1.10	7.11	6.01
2048	10.92	15.92	1.12	6.32	5.20

Table 3.3: Separation results for different FFT frame lengths

NFFT	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
512	9.06	14.96	1.10	7.17	6.06
1024	8.65	14.53	1.10	7.40	6.30
2048	8.60	14.48	1.10	7.44	6.34

noise) by varying the values of λ_{env}, λ_{pitch} and λ_{peak} with the other parameters fixed as: RT =100 msec, window length=512, and NFFT=2048. The values of λ_{env} , λ_{pitch} and λ_{peak} as discussed in section 3.4, were chosen in the range $[0, 0.9]$. The results measured by mSNR_o are given in Figures 3.9, 3.10 and 3.11 respectively. From Figure 3.9, it is observed that mSNR_o after the third stage increases slowly for λ_{env} ranging from 0 to 0.4 and then starts decreasing. Figure 3.10 shows a very slight increase in mSNR_o when λ_{pitch} is between 0 and 0.5 followed by a very slight decrease. In Figure 3.11, mSNR_o first increases slowly when λ_{peak} varies from 0 to 0.4 and then a sharp decrease is observed when λ_{peak} is between 0.5 and 0.9. These experiments show that the separation performance varies to some extent when different values for λ_{env}, λ_{pitch} and λ_{peak} are used.

Table 3.4: Separation results for different RT

RT	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
40	2.16	2.24	1.13	13.22	12.08
60	3.79	4.12	1.15	10.94	9.79
80	5.50	8.30	1.14	9.42	8.27
100	8.60	14.48	1.10	7.44	6.34
120	10.99	19.53	1.03	6.30	5.26
140	13.36	24.14	0.94	5.48	4.53
150	13.86	25.38	0.90	5.29	4.39

Table 3.5: Separation results for different noise levels

Noise	PEL	PNR	mSNR_i	mSNR_o	ΔSNR
-40 dB	8.60	14.48	1.10	7.45	6.34
-30 dB	8.60	14.48	1.10	7.44	6.34
-20 dB	8.62	14.52	1.10	7.43	6.33
-10 dB	9.46	16.49	1.09	6.91	5.81

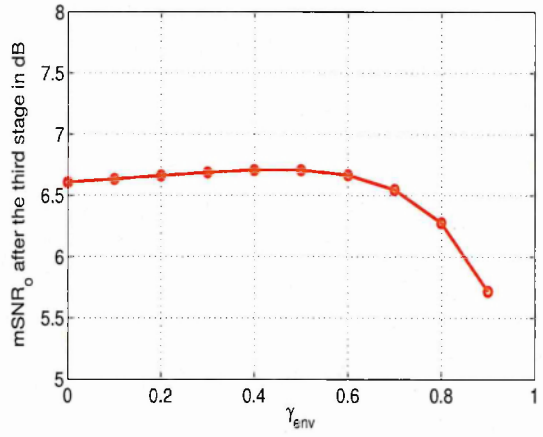


Figure 3.9: Separation performance measured by mSNR_o with different values of λ_{env}.

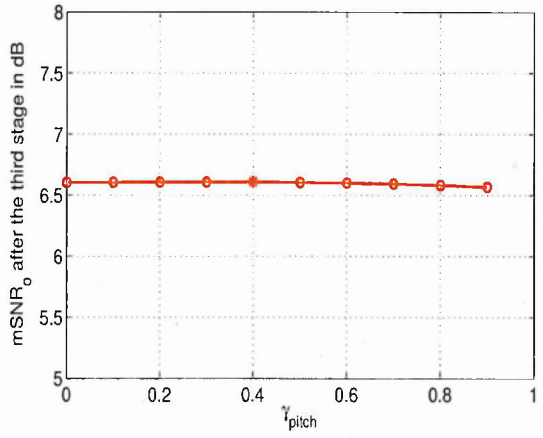


Figure 3.10: Separation performance measured by mSNR_o with different values of λ_{pitch}.

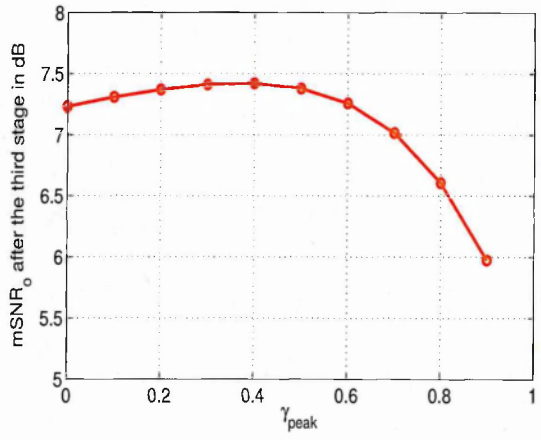


Figure 3.11: Separation performance measured by mSNR_o with different values of λ_{peak}.

3.5.4 Listening tests

As mentioned in the above section that ΔSNR measured from the output of the third stage of the proposed algorithm appears to be slightly lower than that measured from the output of the second stage of the proposed algorithm (see more results and detailed analysis in the next section). This suggests that cepstral smoothing actually does not improve the objective performance in terms of SNR measurement (see also [169]). Nevertheless, the informal listening tests seem to contradict the SNR measurements and confirm that the cepstral smoothing does improve the quality of the separated speech, especially for the musical noise removal. To show this, subjective listening tests have been conducted by recruiting 15 participants with normal hearing. Each of these listeners was asked to give an integer score ranging from 1 (musical noise clearly audible) to 5 (noise not audible) for the final segregated speech signals, as suggested in [7]. During these tests, each participant was asked to listen to 2 groups of separated speech signals obtained in the experiments where RT was set to 50, 100, 150 and 200 msec respectively, with one group containing y_1 and the other group containing y_2. A total of 8 groups of speech signals were evaluated subjectively by these participants. Each group was composed of 3 speech signals, i.e. the estimated source obtained from the output of the second stage, the one from the third stage, and the source signal estimated by Pedersen *et al.*'s method. Note that the listeners had no prior knowledge on which signal was obtained from which algorithm. This ensures a fair

Table 3.6: MOS obtained from subjective listening tests

RT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
50	3.26	3.90	3.01	5.0948	4.1960	0.0320
100	2.12	2.62	2.29	4.7094	4.1960	0.0386
150	1.87	2.39	2.02	5.0995	4.1960	0.0319
200	1.09	2.07	1.82	50.2059	4.1960	0.0000

comparison between the algorithms. The mixtures used in these tests were generated by the simulated room model with RT equal to 50, 100, 150 and 200 msec, respectively. The scores given by the listener are provided on the basis of how clean the separated signals from the two stages are in comparison to each other, or how much musical noise is present in the separated signals. A signal with less musical noise is cleaner, and hence is given a higher mean opinion score (MOS) [7]. The average results of MOS for the 15 listeners are given in Table 3.6. It indicates that using cepstral smoothing gives higher MOS, suggesting the improved quality of the separated speech. To examine whether the improvement in MOS after smoothing is statistically significant, one-way ANOVA based F-test [69] has been performed for the MOS obtained before and after smoothing. The results are given in Table 3.6. The critical value (F_{crit}) is the number that the test statistic must overcome to reject the test. The p-value stands for the probability of a more extreme (positive or negative) result than what is actually achieved, given that the null hypothesis is true. F-value can be defined as the ratio of the variance of the group means to the mean of the within group variances. All the F-tests in this work have been carried out at 5% significance level. If $F < F_{crit}$ and p-value is greater than 0.05 (5% significance level), then the given results are statistically insignificant. It can be observed that the p-values obtained for all the cases of RT in Table 3.6 are smaller than 0.05, suggesting that the improvement in all the four cases is statistically significant.

Additional listening tests have been carried out using the speech signals randomly selected from the experimental results employed for the objective evaluation of the proposed method. 20 volunteers have been recruited to participate the subjective listening tests, including the 15 listeners mentioned earlier. The results have been evaluated

Table 3.7: MOS obtained from subjective listening tests for different window lengths

For RT=100 msec						
Window Length	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
256	2.35	3.70	2.57	64.4233	4.0980	0.00000
512	2.70	3.65	2.90	16.5277	4.0980	0.00023
1024	2.60	3.65	2.81	24.1470	4.0980	0.00001
2048	2.40	3.10	2.64	7.0000	4.0980	0.0118
For RT=200 msec						
Window Length	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
256	1.70	2.80	1.94	16.7810	4.0980	0.00021
512	1.75	2.70	2.04	21.5016	4.0980	0.00004
1024	1.75	2.65	2.01	15.1626	4.0980	0.00038
2048	1.55	2.35	1.78	15.6903	4.0980	0.00031

for different window lengths in Table 3.7, for different FFT frame lengths in Table 3.8 and for different noise levels in Table 3.9. The RT has been set to 100 and 200 msec, respectively. The criteria used in Table 3.6 for the MOS have also been employed here. The results given in Table 3.7 show that for different window lengths at $RT = 100$ and 200 msec, cepstral smoothing offers higher MOS scores, indicating that the quality of the segregated speech signal has been improved. A similar trend can be observed in Table 3.8 and 3.9 where using cepstral smoothing achieves a higher MOS. In all cases the differences of MOS before and after smoothing are statistically significant.

3.5.5 Comparison to other methods

In this section, the proposed multistage method has been compared with two related approaches in [129] and [178]. In [178] speech signals were separated from convolutive mixtures by exploiting the second order non-stationarity of the sources in the frequency domain, where the cross-power spectrum based cost function and a penalty function have been employed to convert the separation problem into a joint diagonalization problem with unconstrained optimization. Pedersen *et al.*'s method [129] combines an instantaneous ICA algorithm with the binary T-F masking for underdetermined blind

Table 3.8: MOS obtained from subjective listening tests for different FFT frame lengths

For RT=100 msec						
NFFT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
512	3.30	4.10	2.88	17.3714	4.0980	0.00017
1024	3.20	4.15	2.87	17.3646	4.0980	0.00017
2048	2.70	3.65	2.90	16.5277	4.0980	0.00023
For RT=200 msec						
NFFT	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
512	2.05	2.80	1.89	8.8509	4.0980	0.00510
1024	1.75	2.50	1.96	10.3012	4.0980	0.00270
2048	1.75	2.70	2.04	21.5016	4.0980	0.00004

Table 3.9: MOS obtained from subjective listening tests for different noise levels

For RT=100 msec						
Noise	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
-40 dB	3.30	4.20	2.84	15.8660	4.0980	0.00029
-30 dB	3.20	4.15	2.70	19.3211	4.0980	0.00008
-20 dB	2.70	3.70	2.09	14.3939	4.0980	0.00051
-10 dB	1.80	2.55	1.84	10.6079	4.0980	0.00240
For RT=200 msec						
Noise	MOS before smoothing	MOS after smoothing	MOS for Pedersen *et al.*	ANOVA based statistical significance evaluation of MOS before & after smoothing		
				F-value	F_{crit}	p-value
-40 dB	2.00	2.80	2.01	16.0000	4.0980	0.00028
-30 dB	2.15	2.85	1.93	12.3311	4.0980	0.00120
-20 dB	1.70	2.50	1.76	18.4242	4.0980	0.00011
-10 dB	1.30	1.90	1.49	9.7714	4.0980	0.0034

Table 3.10: Comparison results for different window lengths

Window Length	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
256	1.10	2.98	7.11	6.81	0.9085	3.9380	0.3429
512	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068
1024	1.10	3.01	7.11	6.09	11.4642	3.9380	0.0010
2048	1.12	2.95	6.32	5.32	12.8289	3.9380	0.0005

source separation, where the outputs of the ICA algorithm were used to estimate the binary mask in an iterative way to extract multiple speech sources from two mixtures.

Comparison between the proposed method and the method in [178] is essentially equivalent to the comparison between the outputs from the third (and/or second stage) and those from the first stage, as the method in [178] is employed in the first stage of the proposed approach. Therefore, without performing additional experiments, more results are shown that were obtained from the experiments already conducted in Section 3.5.3. In parallel with the results shown in Tables 3.2, 3.3, 3.4, and 3.5, the comparison results in terms of $mSNR_o$ is shown in Tables 3.10 for different window lengths, 3.11 for different FFT frame lengths, 3.12 for different RT values and 3.13 for different noise levels. All the results were measured based on 50 random tests. Note that $mSNR_o$ obtained after the first stage of the proposed method is approximately calculated. This is because, according to the definition of SNR_o in Section 3.5.1, the masked output signals should be used for the calculation of output SNR, while the obtained signal from the output of the first stage [178] is not a masked signal. The results in Table 3.10 clearly indicate that the output SNR has been improved at the second and third stage in comparison to the first stage for different window lengths. The objective results from the third stage in terms of $mSNR_o$ measurement are slightly worse than those of the second stage, due to the smoothing operation. According to the subjective listening tests in the previous section, the quality of the speech source from the third stage is actually improved, due to the reduced level of audible musical noise.

Table 3.11 compares the results of the proposed method and the method in [178] for different FFT frame lengths, where the window length was fixed to 512, the overlap

Table 3.11: Comparison results for different FFT frame lengths

NFFT	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
512	1.10	3.01	7.17	6.46	5.8298	3.9380	0.0176
1024	1.10	3.02	7.40	6.57	7.4946	3.9380	0.0074
2048	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068

Table 3.12: Comparison results for different *RT*

RT	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
40	1.13	3.70	13.22	9.44	100.2190	3.9380	0.0000
60	1.15	3.47	10.94	8.48	40.4630	3.9380	0.0000
80	1.14	3.36	9.42	7.75	23.1972	3.9380	0.0000
100	1.10	3.02	7.44	6.59	7.6412	3.9380	0.0068
120	1.03	2.70	6.30	5.82	3.7015	3.9380	0.0573
140	0.94	2.47	5.48	5.23	0.9266	3.9380	0.3381
150	0.90	2.42	5.29	5.11	0.5210	3.9380	0.4721

Table 3.13: Comparison results for different noise levels

Noise	$mSNR_i$	$mSNR_o$ after the 1st stage	$mSNR_o$ after the 2nd stage	$mSNR_o$ after the 3rd stage	ANOVA test for the difference between the SNR_os from the 2nd and 3rd stage		
					F-value	F_{crit}	p-value
-40 dB	1.10	3.02	7.45	6.60	7.6297	3.9380	0.0069
-30 dB	1.10	3.02	7.44	6.60	7.6186	3.9380	0.0069
-20 dB	1.10	3.02	7.43	6.59	7.5950	3.9380	0.0070
-10 dB	1.09	3.06	6.91	6.09	8.2232	3.9380	0.0051

factor and RT remained the same as those used for Table 3.10. From this table, we can also observe the improved performance of the proposed method in terms of SNR measurements, as compared with the method in [178]. Subjective listening tests also show that the results have considerably improved quality over those in [178] for different FFT frame lengths, which are consistent with the SNR measurements. In Table 3.12, comparison has been made for different values of RT, where the window length and the overlap factor were identical to those used in Table 3.11, and the FFT frame length was the same as that in 3.10. The results show that the output SNR decreases with an increase in RT, and the proposed method has better performance in terms of the averaged output SNR. Specifically, when RT equals to 100 msec, mSNR_o of the third stage is approximately 4 dB higher than that of the first stage. The improvement is more prominant when RT is relatively low. In Table 3.13 experiments are performed by considering the microphone noise in the mixture, as discussed already in Table 3.5. In this table, RT was set to 100 msec, and other parameters were the same as those in Table 3.12. It can be observed that the proposed method performs better than the method in [178] for the separation of noisy mixtures. Specifically, comparing mSNR_o between the first and third stages, it has been observed that there is about 3 dB improvement for noise level at -10 dB, and 3.6 dB for noise level at -30 dB. The results discussed above show that the proposed method outperforms the method in [178] in terms of SNR measurements.

To determine whether the relatively small differences of mSNR_o between the second and third stage of the proposed method are statistically significant, one-way ANOVA based F-test [69] is performed as described in Section 3.5.4. The testing results are given in Tables 3.10, 3.11, 3.12 and 3.13. To explain how the F-test was applied to the results, consider the case of NFFT equal to 512 (in Table 3.11) as an example, where mSNR_o after the second and third stage is 7.17 dB and 6.46 dB respectively. Both mSNR_os were calculated by averaging 50 individual SNR_os obtained from the 50 random tests. Each group of 50 SNR_os forms a vector, and hence two vectors can be formed from the second and third stage. The F-value was then computed from these two vectors, which is 5.8298. The F-values in other cases and tables were computed in the same way. From the results in these tables, it can be observed that in many testing cases the

differences of mSNR_o between the second and third stage of the proposed algorithm, although small, are statistically significant whereas in some cases the differences are insignificant.

The performance of the proposed method is also compared with the algorithm in [129] in terms of both computational complexity and separation quality. The separation quality is measured objectively using SNR measurement as in the above experiments, and subjectively by listening tests. To conduct this comparison, the real room recordings were used which were obtained in [129]. The real recordings were made in a reverberant room with $RT = 400$ msec. Two omnidirectional microphones vertically placed and closely spaced are used for the recordings. Different loudspeaker positions are used to measure the room impulse responses. Details about the recordings can be found in [129] and are not given here. Clean speech signals from the pool of 12 speakers were convolved with the room impulses to generate the source signals [129]. The specifications of the computing facilities that were used to perform the experiments include Intel(R) Xeon(TM) 3.00GHz CPU and 31.48 GB memory. The results are given in Table 3.14. The results show that the proposed algorithm is 18 times faster than the Pedersen et al. method. Their method requires 700 minutes for 50 random tests and 14 minutes per test. In contrast the proposed method is much faster and requires 40 minutes for 50 tests and 0.8 minutes per test. The time computational complexity of both methods was also approximately calculated. The order of complexity of the proposed method is $O(I_3(MFKlogK + M)) + O(I_3KMN(2N + M)) + O(MNI_3K) + O(FKlogK) + O(NKF) + O(L)$, where F is the number of frames[1], L is the length of the signal, and I_3 denotes the required number of iterations for the convolutive ICA algorithm [178] to converge. Similarly, the complexity of the Pedersen et al. method is $O(FKlogKI_2) + O(NKFI_2) + O(NMI_1I_2)$, where I_1 is the iteration number for the INFORMAX algorithm (used as a first stage in their method) to converge, while I_2 denotes the total number of iterations for the Pedersen et al. method to segregate the speech mixtures. Although the results for ΔSNR are comparable, listening tests given in Table 3.6 suggest that our results have a better quality than those in [129]. Some demos are available on the website [175] for both real and artificial recordings.

[1]If there is no overlap between adjacent frames then $F \cdot K \approx L$.

Table 3.14: Comparison of separation performance and computational cost between the proposed method and Pedersen Et AL.'s method

Algorithm	PEL	PNR	ΔSNR	Total time	Time per test	Run time memory requirement[2]
Proposed	30.56	9.73	2.50	40min	0.8min	223.28 MB
Pedersen *et al.*	17.14	49.33	2.64	700min	14min	255.17 MB

[2]Note that the results also include the memory required for the matlab software

3.6 Summary

The proposed approach consists of three major steps. A convolutive ICA algorithm [178] is first applied in order to take into account the reverberant mixing environments based on a convolutive unmixing model. Binary T-F masking is used in the second step for improving the SNR of the separated speech signal, due to its effectiveness in rejecting the energy of interference by assigning zeros to the T-F units in the masking matrix in which the energy of the interference is stronger than the target speech. The artifacts (musical noise) due to the error in the estimation of the binary mask in the segregated speech signals are further reduced by applying the cepstral smoothing technique. Compared with smoothing directly in the spectral domain, cesptral smoothing has the advantage of preserving the harmonic structure of the separated speech signal while reducing the musical noise to a lower level by smoothing out the unwanted isolated random peaks.

In comparison to [178], considerable improvement achieved by the proposed method in terms of both objective measurements using SNR and subjective listening tests is mainly due to the introduction of the binary T-F masking operation and the cepstral smoothing. The binary masking contributed mostly to the improvement of interference cancellation, and cepstral smoothing further improves the perceptual quality of the separated speech. For a reverberation time of 100 msec, the proposed algorithm achieves approximately 4 dB SNR gain over the typical convolutive ICA algorithm in [178]. Compared with [178], the computational complexity of the proposed algorithm is higher due to the additional processing of IBM and cepstral smoothing. It is however still computationally efficient as FFT and its inverse are used for the transforms in all the

steps.

Note the difference between the proposed method and Pedersen *et al.*'s method [129] despite a similar combination of an ICA algorithm with the IBM technique. First, the proposed algorithm directly addresses the convolutive BSS model based on the frequency-domain approach, while Pedersen *et al.*'s method is based on an instantaneous model and an instantaneous ICA algorithm, even though their algorithm has also been tested for convolutive mixtures. Second, the algorithm in [129] is iterative, which is computationally demanding. Moreover, cepstral smoothing has been introduced in the proposed method, which has the advantage of reducing the musical artifacts caused by the IBM technique.

As observed in the results, reverberation and noise degrade the performance of the separation for the convolutive speech mixtures. One could analyse reverberation and noise effects and reduce such effects present in the microphone signals before applying the ICA and IBM approaches. This issue will be addressed in the subsequent chapters.

Chapter 4

Empirical Mode Decomposition for Joint Denoising and Dereverberation

In Chapter 3, an algorithm for blind separation of convolutive speech mixtures is proposed. However, the room reverberation effects on the convolutive speech mixtures deteriorate the separation performance of the algorithm developed in Chapter 3. Also the microphone noise could affect the separation performance. Therefore, in this chapter an algorithm is developed to deal with the room reverberation and noise together. The proposed method is for the enhancement of noisy reverberant speech using empirical mode decomposition (EMD) based subband processing without any *prior* information. The proposed algorithm is a one-microphone multistage algorithm. In the first step, noisy reverberant speech is decomposed adaptively into oscillatory components called intrinsic mode functions (IMFs) via an EMD algorithm. Denoising is then applied to selected high frequency IMFs using an EMD-based minimum-mean squared error (MMSE) filter, followed by spectral subtraction of the resulting denoised high-frequency IMFs and low-frequency IMFs. Finally, the enhanced speech signal is reconstructed from the processed IMFs. The method was motivated by the observation that the noise and reverberations are disproportionally distributed across the IMF components. Therefore, different levels of suppression can be applied to the additive noise

and reverberation in each IMF. This leads to an improved enhancement performance as shown later in this chapter in comparison to a related recent approach, based on the measurements by the signal-to-noise ratio (SNR).

4.1 Introduction

As already discussed in Chapter 2 that room reverberation is one of the main causes of performance degradation in automatic speech recognition (ASR) systems. It has also been discussed in detail in Chapter 2 that room reverberation is commonly modeled as the combination of three parts, the direct signal, early reflections and the late reflections. Late reflections degrade the quality and intelligibility of speech and can cause serious problems to ASR performance. Therefore, it is very important to deal with the late reverberations so that ASR performance can be enhanced.

The late reverberations are usually treated as diffusive noise whose variance is estimated and then subtracted from the reverberant speech, for which the spectral subtraction (SS) technique has been widely used [179]. To estimate the late reverberations, a method based on an exponential decay function has been developed in [84]. The main challenge in suppression of late reverberations is to estimate accurately its variance. The presence of noise from the acoustical environments make it more difficult to estimate the power of late reverberations. Therefore, in this chapter, it is considered to enhance the noisy reverberant speech by jointly dealing with the late reverberations and the additive acoustic noise having a Gaussian distribution and white spectrum. Note that early reflections are not considered here and the method developed deals with the late reverberations which can be treated as diffusive noise unlike early reverberations.

A new method is developed here using EMD based subband analysis. An EMD algorithm is used to decompose the noisy reverberant speech into a linear combination of the so-called IMFs, ranging from the high-frequency to low-frequency bands [71], [140], [180], [181], [182]. Then the IMFs that have higher levels of noise are selected and the EMD based MMSE filter [83] is applied to reduce the additive noise. In the next step, the denoised IMF components and the remaining IMF components are used to estimate the power of late reverberations. It has been observed that the energy of the

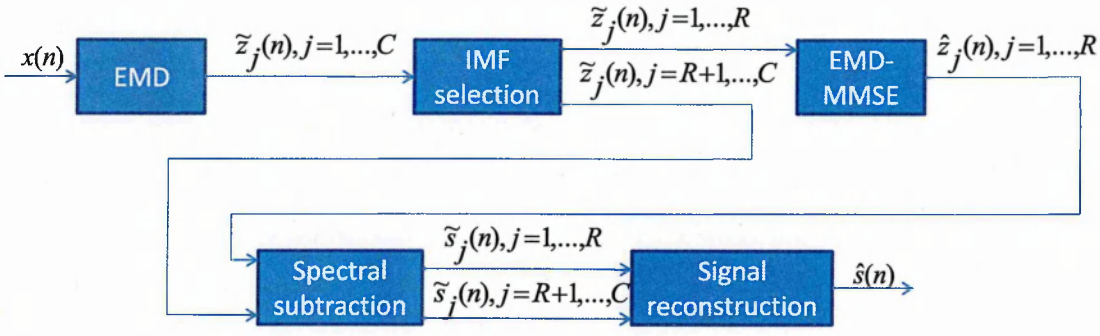


Figure 4.1: Block diagram of the proposed denoising and dereverberation system.

late reverberations is spread over the different IMFs with different magnitude. For this reason, spectral subtraction is applied to each IMF according to the energy of the late reverberations present in the IMF components. The proposed method is evaluated on the simulated and real noisy reverberant speech data, and an improved performance has been observed on the basis of SNR measurements. The next section presents the proposed approach in detail. Section 4.3 shows the evaluation results, followed by a conclusion in Section 4.4.

4.2 System Description

The proposed joint dereverberation and denoising system is depicted in Figure 4.1. First, the EMD algorithm [71] is applied to the noisy reverberant speech $x(n)$ to decompose the signal adaptively into C IMF components $\tilde{z}_j(n)$, $j = 1, ..., C$. In the next step R components are selected from the C IMF components of $\tilde{z}_j(n)$ for denoising. Then, an EMD based MMSE filter [83] is applied to each of the selected IMFs to reduce its noise level. Spectral subtraction with variable scaling factors is applied to the denoised IMFs and the remaining IMFs separately. Finally, the signal is reconstructed as $\hat{s}(n)$.

4.2.1 EMD analysis and its Review

The concept of EMD was introduced by Huang *et al.* in 1998 [71]. The EMD algorithm describes the signal details at certain frequency bands in the form of different IMFs [55].

Each IMF has a distinct time scale and acts as a basis function [71], [140], [182]. There are two main conditions that need to be satisfied by each IMF [71]. First, the difference between the number of extrema and the number of zero crossings should not exceed one. Second, the average value for the envelope assigned to the local maxima and minima is zero.

EMD is a powerful technique for data analysis. In practice, data obtained is an amalgamation of signal and noise such as signals acquired by microphones. Once the noise contaminates the data, it is not a trivial task to remove it. When the acquisition processes are linear and the noise has a distinct time or frequency scale from those of the signal, the spectral filtering method based on Fourier analysis can be employed to separate the noise from the signal. However, the filtering methods will not work properly when the processes are nonlinear. Even if the signal has distinct fundamental frequency from that of the noise, the harmonics of the signal can still mix with the noise. Such type of mixing of harmonics with noise will render the method based on Fourier filtering ineffective as compared to a noise separating method. In such a scenario, the EMD method can offer some benefits [71, 180]. EMD is an adaptive method to decompose data into its IMFs, which act as the basis components for the representation of the given data. While the basis is adaptively obtained, it usually offers a physically meaningful representation of the underlying processes. Also because of the adaptive nature of the basis, there is no need of harmonics and therefore EMD is suitable for analysing data from nonlinear and nonstationary processes.

Neverthless, in [55] and [180] it has been found empirically that EMD works as a dyadic filter bank for the white Gaussian noise and is capable of separating the white noise into IMF components having mean periods, with each having exact twice the value of the previous one. It has also been found that all the IMF components are normally distributed [180]. Hence these findings are the motivation for using EMD to enhance the reverberant speech signal contaminated by white Gaussian noise in this chapter.

EMD is implemented through a *sifting* process that is summarized as follows [55], [71], [140], [182]:

(1) For the given noisy reverberant signal (data), $x(n)$, identify all the local extrema.

(2) Connect all the maxima and minima separately by applying natural cubic spline interpolation to form the upper envelope $u(n)$ and lower envelope $l(n)$.

(3) Calculate the mean of the envelopes as $m(n) = [u(n) + l(n)]/2$.

(4) Find the early-IMF by taking the difference between the data and the mean as $h(n) = x(n) - m(n)$.

(5) Check the early-IMF whether it fulfils the two conditions as mentioned in the begining of this section, to be a candidate IMF.

(6) If the early-IMF does not satisfy the conditions, repeat steps 1-5 on $h(n)$ as many times as required until it satisfies the conditions.

(7) If the early-IMF does meet the conditions, assign the early-IMF as an IMF component, $\tilde{z}(n)$.

(8) Repeat steps 1-7 on the residue signal $r(n) = x(n) - \tilde{z}(n)$, i.e., replacing $x(n)$ in step 1 by $r(n)$.

(9) The iteration terminates when the residue, $r_C(n)$, becomes a monotonic function from which no more IMF can be extracted.

Now the mathematical details are given below to further clarify how the EMD algorithm works. The following equations show the sift process that finds the first IMF component $\tilde{z}_1(n)$, assuming steps 1-5 are repeated l times before this component is found.

$$\begin{aligned} &x(n) - m_{1,1}(n) = h_{1,1}(n); \\ &h_{1,1}(n) - m_{1,2}(n) = h_{1,2}(n); \\ &\vdots \\ &h_{1,l-1}(n) - m_{1,l}(n) = h_{1,l}(n); \end{aligned} \tag{4.1}$$

If $h_{1,l}(n)$ satisfies the sifting conditions, then it is selected as an IMF, i.e., $\tilde{z}_1(n) \leftarrow h_{1,l}(n)$. It is straightforward to reach from (4.1) that

$$\tilde{z}_1(n) = x(n) - (m_{1,1} + m_{1,2} + ... + m_{1,l}) \tag{4.2}$$

The other IMF components can be similarly extracted, i.e.,

$$\begin{aligned} x(n) - \tilde{z}_1(n) &= r_1(n); \\ r_1(n) - \tilde{z}_2(n) &= r_2(n); \\ &\vdots \\ r_{C-1}(n) - \tilde{z}_C(n) &= r_C(n); \end{aligned} \tag{4.3}$$

As a result, $x(n)$ is decomposed into a sum of C IMFs and a residue $r_C(n)$ (assuming $r_C(n)$ is a monotonic function),

$$x(n) = \sum_{j=1}^{C} \tilde{z}_j(n) + r_C(n) \tag{4.4}$$

where $\tilde{z}_j(n)$ represents the jth IMF component. Typically, C was set to 15 in the simulations, where different values of C have also been tested which however give similar results.

4.2.2 IMFs of speech signals for denoising

Only part of IMFs are selected for the denoising in the next subsection 4.2.3. In order to explain the reason behind the selection of these IMFs, an example is given here in which first the noisy speech signal is generated by adding white Gaussian noise to the clean speech signal at SNR= 4 dB. Then, the EMD algorithm is used to derive the IMF components of the clean and its corresponding noisy signal. In Figures 4.2 and 4.3, all the IMF components (ranging from high to low frequencies) derived from the clean speech and its corresponding noisy speech are shown respectively. From the comparison of these two figures, it can be observed that the noise is mainly present in the high frequency components. Motivated from this observation the high frequency IMF components $\tilde{z}_j(n), j = 1, ..., R$ have been chosen for denoising. In this work R=10 is used in the experiments, which is found empirically to be a suitable number.

4.2.3 EMD-MMSE filtering for noise reduction of speech

In this step, denoising is performed for the selected high frequency IMF components $\tilde{z}_j(n)$, where $j = 1, ..., R$, using the MMSE filter [83]. In general, speech noise can be

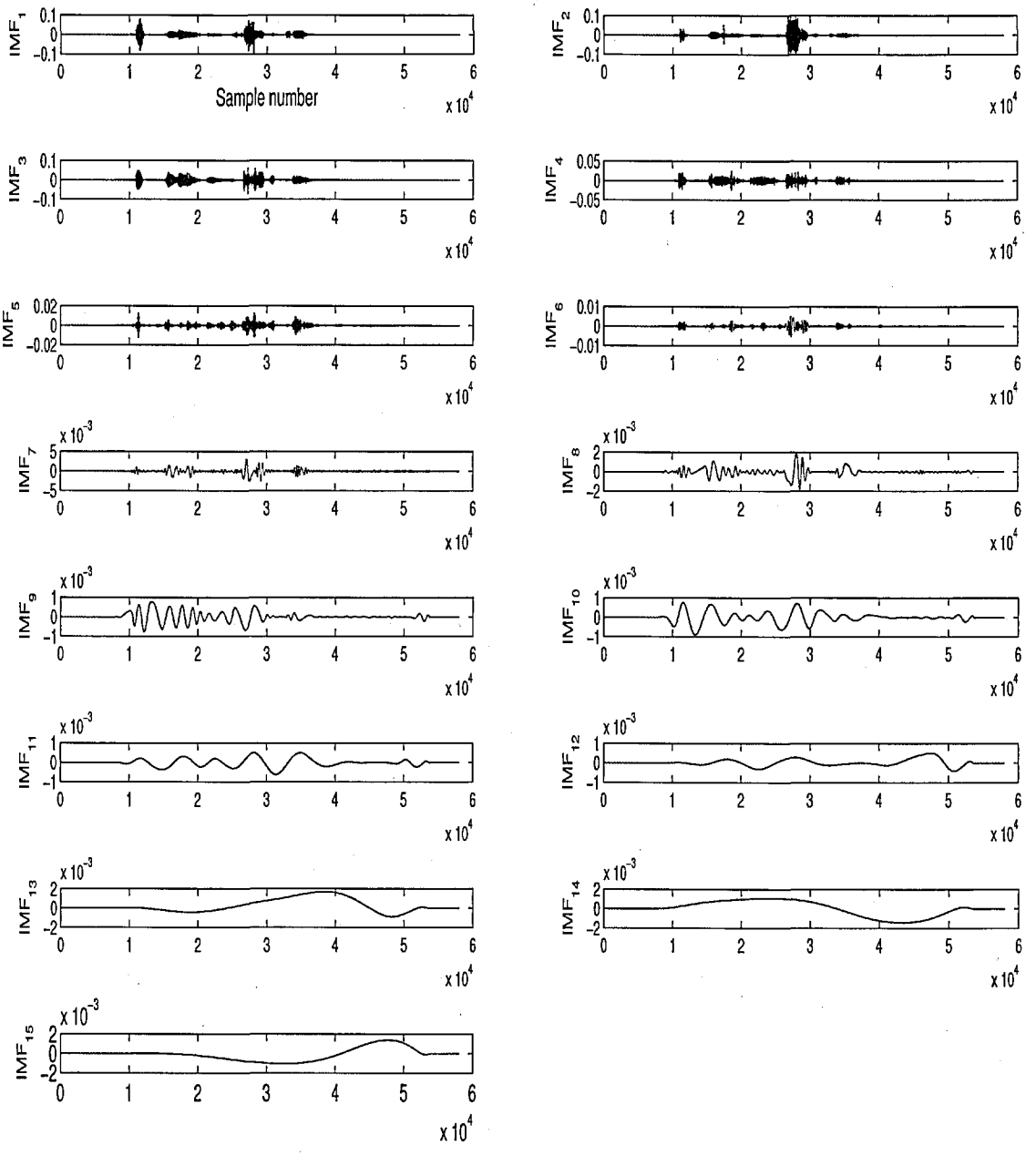


Figure 4.2: The IMF components derived from the clean speech signal. There are 15 IMF components ranging from high to low frequencies.

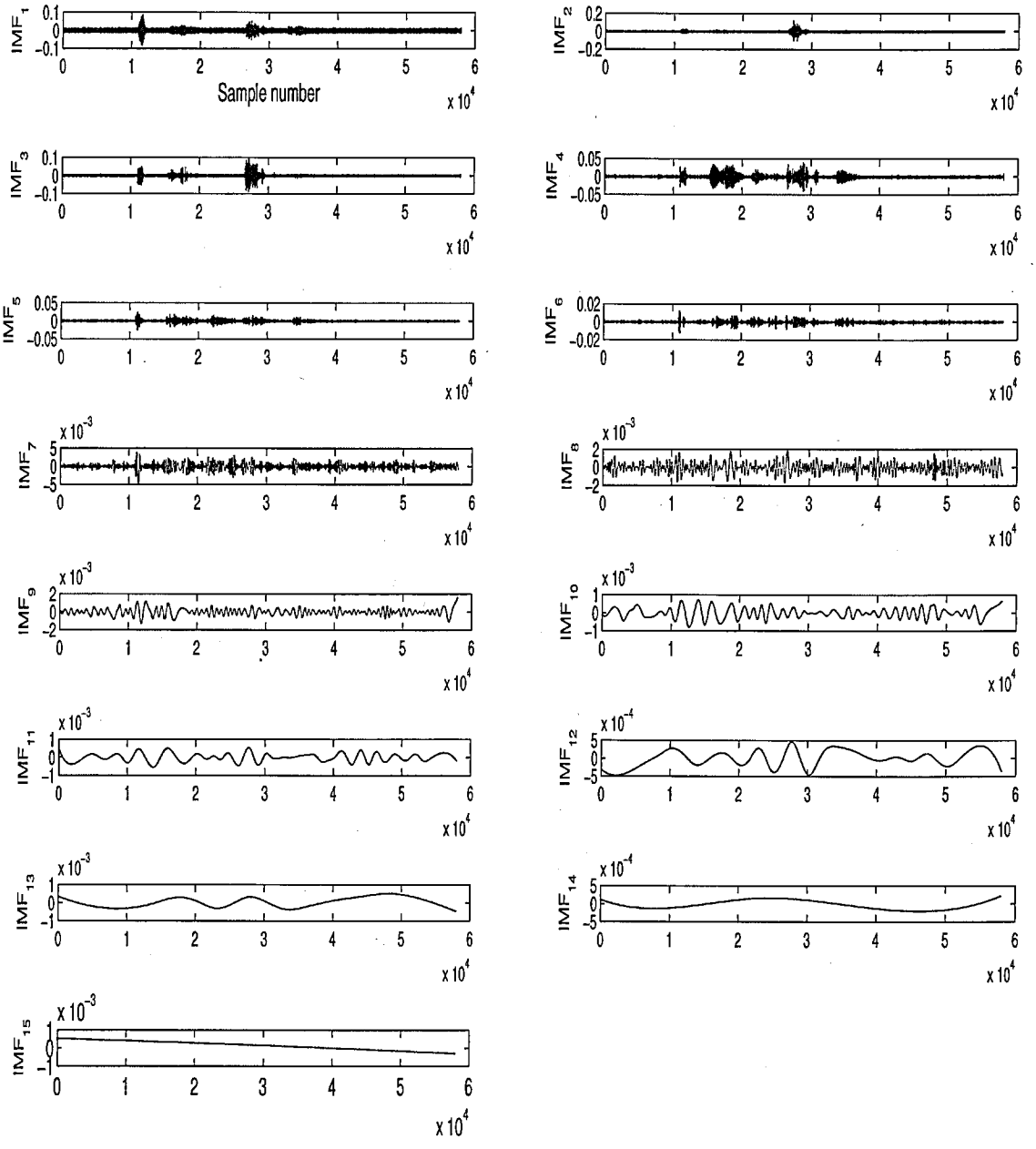


Figure 4.3: The IMF components derived from the noisy speech signal. Again 15 IMF components ranging from high to low frequencies are taken for purpose of comparison with Figure 4.2.

estimated using Boll's method [17]. The silence periods of the signal are detected and then the noise power spectrum is estimated by averaging the power spectra of the noisy signal on the M first temporal frames corresponding to the silence period. Here the first R IMFs are used separately in order to estimate the noise power, following the relation [83]

$$| \hat{B}_j(k) |^2 = \frac{1}{M} \sum_{i=0}^{M-1} | B_j(k;i) |^2, \quad j = 1, ..., R \tag{4.5}$$

where $| B_j(k;i) |$ represents the magnitude spectrum of the jth IMF component at the discrete frequency k and time frame i (index used for the silence period), and $| \hat{B}_j(k) |^2$ is the estimated noise power of the jth IMF component at frequency bin k.

The combined operation of EMD and MMSE filter [47,48] is named as EMD-MMSE. Hence each IMF is filtered by the MMSE filter as follows:

$$\hat{z}_j(k;m) = H_j(k;m)\tilde{z}_j(k;m), \quad j = 1, ..., R \tag{4.6}$$

where $\hat{z}_j(k;m)$ and $\tilde{z}_j(k;m)$ are the spectra of the jth estimated IMF and noisy IMF components respectively, observed at the discrete frequency k and the time frame m. $H_j(k;m)$ can be defined as follows [47]

$$H_j(k;m) = \frac{SNR_{prio}(k;m)}{1 + SNR_{prio}(k;m)} \tag{4.7}$$

The signal to noise ratio, SNR_{prio} can be estimated based on the previous frame of the estimated $\hat{z}_j(k;n-1)$ and a local estimation of SNR_{inst}, given as [47]

$$SNR_{prio}(k;m) = \alpha \frac{\hat{z}_j^2(k;m-1)}{\hat{B}_j^{\,2}(k)} + (1-\alpha) max(SNR_{inst}(k;m), 0) \tag{4.8}$$

where α is a weighting factor (chosen empirically to be 0.98 in this work), max denotes the maximum element of its argument, and SNR_{inst} represents the instantaneous SNR, and can be defined as the local estimation of SNR_{prio},

$$SNR_{inst} = \frac{\tilde{z}_j^2(k;m)}{\hat{B}_j^{\,2}(k)} \tag{4.9}$$

Hence $\hat{z}_j(k;m)$ with $j = 1, ..., R$, obtained in equation (4.6) are the denoised IMF components which are further processed in the next step in order to remove the late reverberations from these components.

4.2.4 IMFs based spectral subtraction for the suppression of late reverberations

It has been observed that the late reverberations lead to the blurring effect on the speech spectrum in the frequency domain, resulting in a smoothed spectrum [179]. Therefore, the power spectrum of the late reverberation components can be estimated as the smoothed and shifted version of the power spectrum of the denoised reverberant IMF components $\hat{z}_j(k,m), j = 1, ..., R$ and remaining low frequency components, $\tilde{z}_j(k,m), j = R+1, ..., C$. For notational simplicity, all of these components are now represented by $\hat{z}_j(k,m)$ where $j = 1, ..., C$.

$$|S_{l_j}(k;m)|^2 = \gamma\omega(m-\rho) * |\hat{z}_j(k;m)|^2 \tag{4.10}$$

where $|S_{l_j}(k;m)|^2$ is the short term power spectrum of the late reverberations in the jth IMF component, γ is the scaling factor specifing the relative strength of the late reverberation components, the symbol $*$ denotes the convolution operation, $\omega(m)$ is a smoothing window, and ρ refers to the relative delay of the late reverberations. The short-term speech spectrum can be obtained by using the Hamming window of length 16 msec with 8 msec overlap for the short-term Fourier analysis.

To estimate the power spectrum of the original speech, the power spectrum of the late reverberation components can be subtracted from that of the IMF components $\hat{z}_j,\ j = 1, ..., R$. Spectral subtraction can be employed for each selected component as follows [179],

$$|\tilde{s}_j(k;m)|^2 = |\hat{z}_j(k;m)|^2 max\left[\frac{|\hat{z}_j(k;m)|^2 - \gamma_j\omega(m-\rho) * |\hat{z}_j(k;m)|^2}{|\hat{z}_j(k;m)|^2}, \varepsilon\right] \tag{4.11}$$

where $|\tilde{s}_j(k;m)|^2$ represents the power spectrum of the jth IMF component of the estimated version of the original speech, ε stands for the floor parameter which was set to be 0.001 in the experiments, corresponding to the maximum attenuation of 30 dB and γ_j is a scaling factor, discussed below. The spectral subtraction procedure discussed above in equation (4.10) and (4.11) were also used for all the IMF components including the remaining low frequency IMFs $\tilde{z}_j,\ j = R+1, ..., C$.

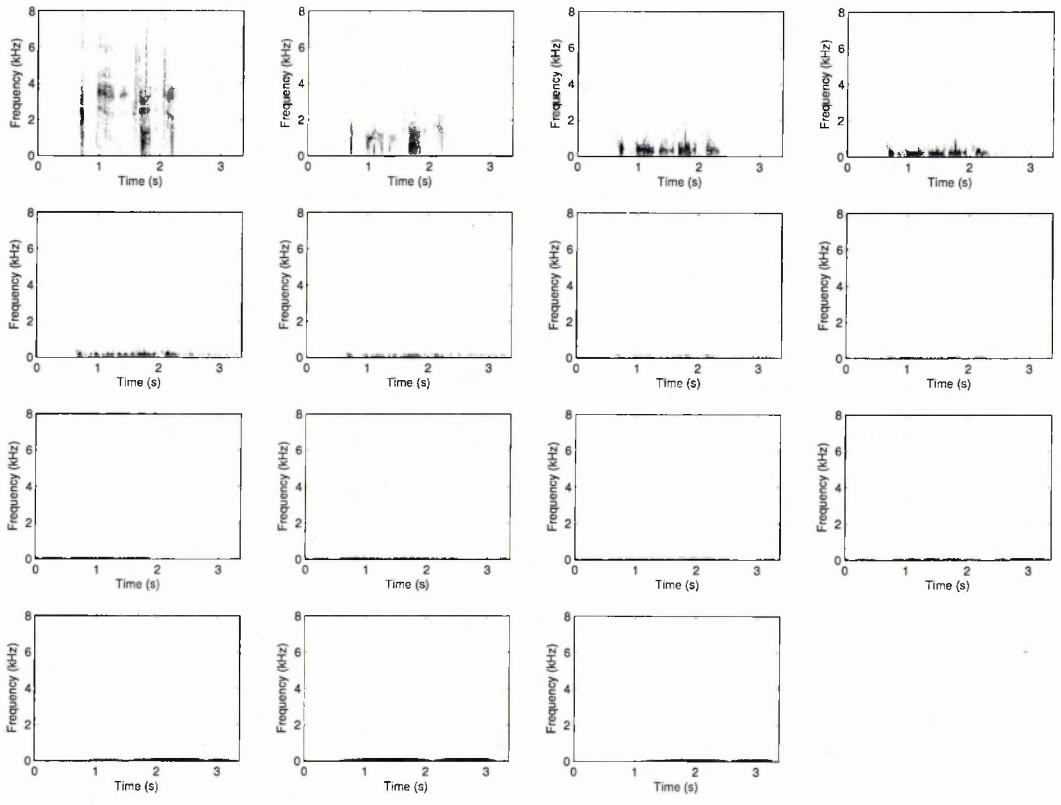


Figure 4.4: The spectrograms of the subtracted IMFs shown in the descending order of frequency patterns with the highest frequency component on the top left and the lowest frequency component on the bottom right.

4.2.5 Selection of variable scaling factor γ_j

The variable scaling factor γ_j is used for the estimation of the late reverberations from the IMF components. To show the motivation for using variable γ_j, an example is provided here in which the IMF components of the reverberant speech signal (at RT= 200 msec) and the clean speech signal are taken. Then, the IMF components of clean speech signal are subtracted from the corresponding IMF components of the reverberant signal to obtain the distribution of the energy of late reverberations. The spectrograms of the subtracted IMF components are shown in Figure 4.4. From this figure, it can be observed that the late reverberations tend to spread over the different

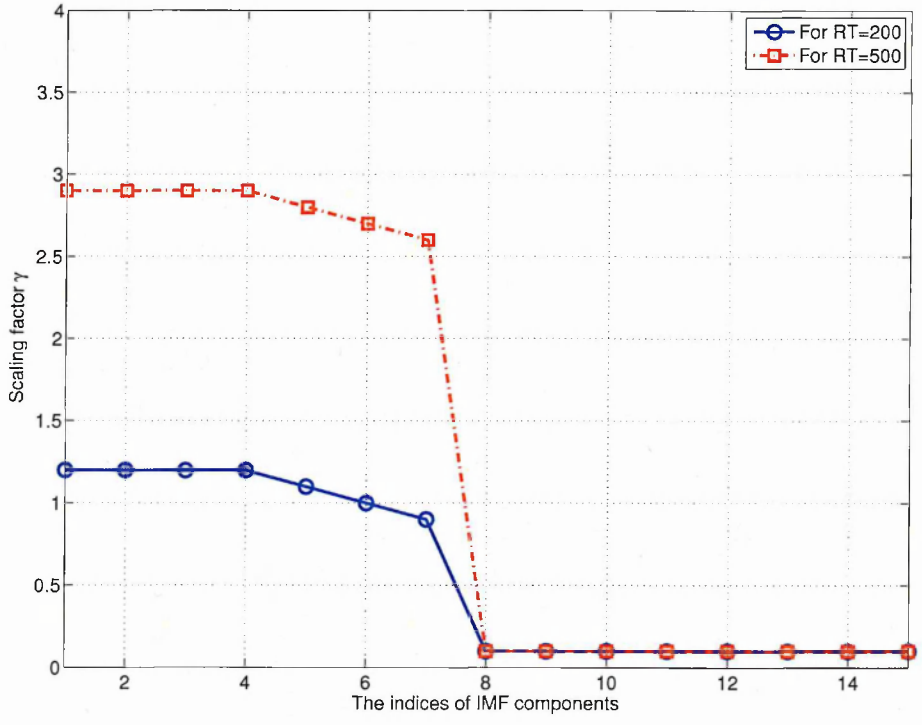


Figure 4.5: Variable scaling factor γ_j, j=1,...,15. Note that the first 7 IMF components contain more diffusive noise, and therefore scaling factor has high values for them.

IMFs with variable energy, i.e. having high energy in the first few high frequency IMFs and decreases in the lower IMFs. Motivated by this fact, it is proposed to use variable scaling factors γ_j instead of a fixed one (as used in method [179]). The high values of γ are selected for the first few high frequency IMF components while being decreased for the lower frequency components. Different values for γ have been tested. The optimized values of γ for each corresponding IMF component are shown in Figure 4.5 where the RT is equal to 200 and 500 msec respectively.

Table 4.1: The proposed EMD based method for joint denoising and dereverberation

Task: Use EMD for the enhancement of noisy reverberant speech.

Input: $x(n)$.

Output: $\hat{s}(n)$.

Initialization: 1) In (4.4), $x(n)$ is split into the sum of $C = 15$ IMFs.

2) In (4.5), $R = 10$ IMFs are used.

3) In (4.8), $\alpha = 0.98$ is used.

Part A: The goal is to denoise $x(n)$. The steps are:

1) Use (4.1)-(4.4) to split $x(n)$ into the sum of C IMFs, i.e., $\tilde{z}_j(n)$, $j = 1, ..., C$.

2) Use (4.5)-(4.9) for the R IMFs (i.e., $\tilde{z}_j(n)$, $j = 1, ..., R$) in order to reduce noise, leaving $(C - R)$ IMFs (i.e., $\tilde{z}_j(n)$, $j = R+1, ..., C$) unprocessed.

Part B: The goal is to dereverberate $x(n)$. The steps are:

- Use (4.10) and (4.11) for the processed R IMFs from Part A (i.e., $\hat{z}_j(n)$, $j = 1, ..., R$) and the unprocessed IMFs (i.e., $\tilde{z}_j(n)$, $j = R+1, ..., C$), to achieve dereverberation. γ_j in (4.11) is used in two ways, i.e., for low and high reverberation conditions.
  - (a) For low reverberant condition
    - If $(j = 1, ..., 4)$, then $\gamma_j = 1.2$
    - Else if $(j = 5, 6, 7)$, then $\gamma_j = 1.1, 1.0, 0.9$
    - Else if $(j = 8, ..., C)$, then $\gamma_j = 0.1$
  - (b) For high reverberant condition
    - If $(j = 1, ..., 4)$, then $\gamma_j = 2.9$
    - Else if $(j = 5, 6, 7)$, then $\gamma_j = 2.8, 2.7, 2.6$
    - Else if $(j = 8, ..., C)$, then $\gamma_j = 0.1$

Output: Compute $\hat{s}(n)$ according to (4.12).

4.2.6 Signal reconstruction

Finally, the enhanced signal $\hat{s}(n)$ can be reconstructed by the superposition of the processed IMFs, and the residue, given as follows,

$$\hat{s}(n) = \sum_{j=1}^{R} \tilde{s}_j(n) + \sum_{j=R+1}^{C} \tilde{s}_j(n) + r_C(n) \tag{4.12}$$

where $\tilde{s}_j(n)$ is computed as the inverse FFT of $\tilde{s}_j(k;m)$ obtained in (4.11). The proposed algorithm is summarized in Table 4.1.

4.3 Experimental Results and Discussions

In this section, the performance of the proposed method is evaluated using simulations. Four clean speech utterances, 2 male and 2 female all sampled at 16 kHz were used. The

simulated RIRs from the image model [4] and the real RIRs from the AIR database [79] were used to generate the reverberant signals from the clean speech signals with different RTs, which were then added by white Gaussian noise with SNR values ranging from -12 dB to 4 dB. The size of the room used in the case of simulated RIRs is $10 \times 10 \times 10$, and the microphone and speaker were positioned at [3, 8, 5] and [2, 2, 5] respectively (the unit is meter) [4]. The performance index used in the evaluations is the SNR [133]. The SNR in dB can be defined as,

$$SNR = 10log_{10}\frac{\sum_{i=1}^{N}(s(n_i))^2}{\sum_{i=1}^{N}(s(n_i) - \hat{s}(n_i))^2} \tag{4.13}$$

where $s(n_i)$ and $\hat{s}(n_i)$ are the original signal and the enhanced signal respectively, and N is the length of the signal.

First, an experiment has been carried out using the proposed method for the noisy speech signals without room reverberations. Four clean speech signals described above have been used in this experiment to generate noisy speech signals with SNR ranging from -12 dB to 4 dB. In total 50 independent random tests have been conducted for each SNR, and the average results were computed. The results are shown in Figure 4.6. It can be observed from this figure that, for the input SNR ranging from -12 dB to 4 dB, the output SNR ranges from 1.5 dB to 6.1 dB (approximately), which shows the reasonably good performance of the proposed method for denoising.

In a further experiment, numerical simulations have been performed using simulated RIRs for RT= 200 and 500 msec respectively, with SNR ranging from -12 dB to 4 dB for each RT. In total, 50 independent random tests have been conducted for each SNR, and the average results were calculated. In order to ensure a fair comparison between the proposed approach and the method in [179] (called for short Wu *et al.* method hereafter), EMD-MMSE has also been applied as a preprocessing step to the Wu *et al.* method. Figure 4.7 shows the comparison of the methods for the signals in terms of SNR obtained for RT= 200 and 500 msec respectively, and for different noise levels. From Figure 4.7, it can be observed that the proposed algorithm offers improvement over the Wu *et al.* method with EMD-MMSE preprocessing, especially for RT equal to 500 msec, and comparable performance is observed for RT equal to 200 msec. As compared to the results obtained by Wu *et al.* method without incorporating EMD-

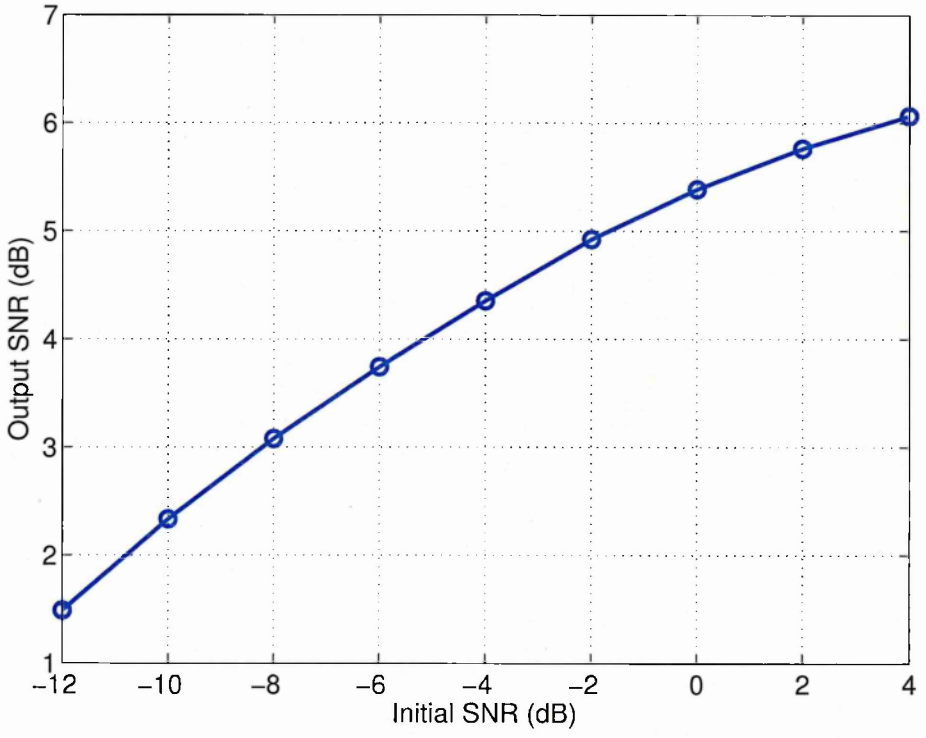


Figure 4.6: Average gain in SNR for the proposed method with different initial noise levels without room reverberation. Results are the average of 50 random tests.

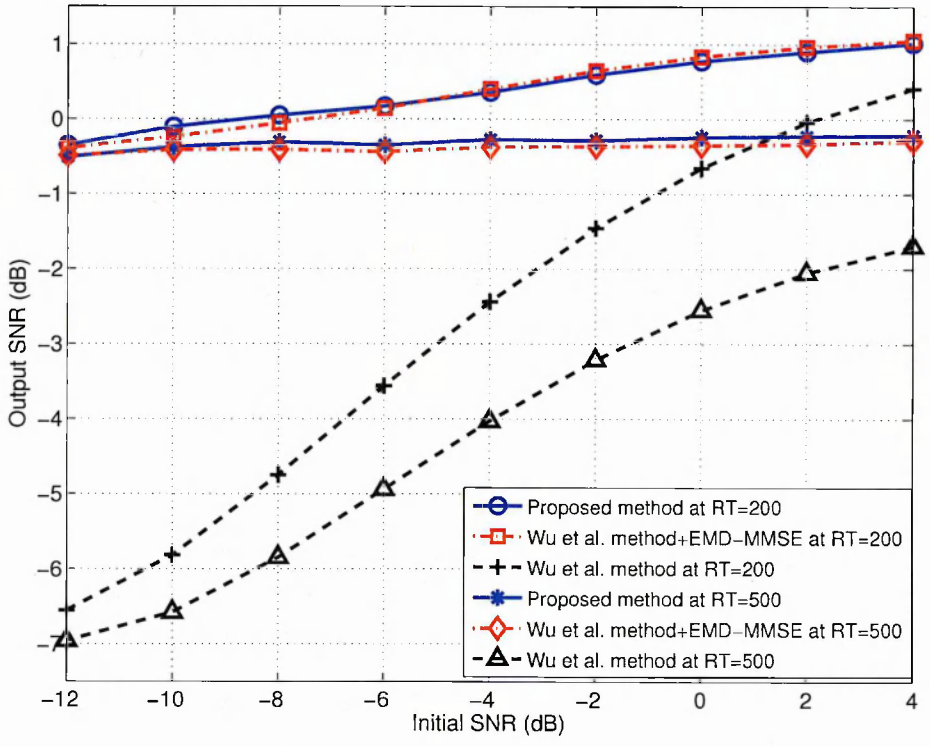


Figure 4.7: Average gain in SNR for RT= 200 msec and 500 msec with different initial noise levels. Results are the average of 50 random tests.

MMSE preprocessing, the proposed method has shown considerably higher performance improvement.

Another set of experiments have been carried out using simulated RIRs from the image model in which the performance of the proposed approach and the Wu *et al.* method is evaluated and compared with and without EMD-MMSE filtering on the basis of different source-microphone distances. The RT used in this set of experiments for all the four signals is 500 msec with initial SNR= -4 dB. Average results for all the speech signals based on 50 random tests, are depicted in Figure 4.8. It can be observed that as the distance between the source and the microphone decreases the average performance of both algorithms increases. In addition, it should be noted that the proposed method performs better for larger source-microphone distances.

Now the performance of the proposed method is evaluated based on the real data

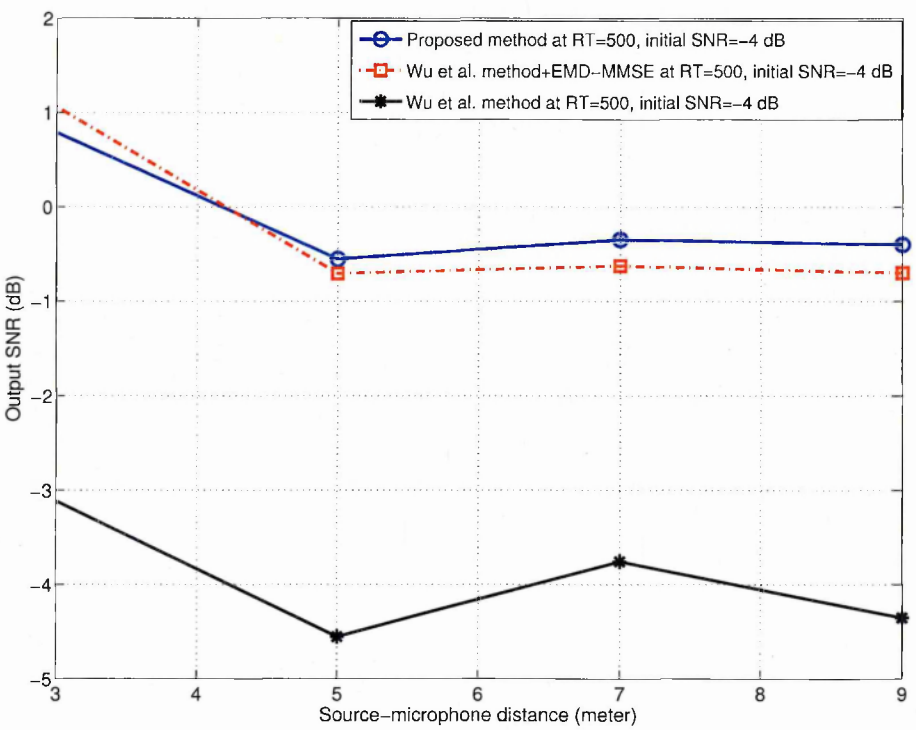


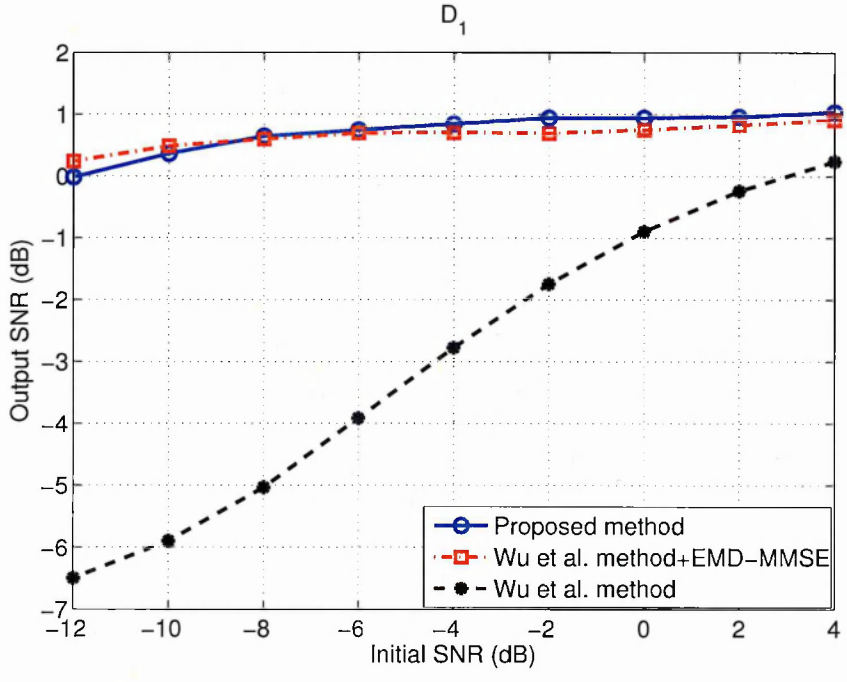
Figure 4.8: Average gain in SNR for different source-microphone distances where RT= 500 msec with initial noise level equal to -4 dB. Results are the average of 50 random tests.

from the AIR database [79]. Five different types of RIRs have been used from the AIR database recorded in five different types of room environments, namely booth, office, meeting, lecture, and stairway. For each room environment, a shorter source-microphone distance and a longer source-microphone distance are used in the experiments, denoted in pair as $\{D_1, D_2\}$ m respectively. Specifically the pair $\{D_1, D_2\}$ used for each room is, $\{0.5, 1.5\}$, $\{1, 3\}$, $\{1.45, 2.8\}$, $\{2.25, 7.1\}$, and $\{1, 3\}$ m, respectively. Four clean speech signals are then convolved with each of these RIRs, with SNR ranging from -12 dB to 4 dB for each RIR to generate the noisy reverberant speech signals. In total 50 independent random tests have been conducted for each SNR, and the average results were computed. The results obtained for the proposed method in comparison to the Wu *et al.* method with and without EMD-MMSE preprocessing are shown for the five different types of rooms at $\{D_1, D_2\}$ m in Figures 4.9, 4.10, 4.11, 4.12, and 4.13 respectively. It can be observed that for different rooms the proposed method offers improvement over the Wu *et al.* method with EMD-MMSE preprocessing, especially for low direct-to-reverberant ratios (i.e., at D_2), and comparable performance is observed at shorter source-microphone distance (i.e., D_1), where the direct-to-reverberant ratio is higher. As compared to the results obtained by Wu *et al.* method without incorporating EMD-MMSE preprocessing, the proposed method has shown considerably higher performance improvement for all the five rooms at both distances (i.e., at D_1 and D_2).

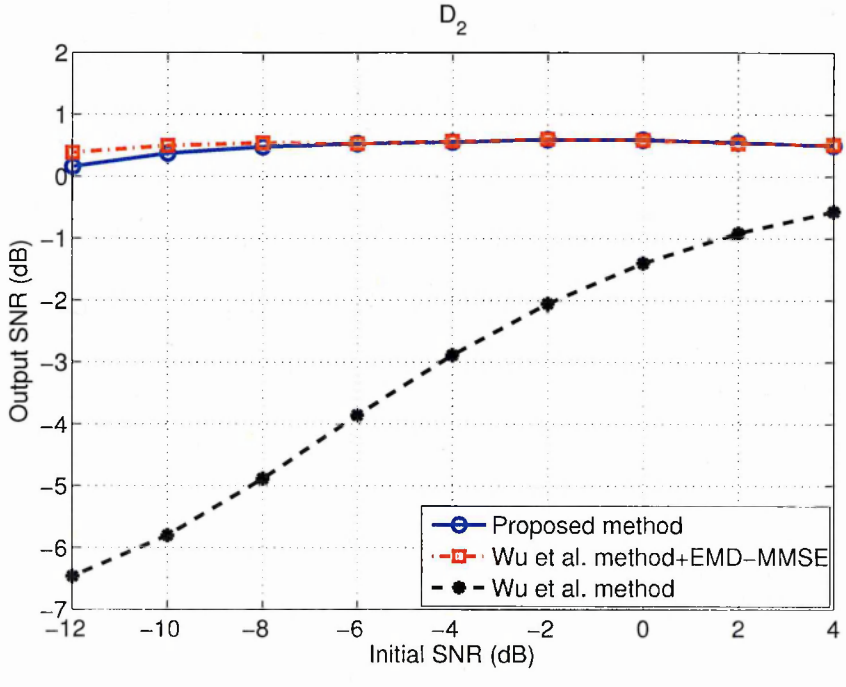
4.4 Summary

In this chapter a novel approach has been presented for speech denoising and dereverberation, based on the EMD decomposition of the noisy reverberant speech. EMD based MMSE and spectral subtraction have been applied to process the IMF components separately. It has been observed that both the additive noise and the late reverberations are spread over the different IMF components in varying magnitudes. As shown in the experiments, performing MMSE and spectral subtraction on individual subband components offers better denoising and dereverberation performance as compared with a related method that directly uses the noisy reverberant speech.

Although it has been shown in this chapter that EMD performs very well in enhance-

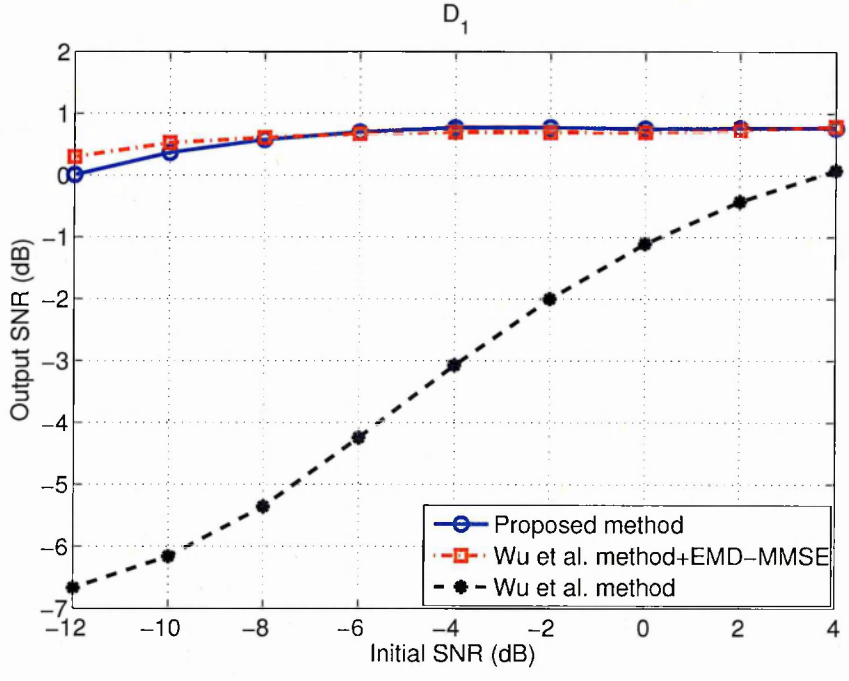


(a)

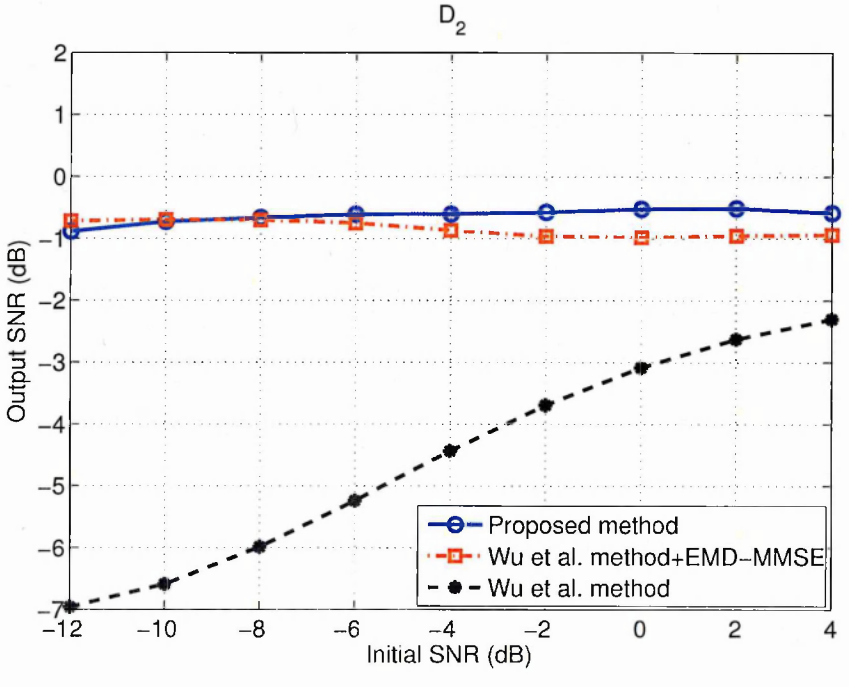


(b)

Figure 4.9: (a) Average output SNR for the booth room from the AIR database with different initial noise levels, based on 50 random tests, at source-microphone distance (a) D_1, (b) D_2.

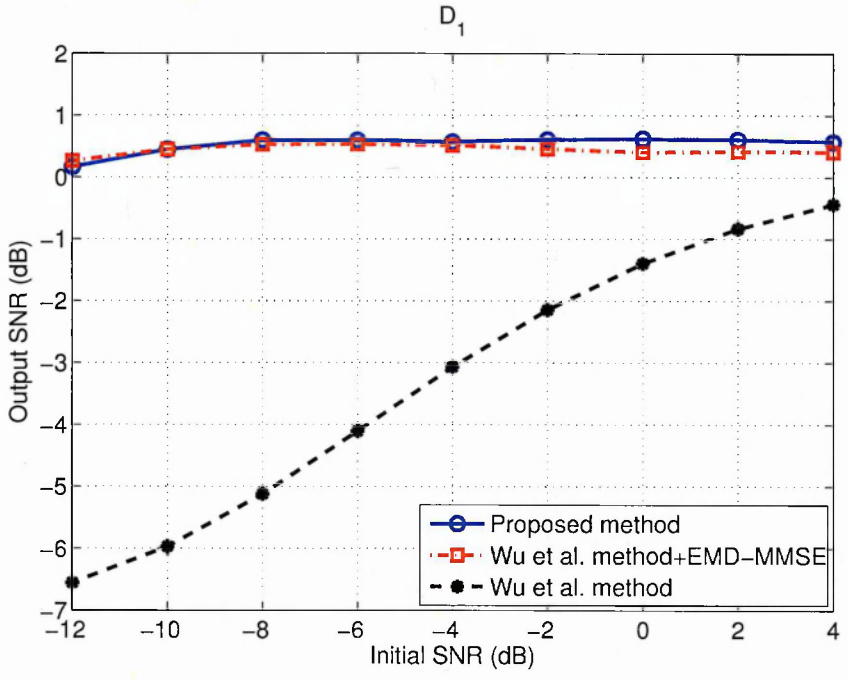


(a)

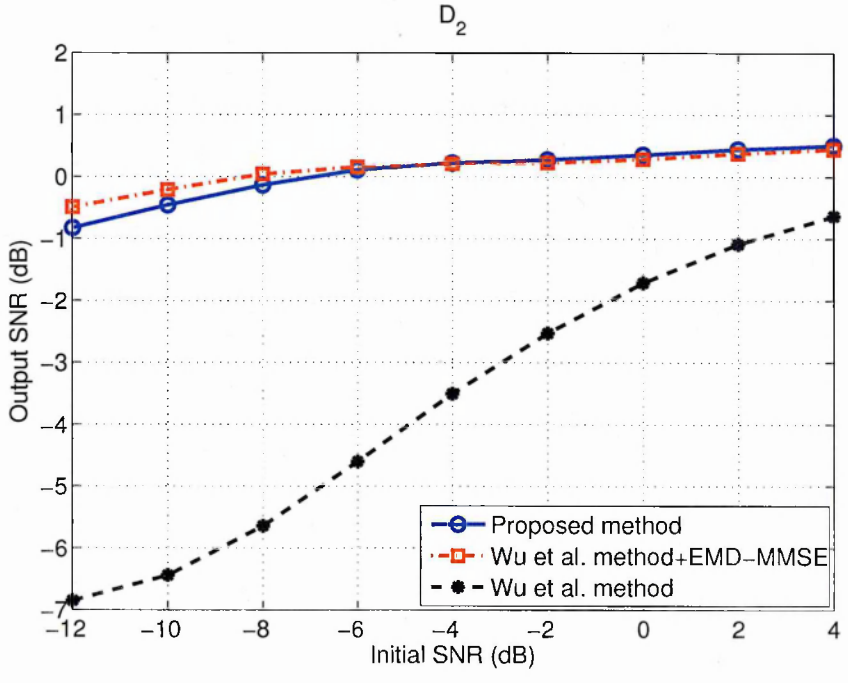


(b)

Figure 4.10: (a) Average output SNR for the office room from the AIR database with different initial noise levels, based on 50 random tests, at source-microphone distance (a) D_1, (b) D_2.

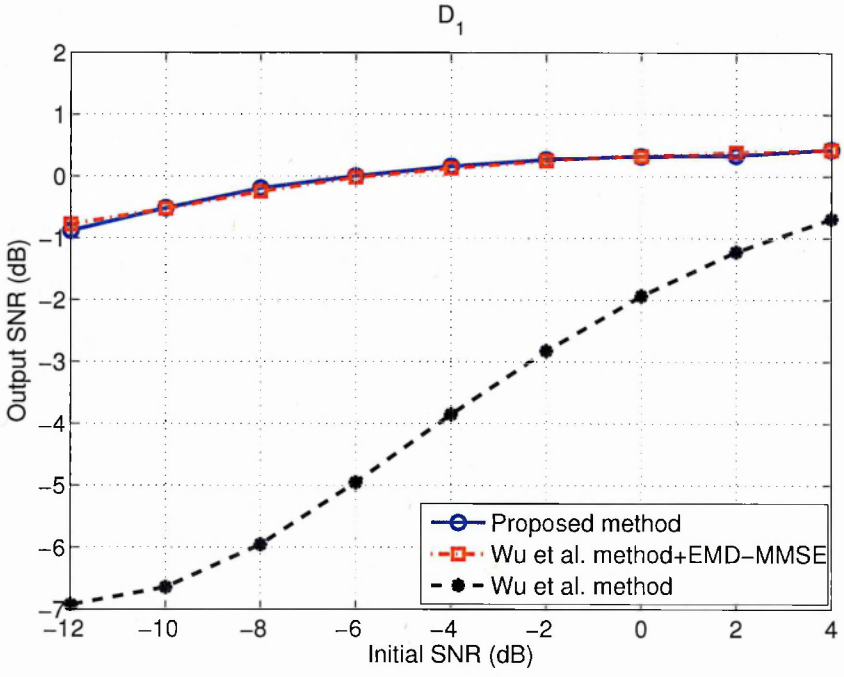


(a)

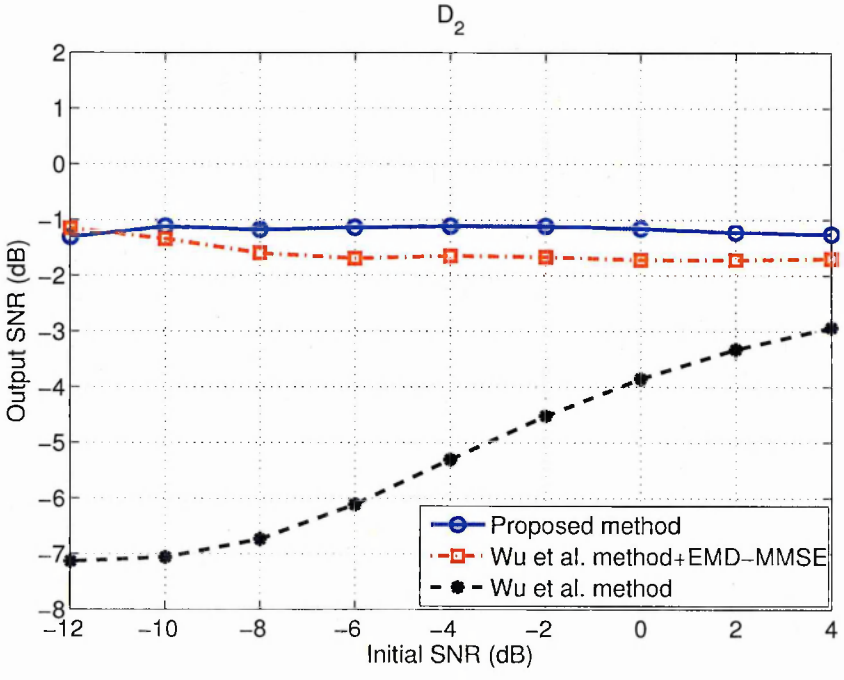


(b)

Figure 4.11: (a) Average output SNR for the meeting room from the AIR database with different initial noise levels, based on 50 random tests, at source-microphone distance (a) D_1, (b) D_2.

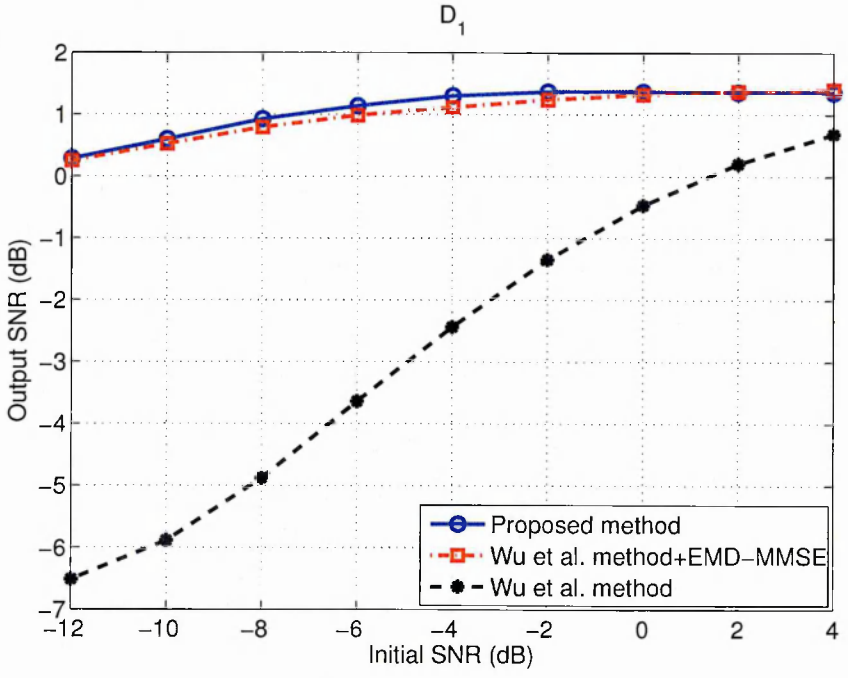


(a)

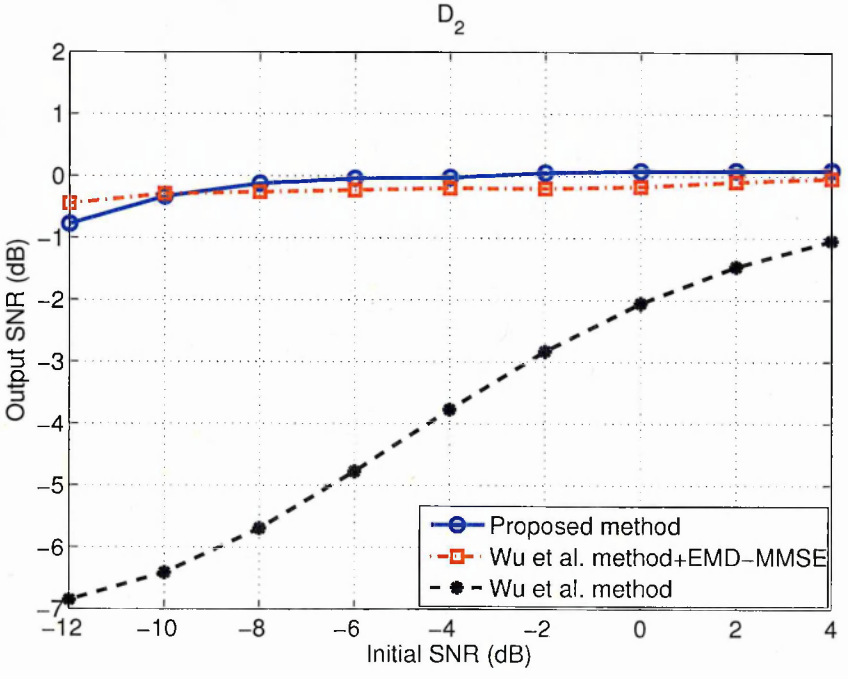


(b)

Figure 4.12: (a) Average output SNR for the lecture room from the AIR database with different initial noise levels, based on 50 random tests, at source-microphone distance (a) D_1, (b) D_2.



(a)



(b)

Figure 4.13: (a) Average output SNR for the stairway from the AIR database with different initial noise levels, based on 50 random tests, at source-microphone distance (a) D_1, (b) D_2.

ment of the noisy reverberant speech, in particular, for the reduction of additive white Gaussian noise, its performance in mitigating the reverberation distortion, as observed in the experiments, is still limited. Therefore, in the next chapter, dereverberation problem is further studied where new solutions are developed to enhance the reverberant speech.

Chapter 5

Suppression of Late and Early Reverberations Using a Frequency Dependent Statistical Model

Suppression of room reverberations is a challenging problem in reverberant speech enhancement. A promising recent approach to this problem is to apply a spectral subtraction mask to the spectrum of the reverberant speech, where the spectral variance of the late reverberations was estimated based on a frequency independent statistical model of the decay rate of the late reverberations, followed by a dual-channel Wiener filter to mitigate the early reflections. In this chapter, a two stage dereverberation algorithm is developed by following a similar process. Instead of using the frequency independent model, however, in this work the frequency dependent reverberation time and decay rate are estimated, and then used for the estimation of the spectral subtraction mask. In order to remove the processing artifacts, the mask is further filtered by a smoothing function, and then applied to reduce the late reverberations from the reverberant speech. In a second stage, a dual-channel Wiener filter is applied such that the early reverberations are attenuated. The performance of the proposed algorithm, measured

by the segmental signal to reverberation ratio (SegSRR) and the signal to distortion ratio (SDR), is evaluated for both simulated and real data. As compared with the related frequency independent algorithm, the proposed algorithm offers a considerable performance improvement.

5.1 Introduction

As mentioned in Chapter 2 the room reverberations degrade speech quality and intelligibilty. Hence a method should be developed to reduce their effects. Different methods have been proposed in the literature (as discussed in Chapter 2) to deal with the detrimental effects of room reverberations. Recently, Lebart *et al.* [93] proposed a statistical model for late reverberations. With this model, the spectral variance of the late reverberations can be estimated from the reverberant speech [93], which was further used by Jeub *et al.* for the suppression of late reverberations [78]. This original model was developed as frequency independent where a fixed reverberant time (T_{60}) was used for all the frequency channels in the estimation of the decay rate of room reverberations. However, it was found by Habets *et al.* [62] that the spectral variance of the late reverberations is frequency dependent. In this chapter, a new dereverberation algorithm is proposed with a frequency dependent model for the late reverberations in the first stage followed by a dual-channel Wiener filter to reduce the early reflections in the second stage, which is based on the coherence model of the reverberant sound field. Section 5.2 formulates the problem and its model. Section 5.3 describes the first stage of the proposed approach which includes the estimation of frequency dependent T_{60} from room impulse responses (RIRs), the estimation of the spectral subtraction mask, and the filtering (smoothing) of the mask. Section 5.4 describes the second stage of the proposed method. Section 5.5 presents the evaluation results, followed by a conclusion in Section 5.6.

5.2 Problem Formulation and Modelling

The reverberant speech signal $x(n)$ can be modelled as the convolution of the anechoic speech signal $s(n)$ and the RIRs $h(n)$ [117],

$$x(n) = \sum_{l=0}^{\infty} h(l)s(n-l) \tag{5.1}$$

where n is the discrete time index. Note that the mathematical formulation provided here will be for single channel case. However, an extension for each of the two channels can be performed in a similar way. The RIR of length T_r in seconds can be modelled as [93]

$$h(n) = \begin{cases} h_{early}(n) & \text{for } 0 \leq n < T_{le} \cdot f_s, \\ h_{late}(n) & \text{for } T_{le} \cdot f_s \leq n \leq T_r \cdot f_s, \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $h_{early}(n)$ denotes the direct and early path, $h_{late}(n)$ is the late reflection path, f_s is the sampling frequency, and T_{le} is the time after which we assume that the late reverberation starts. The range of T_{le} usually lies within 50 to 100 ms.

The reverberant speech signal can now be represented as the combination of two main parts, i.e., $x_{early}(n)$ and $x_{late}(n)$,

$$x(n) = \underbrace{\sum_{l=0}^{T_{le}f_s-1} s(n-l)h(l)}_{x_{early}(n)} + \underbrace{\sum_{l=T_{le}f_s}^{T_r f_s} s(n-l)h(l)}_{x_{late}(n)} \tag{5.3}$$

In order to reduce the effects of early reflections ($x_{early}(n)$), inverse filtering may be used as in [179] and [13]. For the suppression of late reverberations ($x_{late}(n)$), a spectral subtraction technique such as [93], [179], [61] is usually employed, where the spectral variance of the late reverberations is estimated from the reverberant speech. A recent technique for the spectral variance estimation was proposed by Lebart *et al.* [93], [94] in which the late impulse responses are statistically modelled as

$$h_{late}(n) = \begin{cases} \beta(n)e^{-\alpha_1 n} & \text{for } n \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

where $\beta(n)$ is a sequence of zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables, and α_1 denotes the decay rate given as

$$\alpha_1 = \frac{3ln(10)}{T_{60}f_s} \tag{5.5}$$

where ln is the natural logarithm. Using the above model originally proposed by Lebart *et al.* in [93], [94], Jeub *et al.* [79], [78] have recently presented a dereverberation algorithm with a frequency independent α_1. However, it was shown in [62] that a frequency dependent α_1 may provide more accurate estimation of the spectral variance of the late reverberations. In the next section the first stage of a new dereverberation algorithm is presented using this frequency-dependent model.

5.3 The Proposed Frequency Dependent Dereverberation Method for Late Reverberation

5.3.1 Frequency dependent RIR model

Applying the short-time Fourier transform (STFT), equation (5.2) can be rewritten in the T-F domain as

$$H(m,k) = \begin{cases} H_{early}(m,k) & \text{for } 0 \leq m < N_{le}, \\ H_{late}(m,k) & \text{for } N_{le} \leq m \leq N_r, \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

where N_{le} and N_r are the number of frames corresponding to T_{le} and T_r respectively. $H_{late}(m,k)$, the STFT of $h_{late}(n)$, is represented as

$$H_{late}(m,k) = \sum_{n=T_{le}\cdot f_s}^{T_r\cdot f_s} h(n)w(n-mR)e^{-j2\pi/Nk(n-mR)} \tag{5.7}$$

where m is the time frame index, k is the frequency bin index, w is the analysis window of length N, and R denotes the hop size.

With the statistical model (5.4) and a frequency-dependent α_1, $H_{late}(m,k)$ can also be written as [62],

$$H_{late}(m,k) = \begin{cases} \beta(m,k)e^{-\alpha_1(k)mR} & \text{for } m \geq 1, \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

where $\beta(m,k)$ is a sequence of zero-mean mutually i.i.d. Gaussian random variables, and $\alpha_1(k)$ denotes the decay rate which can be obtained from the frequency dependent reverberation time $T_{60}(k)$ as below

$$\alpha_1(k) \triangleq \frac{3ln(10)}{T_{60}(k)f_s} \tag{5.9}$$

5.3.2 Estimation of frequency dependent reverberation time

Robust estimation of $T_{60}(k)$ directly from the reverberant signal is a challenging task to be discussed further in Chapter 6. As a proof of concept in this chapter, $T_{60}(k)$ is estimated from the RIRs which are assumed to be available. To this end, a method similar to the one defined in ISO standard (ISO 3382-1:2009) is used. First, $h(n)$ is passed through a Gammatone filter-bank to obtain sub-band signals $h(p,n)$, where p is the sub-band index. Subsequently, $h(p,n)$ are analysed using Schroeder's method [153] to estimate the reverberation time $\check{T}_{60}(p)$ in each sub-band p. Since this filterbank (indexed by p) is different from the one used in the above section (indexed by k), the $\check{T}_{60}(p)$ values need to be inter- and extra-polated to obtain the estimate of $T_{60}(k)$ in each frequency bin k.

First interpolation is applied to $\check{T}_{60}(p)$ so that $\check{T}_{60}(p)$ from each sub-band p is mapped to $\tilde{T}_{60}(f)$, where $f \in [f_c - \frac{bw}{2}, f_c + \frac{bw}{2}]$ denotes the frequency range (in Hz) of sub-band p, f_c and bw are the centre frequency and the bandwidth of this sub-band respectively. Then, smoothing is applied across the overlapped regions between the neighbouring sub-bands

$$\tilde{T}_{60}(f) = \tilde{T}_{60}(f_1) + \frac{\tilde{T}_{60}(f_2) - \tilde{T}_{60}(f_1)}{f_2 - f_1}(f - f_1) \tag{5.10}$$

where f_1 and f_2 are the frequency points of the neighbouring sub-bands at which their overlap begins and ends respectively. $\tilde{T}_{60}(f_1)$ and $\tilde{T}_{60}(f_2)$ are the reverberation times at frequency points f_1 and f_2 respectively. For non-overlapped regions, no such interpolation as (5.10) is required for $\tilde{T}_{60}(f)$. Finally, $\tilde{T}_{60}(f)$ is then mapped to the STFT sub-bands by an extrapolation method as

$$T_{60}(k) = \frac{\sum_{f=(k-1)\frac{F}{K}+1}^{k\cdot\frac{F}{K}} \tilde{T}_{60}(f)}{(F/K-1)} \tag{5.11}$$

Note that, $f = 1, 2, ..., F$, where F is the whole frequency range and K denotes the number of frequency bins (indexed by k). An alternative method without using the inter- and extra-polation process is to set the hop size as a single sample when calculating the STFT, and then calculate $T_{60}(k)$ directly for each frequency band k, which provides similar performance but is computationally more expensive.

5.3.3 Spectral subtraction mask estimation

The statistical model discussed above in equation (5.8) is valid when the energy of the direct signal is low in comparison to that of all the given reflections. As a result the spectral variance of the late reverberant speech can be estimated as [62]

$$\sigma^2_{x_{late}}(m,k) = e^{-2\alpha_1(k)RN_{le}} \cdot \sigma^2_x(m - N_{le}, k) \tag{5.12}$$

where $\sigma^2_x(m,k)$ is the variance of the reverberant speech which can be estimated by recursive averaging

$$\sigma^2_x(m,k) = e^{-2\alpha_1(k)R}[\tau \cdot \sigma^2_x(m-1,k) + (1-\tau) \cdot \mid X(m,k) \mid^2] \tag{5.13}$$

where $\tau \in [0,1]$ is a forgetting factor and $X(m,k)$ is the T-F representation of $x(n)$ in (5.3). Note that N_{le} is the number of samples after which the late reverberation begins and $e^{-2\alpha_1(k)R}$ measures the reverberation decay rate. The *posteriori* signal-to-distortion ratio (SDR) can then be estimated as follows [78]

$$\varphi(m,k) = \frac{\mid X(m,k) \mid^2}{\sigma^2_{x_{late}}(m,k)} \tag{5.14}$$

To reduce the late reverberations, apply the following spectral subtraction mask [78] to $X(m,k)$

$$\tilde{G}_{late}(m,k) = 1 - \frac{1}{\sqrt{\varphi(m,k)}} \tag{5.15}$$

In order to avoid over-estimation of $\sigma^2_{x_{late}}(m,k)$, a lower bound $\tilde{G}^{min}_{late}$ is applied to all the weighting gains in the mask.

5.3.4 Spectral gain smoothing

A common problem with spectral masking is the processing artifacts, i.e. the so-called musical noise. Therefore, similar to [78], a moving average operation is applied to $\tilde{G}_{late}(m,k)$. To this end, the power ratio between the enhanced signal and the reverberant signal is calculated. However, different from [78] in this work, this power ratio is computed at each frequency bin k and each time frame m

$$\rho_1(m,k) = \frac{\mid \tilde{G}_{late}(m,k) \cdot X_{ref}(m,k) \mid^2}{\mid X_{ref}(m,k) \mid^2} \tag{5.16}$$

where $X_{ref}(m,k)$ is the reference signal and can be obtained from the left channel and right channel microphone signals given as

$$X_{ref}(m,k) = \frac{1}{2} \cdot (X_l(m,k) + X_r(m,k)) \tag{5.17}$$

In the case of a single channel mixture $X(m,k)$, $X_{ref}(m,k)$ is simply replaced by $X(m,k)$. Then a moving average window can be generated, as follows:

$$E_s(m,k) = \begin{cases} 1, & \text{if } \rho_1(m,k) \geq C, \\ 2 \cdot \lfloor (1 - \frac{\rho_1(m,k)}{C}) \cdot \psi \rfloor + 1, & \text{otherwise.} \end{cases} \tag{5.18}$$

where C is a constant controlling the trade off between the speech distortion and reduction of musical noise, ψ is a scaling factor for determining the level of smoothing, and $\lfloor \cdot \rfloor$ rounds the argument to its nearest integer. This window function can now be used to create a smoothing filter as

$$F_s(m,k) = \begin{cases} \frac{1}{E_s(m,k)}, & \text{if } k < E_s(m,k), \\ \frac{1}{2k}, & \text{otherwise} \end{cases} \tag{5.19}$$

By convolving $\tilde{G}_{late}(m,k)$ with $F_s(m,k)$, a smoothed mask can be obtained as follows:

$$G_{late}(m,k) = \tilde{G}_{late}(m,k) * F_s(m,k) \tag{5.20}$$

Finally, the smoothed mask is applied to the T-F representation of the reverberant signals as follows:

$$\hat{S}_l(m,k) = X_l(m,k) \cdot G_{late}(m,k) \tag{5.21a}$$

Table 5.1: The proposed dereverberation method for late reverberation

Task: Use frequency dependent RIR model to suppress the late reverberation.
Input: $X_l(m,k)$ and $X_r(m,k)$.
Output: $\hat{S}_l(m,k)$ and $\hat{S}_r(m,k)$.
Initialization: 1) In (5.6), $N_{le} = 13$ is used.
2) In (5.13), $\tau = 0.1$ is used.
3) In (5.18), $C = 2.5$ and $\psi = 25$ are used.
Part A: The goal is to estimate $T_{60}(k)$ from the RIR. The steps are:
1) Use $h(n)$ from (5.2) and pass it through Gammatone filter-bank to obtain $h(p,n)$.
2) Apply Schroeder's method to $h(p,n)$ to estimate $\check{T}_{60}(p)$.
3) Use (5.10) and (5.11) to map $\check{T}_{60}(p)$ to $T_{60}(k)$.
Part B: The goal is to estimate spectral subtraction mask. The steps are:
1) Use (5.12) and (5.13) to estimate the spectral variance of late reverberant speech, i.e., $\sigma^2_{x_{late}}(m,k)$.
2) Use (5.14) and (5.15) to estimate the spectral subtraction mask, i.e., $\tilde{G}_{late}(m,k)$.
Part C: The goal is to reduce the musical noise from the spectral subtraction mask. The steps are:
1) Use (5.16)-(5.19) to generate a smoothing filter.
2) Use (5.20) to obtain the smoothed spectral subtraction mask, i.e., $G_{late}(m,k)$.
Output: Compute $\hat{S}_l(m,k)$ and $\hat{S}_r(m,k)$ according to (5.21).

$$\hat{S}_r(m,k) = X_r(m,k) \cdot G_{late}(m,k) \tag{5.21b}$$

In the single channel case, similar operation is performed as Equation (5.21) by discarding the subscript (l,r). The proposed dereverberation algorithm used for suppressing late reverberation is summarized in Table 5.1.

5.4 The Dereverberation Method for Early reverberation

The spectral subtraction rule described in Section 5.3 is employed mainly to reduce the late reverberations, and hence the early reverberation remains. Therefore, a second processing step is incorporated here to deal with the effects of early reverberations. Note that the method discussed below will only be applicable to the case of two-channel (stereo) recordings. The subsequent coherence based method exploits the low coherence of the sound field between different microphones to estimate the (direct) speech power spectral density (PSD) and to remove all non-coherent signal parts while keeping the coherent parts unaffected. Since only the direct speech shows a high coherence among

sensors, cf. [78], this approach also reduces early reverberations.

In order to derive this method, consider two general microphone signals $x_{1|2}(n)$ under the assumption that the source-microphone distance should be smaller than the critical distance (The distance between source and microphone at which the direct path energy is equal to the combined energy of the early and late reflections). The coherence between the two signals $x_{1|2}(n)$ is defined as [78],

$$coh_{x_1x_2}(\acute{f}) = \frac{\Upsilon_{x_1x_2}(\acute{f})}{\sqrt{\Upsilon_{x_1x_1}(\acute{f}) \cdot \Upsilon_{x_2x_2}(\acute{f})}} \tag{5.22}$$

where $\Upsilon_{x_1x_1}(\acute{f})$ and $\Upsilon_{x_2x_2}(\acute{f})$ are the auto-power spectral densities of $x_1(n)$ and $x_2(n)$ respectively, $\Upsilon_{x_1x_2}(\acute{f})$ denotes the cross-power spectral density between $x_1(n)$ and $x_2(n)$, and $\acute{f}$ is the frequency range of signals in Hz. The relation between the frequency bin index k and $\acute{f}$ can be described by the bin resolution as f_s/k $[Hz]$, where f_s is the sampling frequency.

Unlike Equation (5.3) in section 5.2, the reverberant signal here can be decomposed into its direct components, early reverberant components, and late reverberant components. For the sake of simplicity, decomposition provided here will be for monaural case only, as an extension for each of the binaural channels can be performed in the same manner. Note that this method can be used for two channel case only. The input signal $x(n)$ can be decomposed as [78]

$$x(n) = \underbrace{\sum_{l=0}^{T_d f_s - 1} s(n-l)h(l)}_{x_{direct}(n)} + \underbrace{\sum_{l=T_d f_s}^{T_{le} f_s - 1} s(n-l)h(l)}_{x_{early}(n)} + \underbrace{\sum_{l=T_{le} f_s}^{T_r f_s} s(n-l)h(l)}_{x_{late}(n)} \tag{5.23}$$

where T_d denotes the time span of the direct sound (including sound propagation). Note that in Section 5.3, the early speech component $x_{early}(n)$ was the target signal, now the direct speech component $x_{direct}(n)$ is the target signal. As a further remark, the early and late reverberant components received by the microphones can be represented by two additive, uncorrelated noise sources, cf. [78, 179], hence the terms noise and reverberant components are used interchangeably in the following discussion. Also the first stage of dereverberation method proposed in this chapter does not affect the coherence and therefore the outputs of the first stage can be used in this second step.

Having described the basic idea of the coherence based dereverberation method, a dual-channel Wiener filter is derived now which takes into account dual-channel coherence. A common framework for speech enhancement is based on the minimum mean square error criterion, cf. [165]. As a result the optimal weighting gains are provided by the Wiener solution [78]

$$G_c(m,k) = \frac{\Upsilon_{ss}(m,k)}{\Upsilon_{ss}(m,k) + \Upsilon_{nn}(m,k)} \tag{5.24}$$

where $\Upsilon_{ss}(m,k)$ and $\Upsilon_{nn}(m,k)$ are the auto-power spectral density of the original (clean) signal and the additive noise component respectively. As discussed previously, the term $\Upsilon_{nn}(m,k)$ is referred to the auto-power spectral density of the reverberant components.

For computing the optimal postfilter coefficients in multichannel system, several approaches have been presented in the past. They all have in common that the estimation procedure is optimized for a specific sound field model. A very well known method developed by Zelinski in [190] assumes a perfectly incoherent sound field and therefore, uncorrelated noise at different sensors. Since this assumption does not hold in real sound fields, an improved approach was developed by McCowan in [106], in which he proposed to use a model of the coherence for diffuse sound field.

First, a brief derivation of this algorithm will be given and second, the estimation of the required power spectra is discussed. Under the assumption of the same noise power spectrum across sensors, the power spectra can be described as

$$\Upsilon_{\hat{s}_r\hat{s}_r}(m,k) = \Upsilon_{ss}(m,k) + \Upsilon_{nn}(m,k) \tag{5.25}$$

$$\Upsilon_{\hat{s}_l\hat{s}_l}(m,k) = \Upsilon_{ss}(m,k) + \Upsilon_{nn}(m,k) \tag{5.26}$$

$$\Upsilon_{\hat{s}_l\hat{s}_r}(m,k) = \Upsilon_{ss}(m,k) + coh_{\hat{s}_l\hat{s}_r}(\acute{f})\Upsilon_{nn}(m,k) \tag{5.27}$$

Note that Equations (5.25) and (5.26) under the assumption of the same noise power spectrum across sensors are used to derive Equation (5.29) of the spectral weights of the Wiener filter. An estimate of the original (clean) signal auto-power spectral density can be obtained as [78, 106]

$$\tilde{\Upsilon}_{ss}(m,k) = \frac{Re\left\{\tilde{\Upsilon}_{\hat{s}_l\hat{s}_r}(m,k)\right\} - \frac{1}{2}Re\left\{coh_{\hat{s}_l\hat{s}_r}(\acute{f})\right\}\left(\tilde{\Upsilon}_{\hat{s}_l\hat{s}_l}(m,k) + \tilde{\Upsilon}_{\hat{s}_r\hat{s}_r}(m,k)\right)}{1 - Re\left\{coh_{\hat{s}_l\hat{s}_r}(\acute{f})\right\}} \tag{5.28}$$

where the tilde-operator $\{\tilde{\cdot}\}$ indicates an estimate as shown later. The function $Re\{\cdot\}$ returns the real part of its argument. Since the estimated auto-power spectral density of the signal may not be negative, a maximum threshold (coh_{max}) for the coherence function has to be applied to ensure that $1 - Re\left\{coh_{\hat{s}_l\hat{s}_r}(\acute{f})\right\} > 0$ holds for the denominator. The resulting spectral weights of the Wiener filter can now be computed by

$$G_c(m,k) = \frac{\tilde{\Upsilon}_{ss}(m,k)}{\frac{1}{2}\cdot\left(\tilde{\Upsilon}_{\hat{s}_l\hat{s}_l}(m,k) + \tilde{\Upsilon}_{\hat{s}_r\hat{s}_r}(m,k)\right)} \tag{5.29}$$

The spectral weights are further confined by a lower threshold G^c_{min} for robustness against overestimation errors (i.e., biases in measurements) and to control the amount by which reverberation is attenuated. The spectral weights are then applied to each of the two channels (i.e., left and right) by

$$\overline{S}_l(m,k) = \hat{S}_l(m,k)\cdot G_c(m,k) \tag{5.30a}$$

$$\overline{S}_r(m,k) = \hat{S}_r(m,k)\cdot G_c(m,k) \tag{5.30b}$$

After transforming $\overline{S}_l(m,k)$ and $\overline{S}_r(m,k)$ back to the time domain using the inverse STFT, the dereverberated signals $\overline{s}_l(n)$ and $\overline{s}_r(n)$ can be obtained.

The calculation of the weighting gains $G_c(m,k)$ comprises an estimation of the auto-power spectral densities, i.e., $\Upsilon_{\hat{s}_l\hat{s}_l}(m,k)$, $\Upsilon_{\hat{s}_r\hat{s}_r}(m,k)$ and cross-power spectral density $\Upsilon_{\hat{s}_l\hat{s}_r}(m,k)$ of the two input channels (i.e., left and right). A recursive approach has been used here for this purpose given as [78]

$$\tilde{\Upsilon}_{\hat{s}_l\hat{s}_l|\hat{s}_r\hat{s}_r}(m,k) = \alpha_2\tilde{\Upsilon}_{\hat{s}_l\hat{s}_l|\hat{s}_r\hat{s}_r}(m-1,k) + (1-\alpha_2)\mid\hat{S}_{l|r}(m,k)\mid^2 \tag{5.31}$$

$$\tilde{\Upsilon}_{\hat{s}_l\hat{s}_r}(m,k) = \alpha_2\tilde{\Upsilon}_{\hat{s}_l\hat{s}_r}(m-1,k) + (1-\alpha_2)\hat{S}_l(m,k)\cdot\hat{S}^*_r(m,k) \tag{5.32}$$

where $\alpha_2 \in [0,1]$ is a smoothing factor, $\hat{S}_{l|r}(m,k)$ are the left/right microphone signals obtained in (5.21), and $\hat{S}^*_r(m,k)$ is the complex conjugate of $\hat{S}_r(m,k)$.

The essential part of this work is to choose a suitable model for the sound field coherence in (5.28). The coherence model selected here is based on the binaural sound field and can be expressed as [78]

$$\tilde{coh}^{(head)}_{x_lx_r}(\acute{f}) = \sum_{q=1}^{Q} a_q\cdot exp\left(-\left(\frac{\acute{f}-b_q}{c_q}\right)^2\right) \tag{5.33}$$

Table 5.2: Coefficients and order of the binaural coherence model

q	a_q	b_q	c_q
1	1	18.97	291.1
2	$14.5 \cdot 10^{-3}$	875.2	105.7
3	$2.38 \cdot 10^{-3}$	1371	151.5

a_q, b_q, and c_q are the coefficients of the model, while q shows the order of the model. Note that this model is based on the sum of Gaussians and provide an approximation of the sound field coherence. The coefficients a_q, b_q, c_q for natural ear spacing of 0.17 m and a mixture of $Q = 3$ Gaussians are calculated using the MATLAB Curve Fitting Toolbox. The values used here for a_q, b_q, c_q, for $Q = 3$ are given in Table 5.2 (Further details can be found in [78]). The dereverberation algorithm used for reducing early reverberation is summarized in Table 5.3.

Table 5.3: The dereverberation method for early reverberation

Task: Use Wiener filtering approach to suppress the early reverberation.

Input: $\hat{S}_l(m,k)$ and $\hat{S}_r(m,k)$.

Output: $\overline{S}_l(m,k)$ and $\overline{S}_r(m,k)$.

Initialization: In (5.31) and (5.32), $\alpha_2 = 0.8$ is used.

Case: The goal is to estimate the spectral weights of the Wiener filter. The steps are:

1) Use (5.31) and (5.32) to estimate $\tilde{Y}_{\hat{s}_l\hat{s}_l}(m,k)$, $\tilde{Y}_{\hat{s}_r\hat{s}_r}(m,k)$, and $\tilde{Y}_{\hat{s}_l\hat{s}_r}(m,k)$.

2) Use (5.33) to obtain the sound field coherence, i.e., $coh_{\hat{s}_l\hat{s}_r}(f)$.

3) Use (5.28) to obtain $\tilde{Y}_{ss}(m,k)$.

4) Use (5.29) to estimate the spectral weights of the Wiener filter, i.e., $G_c(m,k)$.

Output: Compute $\overline{S}_l(m,k)$ and $\overline{S}_r(m,k)$ according to (5.30).

5.5 Experimental Results and Discussion

In this section, the performance of the proposed method is evaluated using the simulated RIRs from the image model [4] and the real RIRs from the acoustic impulse response (AIR) database [79]. Ten different anechoic speech signals from the TIMIT database, uttered by 5 males and 5 females all sampled at 16 KHz, are convolved with the RIRs to

generate the reverberant speech files. The size of the room used in the case of simulated RIRs is 10 x 10 x 10 (m^3). The Hanning window of 256 samples is used with an overlap factor set to 50%. The STFT length is 256. The rest of the parameters are set as: τ= 0.1, C= 2.5, N_{le}=13, R=128, ψ= 25, $\tilde{G}_{late}^{min}$= 2.22×10^{-16}, α_2 = 0.8, coh_{max} = 0.99, G_{min}^{c} = 0.3 . Performance indices used in the evaluations are the segmental signal to reverberation ratio (SegSRR) [88], and the signal to distortion ratio (SDR) [103]. SegSRR is defined as,

$$SegSRR(m) = 10log_{10}\frac{\sum_{n=mR}^{mR+N-1} s_d^2(n)}{\sum_{n=mR}^{mR+N-1}(s_d(n)-\hat{s}(n))^2} \tag{5.34}$$

where $s_d(n) = s(n) * h_d(n)$ represents the direct signal (delayed version of the clean signal), $h_d(n)$ is obtained from the known impulse response and $\hat{s}(n)$ is the enhanced speech signal. N and R are the number of samples per frame and frame rate in samples respectively. The mean SegSRR can be obtained by averaging (5.34) over the total frames. The SDR can be defined as [103],

$$SDR = 10log_{10}\frac{\sum_{n=0}^{L} s^2(n)}{\sum_{n=0}^{L}(s(n)-\hat{s}(n))^2} \tag{5.35}$$

where $s(n)$ and $\hat{s}(n)$ are the original signal and the enhanced signal respectively, and L is the length of the signal. Note that, $SegSRR$ and SDR are calculated in this work for $\overline{s}_l(n)$ and $\overline{s}_r(n)$ separately and then averaged.

For performance comparison the method in [78] (called for short Jeub *et al.* method hereafter) is used as the baseline which represents the state-of-the-art and uses the frequency-independent model for decay rate estimation.

First, a dereverberation example is presented here for the real data recorded in a booth and lecture room [79], where the T_{60} is approximately 400 ms and 900 ms respectively, and the source-microphone distance is 1 m and 2.25 m respectively. The spectrograms of the signals for the booth and lecture room are shown in Figure 5.1 and 5.2 respectively. For comparison 3 different regions are highlighted which are marked as A_i, B_i and C_i, where $i = 1$ is for the clean signal, $i = 2$ for the dereverberated signal by Jeub *et al.* method and $i = 3$ for the dereverberated signal from the proposed method. From the highlighted regions it can be observed that the signal obtained by the proposed method is closer to the clean one as compared to the Jeub *et al.* method in both the cases.

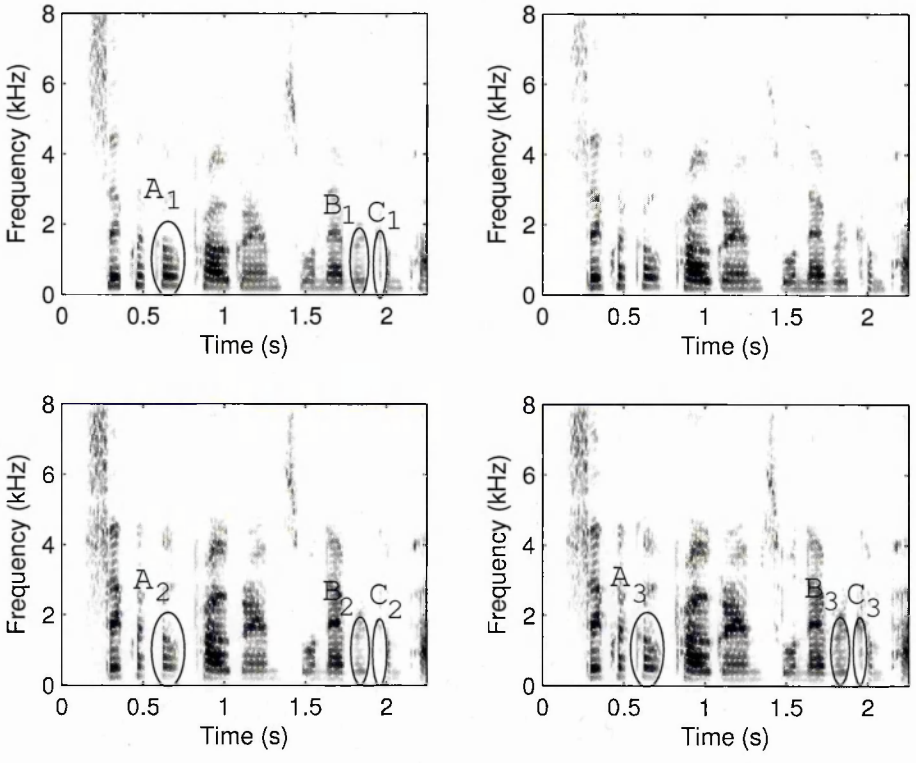


Figure 5.1: Comparison of the spectrograms of the clean signal (top left) with the enhanced signals obtained by the proposed method (bottom right) and the Jeub *et al.* method (bottom left) for the real data recorded in a booth. The top right plot shows the reverberant signal. The RIRs used to generate the reverberant signal were recorded from the booth room with source-microphone distance equal to 1 m.

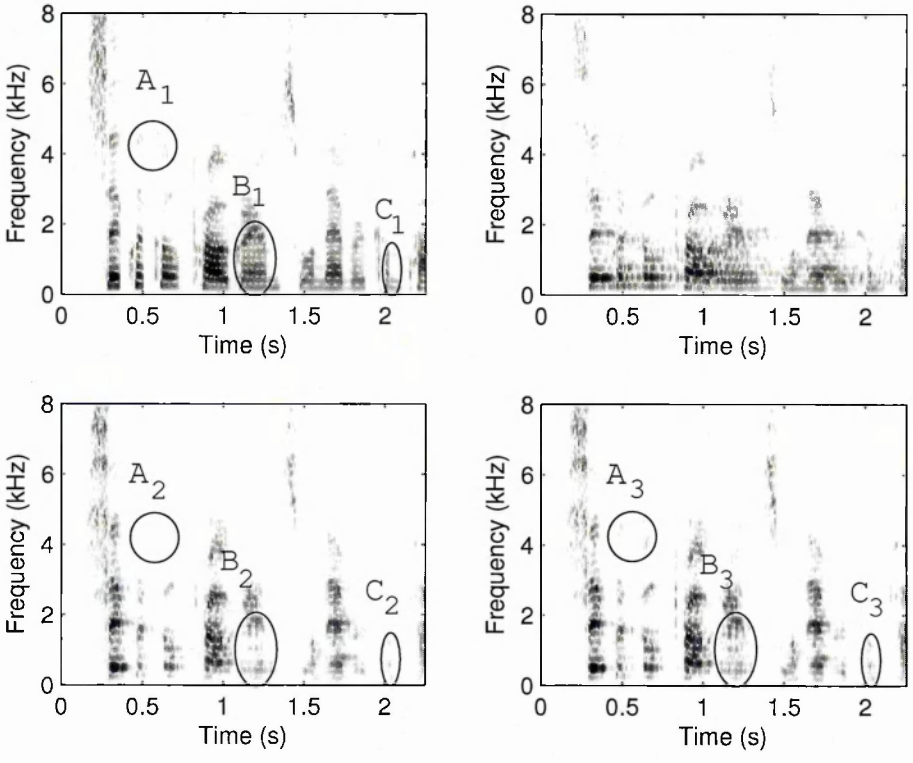


Figure 5.2: Comparison of the spectrograms of the clean signal (top left) with the enhanced signals obtained by the proposed method (bottom right) and the Jeub *et al.* method (bottom left) for the real data recorded in a lecture room. The top right plot shows the reverberant signal. The RIRs used to generate the reverberant signal were recorded from the lecture room with source-microphone distance equal to 2.25 m.

In a further experiment, the performance of the proposed method is evaluated in comparison to the Jeub *et al.* method using SDR and mean SegSRR. First the simulated RIRs are used to generate the reverberant signals from the anechoic speech signals at three different reverberation times, i.e., $T_{60} = \{300, 500, 600\}$ ms, and two different source-microphone distances, i.e., 0.5 and 2.5 m respectively. For each T_{60} and source-microphone distance, 5 different source-microphone positions and the 10 anechoic signals from the TIMIT database, resulting in 100 different reverberant signals for both left and right channel, were used for testing the algorithms. In total, 300 independent tests were run for the simulated data generating 600 different reverberant signals for both left and right channel. Figure 5.3 shows for each T_{60} and source-microphone distance the results (mean values $\pm$ standard deviations) averaged over the 100 tests. The results indicate that the proposed method gives consistently higher SDRs and SegSRRs than Jeub *et al.* method for various source-microphone distances and reverberation times.

In another set of experiments, the real binaural RIRs from the AIR database [79] are used which contain five different types of RIRs, recorded in five different room environments, namely booth, office, meeting, lecture, and stairway. For each room environment, a pair of source-microphone distances $\{D_1, D_2\}$ m, $\{0.5, 1.5\}$, $\{1, 3\}$, $\{1.45, 2.8\}$, $\{2.25, 7.1\}$, and $\{1, 3\}$ are selected respectively. The 10 anechoic signals from the TIMIT database are then convolved with each of these RIRs, resulting in 200 reverberant signals in total for both left and right channels. For each room type and source-microphone distance, the average results of SDR and SegSRR over the 10 different signals, are given in Figure 5.4. The proposed method performs significantly better than Jeub *et al.* method for shorter source-microphone distances. For example, for the booth and $D_1 = 0.5$ m, both SDR and SegSRR obtained by the proposed method are about 8 dB higher than those by Jeub *et al.* method. Such an improvement, observed for nearly all the testing cases, decreases when the source-microphone distance increases. Averaged over all the 200 tests, the SDR and SegSRR of the proposed method are respectively 1.82 dB and 1.90 dB higher than those of the Jeub *et al.* method. These results demonstrate the advantage of the frequency dependent model in particular for shorter source-microphone distances. Note that the output SDR and

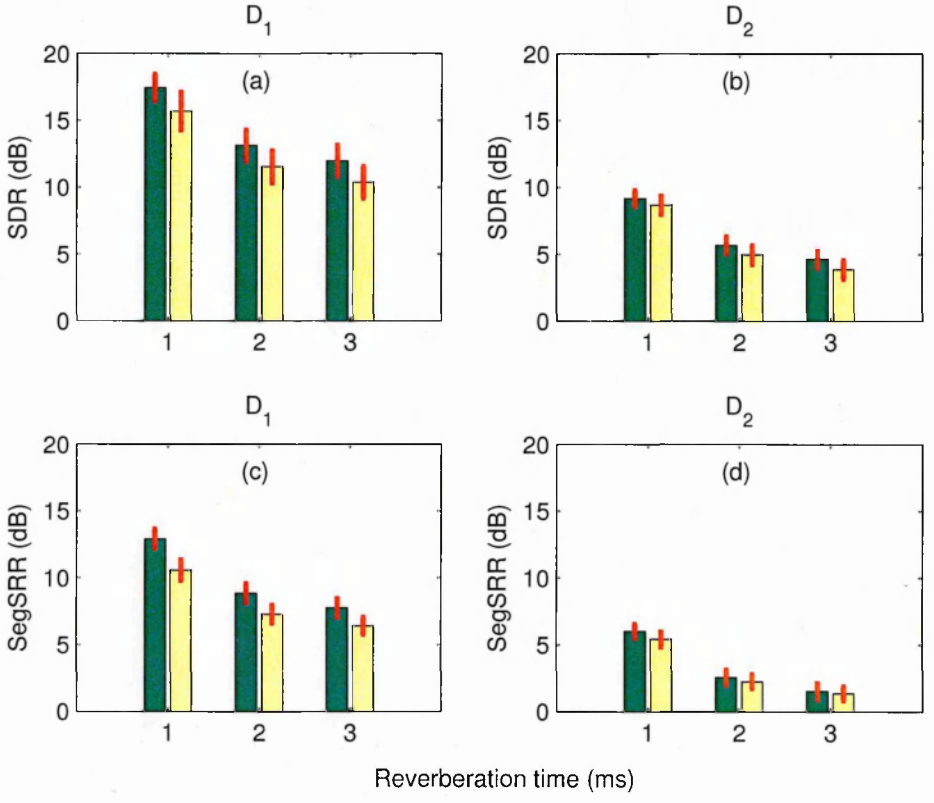


Figure 5.3: SDR and SegSRR of the proposed method (green bars) and Jeub *et al.* method (yellow bars) for the simulated data. The labels on the horizontal axis represent different reverberation times, namely, 1 - 300 ms, 2 - 500 ms, 3 - 600 ms. For each of the reverberation times, two different source-microphone distances were tested, respectively $D_1 = \{0.5\}$ m and $D_2 = \{2.5\}$ m. The standard deviations are also plotted as short lines on top of the bars.

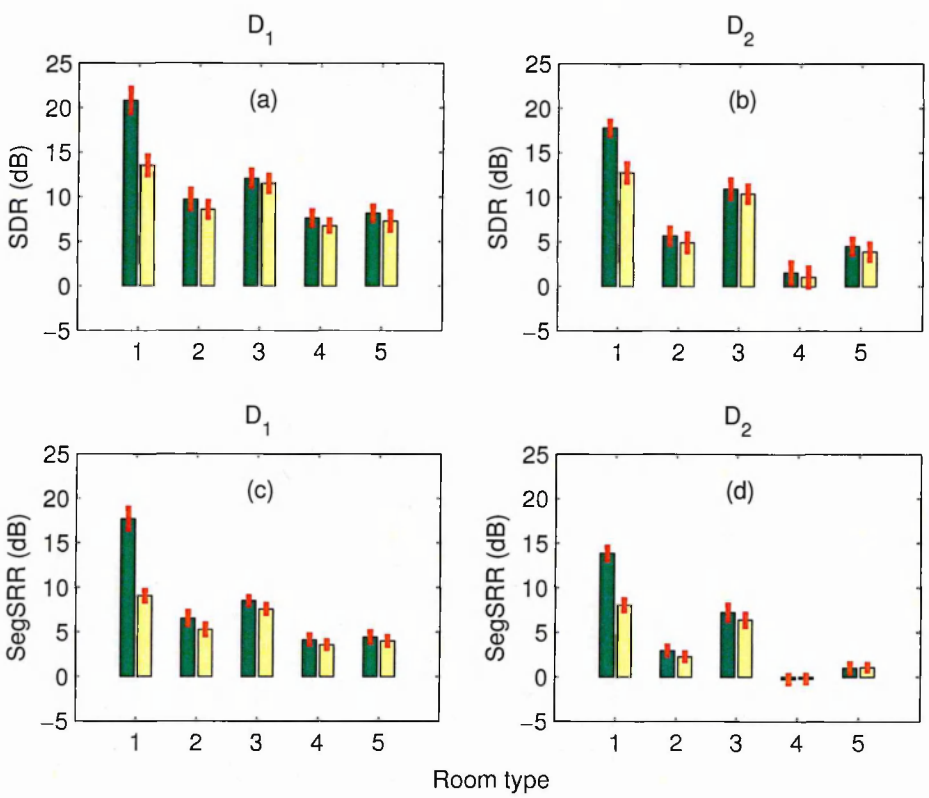


Figure 5.4: SDR and SegSRR for the AIR database of the proposed method (green bars) and Jeub *et al.* method (yellow bars). The labels on the horizontal axis represent different room types, namely, 1 - booth, 2 - office, 3 - meeting, 4 - lecture, 5 - stairway. For each of the five rooms, two different source-microphone distances were tested, respectively $D_1 = \{0.5, 1, 1.45, 2.25, 1\}$ m and $D_2 = \{1.5, 3, 2.8, 7.1, 3\}$ m. The standard deviations are also plotted as short lines on top of the bars.

SegSRR are reported here in the results. It has been observed in the experiments that ΔSDR and ΔSegSRR for the proposed method is higher when direct to reverberation ratio is negative (i.e., for higher source-microphone distances) in comparison to when direct to reverberation ratio is positive (i.e., for shorter source-microphone distances). Also the proposed method is giving improvement over the Jeub *et al.* method in terms of ΔSDR and ΔSegSRR for both positive and negative direct to reverberation ratio.

5.6 Summary

In this chapter a dereverberation algorithm based on a frequency dependent statistical model of the reverberation time has been proposed. The algorithm is composed of the estimation of the decay rate of the late reverberations based on this model, the estimation of the mask containing spectral subtraction gains, the smoothing of the spectral mask by a frequency dependent filter, followed by Wiener filtering for reducing early reflections. It has been shown that the proposed algorithm offers considerably higher dereverberation performance as compared with a related recent approach using the frequency independent model. However, the frequency dependent reverberation time and decay rate required in the proposed model are estimated from the RIRs, which can be limited in practical applications, where RIRs may not be available. To this end, the next chapter further addresses this problem and proposes a method that can directly estimate them from reverberant speech signals.

Chapter 6

Blind Estimation of Reverberation Time For Blind Dereverberation and Separation of Speech Mixtures

In previous chapters source separation and dereverberation issues have been analysed separately. This chapter proposes a method for performing blind dereverberation (BD) and blind source separation together for the speech mixtures. It is common that the performance of the speech separation algorithms deteriorates with the increase of room reverberations. Therefore in this chapter the dereverberation algorithm developed in Chapter 5 is combined with the separation method presented in Chapter 3 to mitigate the effects of room reverberations on the mixtures and hence to improve the separation performance. The dereverberation algorithm presented in Chapter 5 assumes that the RIRs are known as a *priori*, which however are not directly accessible from the speech mixtures in practice. To address this problem, a method consisting of a step for blind estimation of reverberation time (RT) is proposed to estimate the decay rate (i.e., $\alpha(k)$ in Equation (5.9)) of reverberation directly from the reverberant speech signal (i.e., mixtures). Based on the analysis of an existing RT estimation method,

which models the reverberation decay as a Gaussian random process modulated by a deterministic envelope, a Laplacian distribution based decay model is proposed in which an efficient procedure for locating free decay from reverberant speech is also incorporated. Hence the developed algorithm works in a blind manner, i.e., directly dealing with the reverberant speech signals without the information from the RIRs. Evaluation results in terms of SDR and SegSRR reported in this chapter reveal that using this method the performance of the separation algorithm developed in Chapter 3 can be further enhanced.

6.1 Introduction

The speech signals captured by the microphone in a closed environment are often reverberated and also contaminated by the intereferences from the nearby sound sources. The separation of the target speech from the microphone signal is a challenging task because of the interfering speech signals, and the presence of reverberation makes it more challenging. Therefore, it is very important to devise a method which can separate the target speech from the interfering ones and can also reduce the adverse acoustic disturbances.

In Chapter 3, a source separation algorithm has been developed, however its performance deteriorates in the presence of room reverberations. Therefore, in Chapter 5 of this thesis a dereverberation algorithm has been developed to suppress the room reverberation, and here this dereverberation algorithm is combined with the separation algorithm developed in Chapter 3 to enhance the separation performance. However the dereverberation algorithm developed in Chapter 5 assumes the RIRs to be known *a priori*, which in reality are not available. To address this problem, a method is proposed in this chapter for the blind estimation of RT and then incorporated with the algorithm developed in Chapter 5. The proposed blind RT estimation method uses the reverberant speech (i.e., mixture) directly to estimate the decay rate instead of the RIRs as done in Chapter 5. In the proposed method, a Laplacian distribution based decay model for room reverberation is used along with an efficient procedure for locating the free decay in reverberant speech. Finally, the proposed RT estimation method

is incorporated with the algorithms developed in Chapters 3 and 5 to obtain a joint blind dereverberation and separation algorithm for the speech mixtures.

The developed joint algorithm which is a two channel method has been employed in three different ways. Firstly, the available mixture signals are used to estimate blindly the RT based on a maximum-likelihood (ML) method and statistical modelling of the sound decay rate of the reverberant speech, followed by the dereverberation of the mixture signals using the method based on the frequency depenedent statistical model as described in Chapter 5. Then the separation algorithm proposed in Chapter 3 is applied to these resultant mixtures so that the source speech signals can be obtained. Secondly, the separation algorithm is applied first to the mixtures to segregate the speech signals, followed by the blind estimation of RT from the separated speech signal. Then dereverberation is employed to the segregated speech signals. In the third scheme, the multistage separation algorithm proposed in Chapter 3 is split such that the convolutive ICA is first applied to the mixtures to obtain the estimated source signals. Then, the signal obtained from the convolutive ICA is used to estimate the RT followed by the blind dereverberation of the signals obtained from convolutive ICA. Then the T-F representation of dereverberated signals are used to estimate the IBM followed by cepstral smoothing to enhance the separated speech signals.

The rest of the chapter is organized as follows. Section 6.2 presents the proposed and related method for blind estimation of RT from the reverberant speech signal. In Section 6.3, the proposed blind dereverberation method will be described and evaluated. Section 6.4 evaluates the performance of the proposed joint blind dereverberation and separation algorithm and reports the experimental results followed by a conclusion in Section 6.5.

6.2 Blind Reverberation Time Estimation

The concept of measuring RT was coined for the first time by Sabine in 1922 [144]. Robust estimation of RT directly from the reverberant signal is a challenging task. In this work a method is proposed to estimate RT directly from the reverberant signal, which is based on the ML estimation of the unknown sound decay rate modelled by

a Laplace distribution. Before describing the proposed method, a brief overview is provided for the RT and its measurements.

6.2.1 Theory and background

Estimation of RT has been investigated for a long time. The RT of an enclosed environment is defined as the time for which a sound prevails after it has been turned off, due to its multiple reflections from the different surfaces within the enclosed environment. The RT is usually referred to as the time for the sound level to drop to 60 dB below its original value [137], [138], [144]. Reverberation leads to speech distortion both in terms of its envelop and fine structure, therefore RT is an important parameter that measures the listening quality of the enclosed environment, i.e., room. The effect of reverberation is most perceptible when speech recorded by microphones is played back via headphones. The distortions previously unseen in the sound pattern are now clearly noticed even by normal listeners, pointing the extraordinary echo suppression and dereverberation capabilities of the normal auditory system when the ears receive sounds directly [66]. For hearing impaired listeners, the reception of a reverberant signal via the microphone of a hearing aid intensify the problem of listening in challenging environments.

Although dereverberation is an active area of investigation, state-of-the-art hearing aids or other audio processing instruments, apply signal processing strategies complying to specific listening environments. These instruments are anticipated to have the ability to assess the characteristics of the environment, and to trigger the most suitable signal processing strategy. Hence a method that can characterize the RT of a room from passively received microphone signals represents an important area of research.

In the early days of 20th century, Sabine [144] implemented an empirical formula for the calculation of RT based entirely on the geometry of the environment (i.e., volume and surface area) and the absorption attributes of its surfaces. Later on, Sabine's RT equation has been greatly modified and its accuracy improved (refer to [89] for the details of the modifications), and thats why currently it has been used in numerous commercial software packages for the acoustic design of interiors, anechoic chamber

measurements, design of concert halls, classrooms, and other acoustic environments where the quality of the received sound is of high importance and magnitude of reverberations must be controlled. However, such methods require that the room geometry and absorptive characteristics of the room be determined first. When these can not be determined easily, it is important then to find some method which is based on the test sound signal radiated in the enclosed environment.

Methods using the test sound signal for measuring RT are based on sound decay curves. In the interrupted noise method [75], a burst of noise having broad or narrow band is radiated into the test room. In the time instant where the sound field attains the steady state, the noise source is switched off and the decay curve is obtained. The slope of the decay curve is used to estimate the RT. As the noise source signal has fluctuations, the decay curve obtained will differ from trial to trial. Hence to estimate the reliable RT averaging must be applied to the large number of obtained decay curves. In order to overcome this issue, Schroeder developed an integrated impulse response method in 1965 [153] in which the excitation signal is a pulse either broad band or narrow band. The enclosure (room) output for a pulse is simply the impulse response of the room in the specified frequency band. Schroeder proved that the impulse response of the room is related via a certain integral to the overall average of the decay curve obtained using the interrupted noise method, and hence the repeated trials were inessential. Both the methods require controlling environment for the experiment, specifically a suitable excitation signal must be available *a priori*.

While Schroeder's method has been used immensely over the past few decades for the estimation of RT, and has been improved over the years (see for example, [31, 183]), there is a need of some *blind* method that can estimate room RT from the available microphone signals, i.e., without any information about the room geometry and absorption attributes, or when the test sound signal is not available. Such blind method which works with speech sound directly will be very useful for incorporating in hearing aids or hands free telephony devices. Some partial blind methods have also been developed in which the room characteristics are learned using neural network approaches [36, 113, 162], or some sort of segmentation procedure is used for detecting gaps in the sounds so that the sound decay curve can be tracked [94]. Several meth-

ods have been developed recently that can estimate RT *blindly*, i.e., directly from the recorded reverberant signals [99,100,137,138]. These methods are based on the statistical modelling of the sound decay such that the ML estimator can be used to determine the RT.

Ratnam *et al.* [138] developed an algorithm for the blind estimation of RT based entirely on the available recorded sound. The estimator is based on a noise decay curve model explaining the reverberation characteristics of the enclosure. Sounds in the test environment are processed such that a running estimate of RT is achieved by the system employing the ML parameter estimation procedure. A decision making step is then applied to collect the estimate of RT over a period of time and attains the most likely RT using an order statistics filter. However detecting the correct sound decay from a reverberant speech signal is a challenging problem and a method in [138] used an iterative approach for that purpose, which makes the algorithm computationally expensive. Later on Ratnam *et al.* presented another algorithm in [137] based on their original model in [138] in order to improve the computational efficiency of the original method. Very recently Lollmann *et al.* [100] presented an algorithm for the blind estimation of RT from reverberant speech signals. The method is using a statistical model for the sound decay based on the sound decay model developed in [138], followed by the ML estimation approach to estimate the decay rate presented in [137]. However, the method of Lollmann *et al.* is employing a pre-selection mechanism to detect the possible sound decay which makes it robust and computationally efficient. The method presented in this chapter for the blind estimation of RT is based on Lollmann *et al.* method. Therefore, the next subsections will describe in detail the sound decay model and ML estimation approach presented in Ratnam *et al.* method, the pre-selection mechanism to detect the possible sound decay presented in Lollmann *et al.* method, and our proposed method based on using Laplace distribution for modelling the decay rate.

6.2.2 Sound decay model and ML estimation

The sound decay model used by the Lollmann *et al.* method [100] is based on the original model presented in [138]. The model is based on the assumption that the reverberation tail of a decaying sound denoted here as y is the product of a fine structure denoted as x that is a random process, and an envelop a that is deterministic. Suppose $x(n)$ is a random sequence for $n \geq 0$, of independent and identically distributed (i.i.d.) random variables having normal distribution with zero mean and variance σ, $\mathcal{N}(0, \sigma)$. Similarly for each n a deterministic sequence is defined as $a(n) > 0$. As a result the model is obtained for the room decay in which the observations y are represented as $y(n) = a(n)x(n)$. As $a(n)$ is a time varying term, therefore $y(n)$ are independent but not identically distributed, and hence their probability density function is $\mathcal{N}(0, \sigma a(n))$.

In order to estimate the decay rate, consider a finite sequence of observations, $n = 0, ..., N-1$. For notational convenience, N-dimensional vectors of y and a are denoted as $\mathbf{y}$ and $\mathbf{a}$ respectively. Hence the likelihood function of $\mathbf{y}$ (the joint probability density), parameterized by $\mathbf{a}$ and σ, is [138]

$$L(\mathbf{y}; \mathbf{a}, \sigma) = \frac{1}{a(0) \cdots a(N-1)} \left(\frac{1}{2\pi\sigma^2} \right)^{N/2} \times exp\left(- \frac{\sum_{n=0}^{N-1} (y(n)/a(n))^2}{2\sigma^2} \right) \quad (6.1)$$

where $\mathbf{a}$ and σ are the $(N+1)$ unknown parameters that are required to be estimated from the observation $\mathbf{y}$. As the main goal here is to model the sound decay in a room and the likelihood function obtained in Equation (6.1) can be further simplified. Suppose a single decay rate ρ_2 define the damping of the sound envelop during the regions of free decay (i.e., the period following the sharp offset of a speech sound) instead of those regions where sound is actually ongoing, onset, or gradually declining speech offsets. As a result the sequence $a(n)$ is determined by

$$a(n) = exp(-n/\rho_2) \quad (6.2)$$

Hence, the N-dimensional parameter $a(n)$ can be replaced by a single scalar parameter a which is denoted by ρ_2 as

$$a = exp(-1/\rho_2) \quad (6.3)$$

As a result Equation (6.2) can be written as

$$a(n) = a^n \quad (6.4)$$

Now Equation (6.1), after incorporating Equation (6.4) becomes

$$L(\mathbf{y};a,\sigma)=\left(\frac{1}{2\pi a^{(N-1)}\sigma^2}\right)^{N/2}\times exp\left(-\frac{\sum_{n=0}^{N-1}a^{-2n}y(n)^2}{2\sigma^2}\right) \tag{6.5}$$

ML approach is then used to estimate the parameters a and σ [131,138]. Firstly, the logarithm of Equation (6.5) is taken to obtain the log-likelihood function

$$lnL(\mathbf{y};a,\sigma)=-\frac{N(N-1)}{2}ln(a)-\frac{N}{2}ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}a^{-2n}y(n)^2 \tag{6.6}$$

To find the maximum of $ln(L)$, differentiate the log-likelihood function in Equation (6.6) with respect to a to obtain the score function SF_a [131]

$$SF_a(a;\mathbf{y},\sigma)=\frac{\partial lnL(\mathbf{y};a,\sigma)}{\partial a}=-\frac{N(N-1)}{2a}+\frac{1}{a\sigma^2}\sum_{n=0}^{N-1}na^{-2n}y(n)^2 \tag{6.7}$$

Let $\partial lnL(\mathbf{y};a,\sigma)/\partial a=0$, then the log-likelihood function achieves the extremum, given as [138]

$$-\frac{N(N-1)}{2a}+\frac{1}{a\sigma^2}\sum_{n=0}^{N-1}na^{-2n}y(n)^2=0 \tag{6.8}$$

The zero of the score function achieves the best estimate in the sense that $E[SF_a]=0$, which is denoted by $\hat{a}^{(ML)}$. It can be demonstrated that the second derivative $\partial^2 lnL(\mathbf{y};a,\sigma)/\partial a^2\mid_{a=\hat{a}^{(ML)}}<0$, i.e., the estimate $\hat{a}^{(ML)}$ maximizes the log-likelihood function.

Similarly, the variance σ^2 can be estimated by differentiating the log-likelihood function in Equation (6.6) with respect to σ,

$$SF_\sigma(\sigma;\mathbf{y},a)=\frac{\partial lnL(\mathbf{y};a,\sigma)}{\partial\sigma}=-\frac{N}{\sigma}+\frac{1}{\sigma^3}\sum_{n=0}^{N-1}a^{-2n}y(n)^2 \tag{6.9}$$

Now again the log-likelihood function achieves the extremum when $\partial lnL(\mathbf{y};a,\sigma)/\partial\sigma=0$, which results in

$$\sigma^2=\frac{1}{N}\sum_{n=0}^{N-1}a^{-2n}y(n)^2 \tag{6.10}$$

As done above, it can be also shown here that $E[SF_\sigma]=0$, which leads to the optimal estimate of the variance, denoted by $\hat{\sigma^2}^{(ML)}$. It can be shown that the second derivative $\partial^2 lnL(\mathbf{y};a,\sigma)/\partial\sigma^2\mid_{\sigma=\hat{\sigma}^{(ML)}}<0$, i.e., the estimate $\hat{\sigma^2}^{(ML)}$ maximizes the log-likelihood

function. Note that (6.8) is an implicit expression for a and hence a can not be solved directly, whereas (6.10) provides the ML estimate of σ directly if a is known. Now if the solution for σ^2 in Equation (6.10) is substituted into Equation (6.6), the log-likelihood function can be rewritten as [100, 138]

$$lnL(a;\mathbf{y}) = -\frac{N}{2}\left((N-1)\,ln(a) + ln\left(\frac{2\pi}{N}\sum_{n=0}^{N-1} a^{-2n}y(n)^2\right) + 1\right) \tag{6.11}$$

Therefore, Equation (6.11) is used to find the estimate of a, i.e., $\hat{a}^{(ML)}$. The approach proposed in [137] is implemented by quantizing the range of a. As in Equation (6.3) defined already, ρ_2 is a time constant to be estimated. It is noted that $a \in [0, 1)$ maps one-to-one onto $\rho_2 \in [0, \infty)$. Now the given range of a is quantized such that the bins of the histogram of a are formed. Then the likelihood values are calculated, and the highest likelihood is assigned to that bin in the histogram.

Let the range of $a \in [0, 1)$ be quantized into Q values, so that a_j is obtained with $j = 1, ..., Q$. Then, for each a_j, the log-likelihood function given by Equation (6.11) can be written as

$$lnL(a_j;\mathbf{y}) = -\frac{N}{2}\left((N-1)\,ln(a_j) + ln\left(\frac{2\pi}{N}\sum_{n=0}^{N-1} a_j^{-2n}y(n)^2\right) + 1\right) \tag{6.12}$$

The best estimate of a, i.e., $\hat{a}^{(ML)}$ is selected as

$$\hat{a}^{(ML)} = \max_{a}\{lnL(a_j;\mathbf{y})\} \tag{6.13}$$

Then Equation (6.3) is used to obtain the estimate of the decay rate $\hat{\rho_2}^{ML}$, followed by the calculation of the RT value, i.e., $\hat{T}_{60}^{(ML)}$ using the following formula [138]

$$T_{60} = 6.908 \times \rho_2 \tag{6.14}$$

6.2.3 Effective RT estimation

As the original method presented in [138] used an iterative approach to estimate the sound decay rate which makes the algorithm computationally very demanding. The method presented in [137] improves the computational efficiency of the original method, however it considers the whole recorded reverberant speech signal during the process

of ML estimation of the sound decay rate instead of using only the free sound decay regions. Hence there is a need for some method which can capture the free sound decay regions first in the reverberant speech signal so that only the detected sound decay regions can be used for ML estimation of decay rate. Therefore, Lollmann *et al.* devised an efficient estimation procedure which can capture correctly the regions of free decay in the reverberant speech first, and then used such detected regions only for the ML estimation of decay rate, which improves the computational efficiency of the algorithm as well as reduces the effects of the outliers on the estimated RT value. The sequence of the reverberant signal defined in Chapter 5 (Equation(5.1)) is processed within the frames of B samples shifted by instants of ΔB samples [100], given as

$$Y(\lambda, b) = y(\lambda \Delta B + b) \quad with \;\; b = 0, 1, ..., B-1 \tag{6.15}$$

where $\lambda \in \mathbb{N}$. In the first step, pre-selection is carried out to detect the possible sound decays. In order to achieve this, the current frame $Y(\lambda, b)$ is divided into $L = B/P \in \mathbb{N}$ sub frames

$$V(\lambda, l_{sub}, k_{sub}) = Y(\lambda, l_{sub}P + k_{sub}) \tag{6.16}$$

where $k_{sub} = 0, 1, ..., P-1$ and sub-frame index $l_{sub} = 0, 1, ..., L-1$. Now it is examined whether the maximum energy and minimum energy values of a sub-frame deviates from the succeeding sub-frames according to [100]

$$\sum_{k_{sub}=0}^{P-1} V^2(\lambda, l_{sub}, k_{sub}) > \tau_{l_{sub}} \cdot \sum_{k_{sub}=0}^{P-1} V^2(\lambda, l_{sub}+1, k_{sub}) \tag{6.17a}$$

$$\max_{k_{sub}}\{V(\lambda, l_{sub}, k_{sub})\} > \tau_{l_{sub}} \cdot \max_{k_{sub}}\{V(\lambda, l_{sub}+1, k_{sub})\} \tag{6.17b}$$

$$\min_{k_{sub}}\{V(\lambda, l_{sub}, k_{sub})\} < \tau_{l_{sub}} \cdot \min_{k_{sub}}\{V(\lambda, l_{sub}+1, k_{sub})\} \tag{6.17c}$$

where $0 \leq \tau_{l_{sub}} \leq 1$ is a weighting factor. If one of these conditions is violated, it is examined whether the counter l_{sub} has reached a minimum value $1 < l_{submin} < L-2$. If this is not the case, the comparison is terminated and the next signal frame $Y(\lambda+1, b)$ is processed. Otherwise, the sequence of sub-frames for which Equation (6.17) applies is detected as a possible sound decay. For this detected frame, the RT, i.e., $\hat{T}_{60}^{(ML)}$ is calculated using Equations (6.12), (6.13), (6.3), and (6.14) for a finite set of RT values (decay rates).

A new ML estimate is used now in which a histogram with a bin size 10 is generated and contains the estimated RT values obtained above (i.e., $\hat{T}_{60}^{(ML)}$), and updated each time when another RT value (i.e., $\hat{T}_{60}^{(ML)}$) is obtained. The current RT estimate denoted here as $\hat{T}_{60}^{(1)}$ is associated with the maximum of this histogram (The maximum instead of the first peak can be taken as this histogram contains no significant number of outliers due to the pre-selection). The variance for the estimated RT is reduced by a recursive smoothing such that the final estimate is given by

$$\hat{T}_{60}(\lambda) = \alpha \cdot \hat{T}_{60}(\lambda - 1) + (1 - \alpha) \cdot \hat{T}_{60}^{(1)}(\lambda) \tag{6.18}$$

where $0.9 < \alpha < 1$. The final RT value is estimated by

$$\hat{T}_{60} = mean(\hat{T}_{60}(\lambda)) \tag{6.19}$$

6.2.4 Proposed method

In this section a new method is proposed for RT estimation based on the Laplacian distribution. The method is motivated from the findings in [130], where it has been shown that the amplitude distribution of the reverberant speech is better modeled by Laplace distribution. Therefore, the reverberant tail of a decaying sound is modeled using a sequence of random variables with Laplace distribution $\mathcal{L}(\theta, \varrho)$, where θ is the mean considered as zero here and ϱ is the variance of the Laplace distribution. Consider again the random sequence as $x(n)$ for $n \geq 0$ of i.i.d. random variables having laplace distribution $\mathcal{L}(0, \varrho)$. Based on the model described above in Section 6.2.2 for the observations $y(n)$, a new model is proposed in this work for the observations $y(n)$ whose probability density function is $\mathcal{L}(0, \varrho a(n))$.

In order to estimate the decay rate, consider again a finite sequence of observations, $n = 0, ..., N-1$. Hence the likelihood function of N-dimensional vector of y, i.e., $\mathbf{y}$ (the joint probability density), parameterized by N-dimensional vector of a, i.e., $\mathbf{a}$ and ϱ, is [87]

$$L(\mathbf{y}; \mathbf{a}, \varrho) = \frac{1}{a(0) \cdots a(N-1)} \left(\frac{1}{2\varrho}\right)^N \times exp\left(-\frac{\sum_{n=0}^{N-1} \mid y(n)/a(n) \mid}{\varrho}\right) \tag{6.20}$$

where $\mathbf{a}$ and ϱ are the $(N+1)$ unknown parameters that are required to be estimated from the observation $\mathbf{y}$. Based on Equation (6.4) for the sequence $a(n)$, Equation (6.20) can be written as

$$L(\mathbf{y};a,\varrho)=\left(\frac{1}{2a^{(N-1)/2}\varrho}\right)^{N}\times exp\left(-\frac{\sum_{n=0}^{N-1}\mid a^{-n}y(n)\mid}{\varrho}\right) \tag{6.21}$$

ML approach is then used to estimate the parameters a and ϱ. Firstly, the logarithm of Equation (6.21) is taken to obtain the log-likelihood function

$$lnL(\mathbf{y};a,\varrho)=-Nln(2)-\sum_{n=0}^{N-1}ln(a^{n}\cdot\varrho)-\frac{1}{\varrho}\sum_{n=0}^{N-1}a^{-n}\mid y(n)\mid \tag{6.22}$$

To get the maximum of $ln(L)$, differentiate the log-likelihood function in Equation (6.22) with respect to a to achieve the score function SF_a [131]

$$SF_a(a;\mathbf{y},\varrho)=\frac{\partial lnL(\mathbf{y};a,\varrho)}{\partial a}=-\frac{1}{a}\sum_{n=0}^{N-1}n-\sum_{n=0}^{N-1}n\mid y(n)\mid a^{n-1} \tag{6.23}$$

Let $\partial lnL(\mathbf{y};a,\varrho)/\partial a=0$, then the log-likelihood function attains the extremum, as given

$$-\frac{1}{a}\sum_{n=0}^{N-1}n-\sum_{n=0}^{N-1}n\mid y(n)\mid a^{n-1}=0 \tag{6.24}$$

Denote the zero of the score function SF_a, and satisfying Equation (6.24), by $\hat{a}^{(ML)}$. It can be verified that the second derivative $\partial^2 lnL(\mathbf{y};a,\varrho)/\partial a^2\mid_{a=\hat{a}^{(ML)}}<0$, i.e., the estimate $\hat{a}^{(ML)}$ maximizes the log-likelihood function.

Similarly differentiate the log-likelihood function in Equation (6.22) with respect to ϱ,

$$SF_\varrho(\varrho;\mathbf{y},a)=\frac{\partial lnL(\mathbf{y};a,\varrho)}{\partial\varrho}=-\frac{N}{\varrho}+\frac{1}{\varrho^2}\sum_{n=0}^{N-1}a^{-n}\mid y(n)\mid \tag{6.25}$$

When $\partial lnL(\mathbf{y};a,\varrho)/\partial\varrho=0$, the log-likelihood function achieves the extremum, which results in

$$\varrho=\frac{1}{N}\sum_{n=0}^{N-1}a^{-n}\mid y(n)\mid \tag{6.26}$$

Using the score function SF_ϱ, the log-likelihood function can be maximized for ϱ also in the same way as done above by taking the second derivative.

It can be observed that Equation (6.24) is an implicit expression and a can not be solved explicitly, while Equation (6.26) provides the explicit estimate of ϱ if a is known.

Table 6.1: The proposed blind RT estimation method

Task: Use Laplacian distribution based energy decay model for the estimation of RT.
Input: Reverberant speech, i.e., $x(n)$.
Output: Estimated RT, i.e., $\hat{T}_{60}$.
Initialization: 1) In (6.15), $B = 1631$ and $\Delta B = 67$ are used.
2) In (6.16), $P = 233$ is used.
3) In (6.18), $\alpha = 0.995$ is used.
4) In (6.27) and (6.28), $j = 1, ..., Q$ while $Q = 10$ is used.
Case: The goal is to estimate the RT from reverberant speech signal. The steps are:
1) Use (6.15)-(6.17) to detect the free decay regions indexed by frame number λ.
2) For the detected regions, use (6.27), (6.28), (6.3), and (6.14) to obtain $\hat{T}_{60}^{(ML)}(\lambda)$.
3) Apply recursive smoothing via (6.18) to the estimated RT values, i.e., $\hat{T}_{60}^{(ML)}(\lambda)$.
Output: Compute $\hat{T}_{60}$ according to (6.19).

Based on the derivation pattern of Equation (6.12) from (6.11), a log-likelihood function used here in Equation (6.22) can be re-written as to select the best estimate of a, (i.e., $\hat{a}^{(ML)}$), given as

$$lnL(a_j; \mathbf{y}) = -Nln(2) - \sum_{n=0}^{N-1} ln(a_j^n \cdot \varrho) - \frac{1}{\varrho} \sum_{n=0}^{N-1} a_j^{-n} \mid y(n) \mid \tag{6.27}$$

Now $\hat{a}^{(ML)}$ can be selected as

$$\hat{a}^{(ML)} = \max_{a} \{lnL(a_j; \mathbf{y})\} \tag{6.28}$$

Now the estimate of the decay rate $\hat{\rho}_2{}^{ML}$ is obtained using Equation (6.3). Finally the RT value, i.e., $\hat{T}_{60}^{(ML)}$ is estimated using the formula in Equation (6.14). The effective RT estimation procedure described in Section 6.2.3 is applied then to obtain the final estimated single RT value for the reverberant speech signal. The proposed blind RT estimation algorithm using the Laplacian distribution based energy decay model is summarized in Table 6.1.

6.2.5 Simulation example

The performance of the proposed method for blind estimation of RT shall be illustrated by some simulation examples. To this end, similar to the experiments performed in Chapter 5, 10 different anechoic speech signals randomly selected from the TIMIT

database, uttered by 5 males and 5 females all sampled at 16 KHz, are convolved with the real RIRs from the AIR database [79] to generate the different reverberant speech files. The employed RIRs were recorded in four different room environments, namely booth, office, meeting, and lecture (Note that the stairway case is not considered from the AIR database in this example, as the mean RT values for the stairway are not reported in the original paper that describes the AIR database [79]). For each room environment, a pair of source-microphone distances $\{D_1,D_2\}$ m respectively, are selected, i.e., {0.5, 1.5}, {1, 3}, {1.45, 2.8}, and {2.25, 7.1}. The rest of the parameters used are given as : $Q = 10$, $L = 7$, $l_{submin} = 3$, $\alpha = 0.995$, $B = 1631$ (corresponds approximately to a time span of 0.10 s), $P = 233$, $\Delta B = 67$ (corresponds approximately to a frame shift of 0.0042 s), $\tau_{l_{sub}} = 1$.

For each room environment and each source-microphone distance, 10 different reverberant speech signals have been generated and then tested for the RT estimation. For each room type and source-microphone distance, the average results of estimated RT over the 10 different signals, are given in Figures 6.1 and 6.2 respectively, where RT estimated directly from RIRs based on Schroeder's method [153] and mean RT reported in [79] are also plotted for comparison purpose. For estimated RT based on Schroeder's method, the recorded RIRs in four different rooms for distances D_1 and D_2 have been used to estimate the RT value. On the other hand, the actual RT values are obtained from the results reported in [79], which are calculated for each room by taking the average of the RT values obtained over all measured positions of source-microphone in the room (further details can be found in [79]). The standard deviations are also plotted as short lines on top of the different color bars symbolizing the different methods.

Note that the results shown in Figure 6.1 are obtained for the shorter source-microphone distances from the above used pairs, i.e., D_1, while the results in Figure 6.2 are obtained for the longer source-microphone distances from the pairs, i.e., D_2. It can be observed that the difference between the estimated RT obtained using the proposed method and the actual RT (shown by red bars) is small in different room environments. For example, for the office room at D_1, the RT value obtained by the proposed method is 0.43 seconds and the actual RT value is 0.37 seconds, and similarly for the office room at D_2, the RT value estimated by the proposed method is 0.46 seconds and the actual

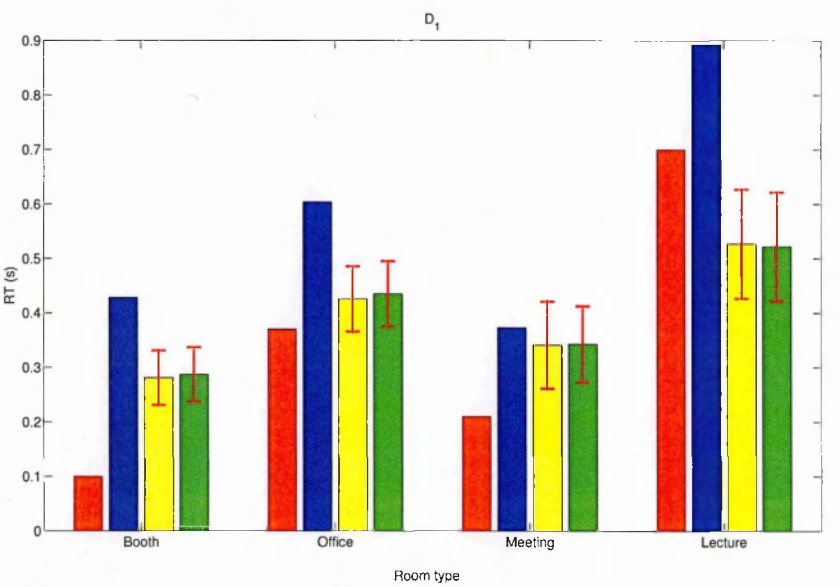


Figure 6.1: Performance measurement of different RT estimation methods in terms of accuracy obtained for different room environments from the AIR database. The mean RT is shown by red bars, the RT estimated from the RIRs by Schroeder's method is shown by blue bars, RT estimated by the Lollmann *et al.* method is shown by yellow bars, and RT estimated by the proposed method is shown by green bars. The distances between source and microphone for all of the four rooms are D_1={0.5, 1.0, 1.45, 2.25} m respectively. The standard deviations are also plotted as short lines on top of the yellow and green bars.

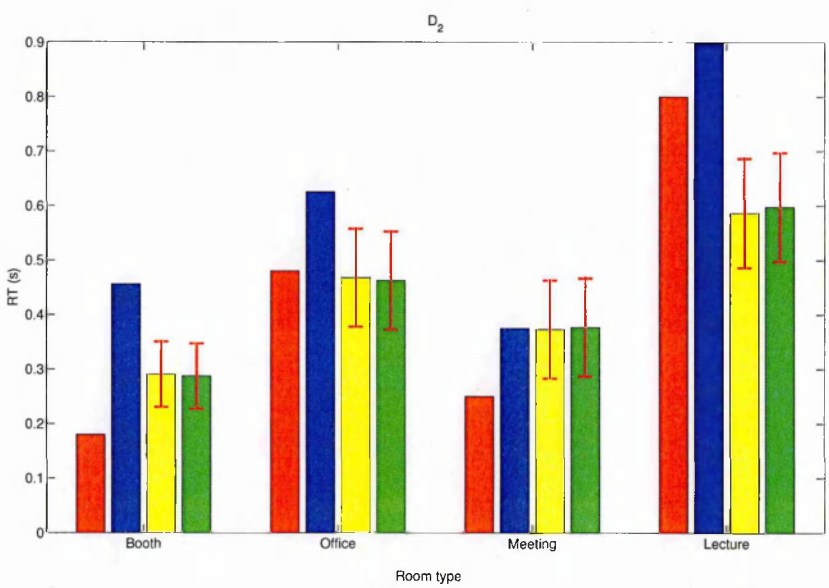


Figure 6.2: Performance measurement of different RT estimation methods in terms of accuracy obtained for different room environments from the AIR database. The mean RT is shown by red bars, the RT estimated from the RIRs by Schroeder's method is shown by blue bars, RT estimated by the Lollmann *et al.* method is shown by yellow bars, and RT estimated by the proposed method is shown by green bars. The distances between source and microphone for all of the four rooms are D_2={1.5, 3.0, 2.8, 7.1} m respectively. The standard deviations are also plotted as short lines on top of the yellow and green bars.

RT value is 0.48 seconds. Therefore, in the next section the proposed RT estimatiom method is used for blind dereverberation.

6.3 Blind Dereverberation

With the RT estimated by the methods described in Section 6.2, the dereverberation method which was already discussed in detail in Section 5.3 of Chapter 5, can be devised to work in a blind manner, i.e., without knowing the RIRs. Equation (5.8) in Chapter 5, which represents the model used to estimate the spectral variance of late reverberation from the RIRs, however, is devised here such that the spectral variance of late reverberation can be estimated from the available reverberant speech signal instead of the RIRs (which are not available in practice).

A dereverberation example is presented here for the real data from the AIR database [79]. This example will focus on the comparison between the dereverberation based on the frequency dependent statistical model with the knowledge of RIRs (the method developed in Chapter 5), the proposed blind dereverberation method based on the frequency dependent statistical model and the RT estimation using the Laplacian model, and the dereverberation achieved from the Jeub *et al.* method employing the frequency independent statistical model [78]. For comparison purpose, a revised version of both the proposed method and Jeub *et al.* method were also tested. Note that the revised version of the proposed method is a blind dereverberation method based on the frequency dependent statistical model and the RT estimation using the Gaussian model. Similarly the revised version of the Jeub *et al.* method is employing the reverberant speech for the estimation of RT instead of the RIRs used in the original version. The real RIRs used in this example are from the AIR database [79] which contains five different types of RIRs, recorded in five different room environments, namely booth, office, meeting, lecture, and stairway. Ten different anechoic speech signals from the TIMIT database, pronounced by 5 male and 5 female speakers with sampling frequency of 16 KHz, have been used here to generate the different reverberant speech signals. To establish the comparison between different dereverberation methods in this example, a pair of source-microphone distances $\{D_1, D_2\}$ m, {0.5, 1.5}, {1, 3}, {1.45, 2.8},

{2.25, 7.1}, and {1, 3} are selected respectively for the five different room environments. Performance indices used in the evaluation and comparison in this example are the segmental signal to reverberation ratio (SegSRR) [88], and the signal to distortion ratio (SDR) [103], as already defined in Chapter 5 (section 5.5). As 10 signals have been used in this example to generate different reverberant speech signals after convolving with the RIRs for five different room environments, and each environment is tested for a pair of source-microphone distances, in total 100 different reverberant speech signals have been tested. For each room type and source-microphone distance, the average results of SDR and SegSRR over the 10 different signals, are given in Figures 6.3 and 6.4 respectively.

It can be observed that for all the testing cases, dereverberation performance for the proposed blind dereverberation method (shown by the green bars) both in terms of SDR and SegSRR is better than the Jeub *et al.* method [78] (shown by the gray bars) especially for shorter source-microphone distances. Similarly, the proposed method is giving improvement for nearly all the testing cases in comparison to the Jeub *et al.* method [78], however the improvement decreases when the source-microphone distance increases. Also it can be seen that the dereverberation performance of the proposed blind dereverberation method is comparable to the dereverberation method using the RIRs. Hence it is feasible to use the proposed blind dereverberation method instead of the one employing the assumption of the RIR to be known *a priori*.

6.4 Joint Blind Dereverberation and Separation

This section presents results of joint blind dereverberation and separation algorithm for speech mixtures based on the algorithms developed in Chapter 3, Chapter 5, and the previous sections of this chapter. The proposed method is assessed in three different ways. In the first scheme, mixture signals are employed to estimate the RT blindly using the proposed blind RT estimation method followed by the blind dereverberation using frequency dependent statistical model employing the RT obtain from the previous step to estimate the spectral variance of room reverberation and then the spectral subtraction mask and the smoothed mask which is used to dereverberate the mixtures.

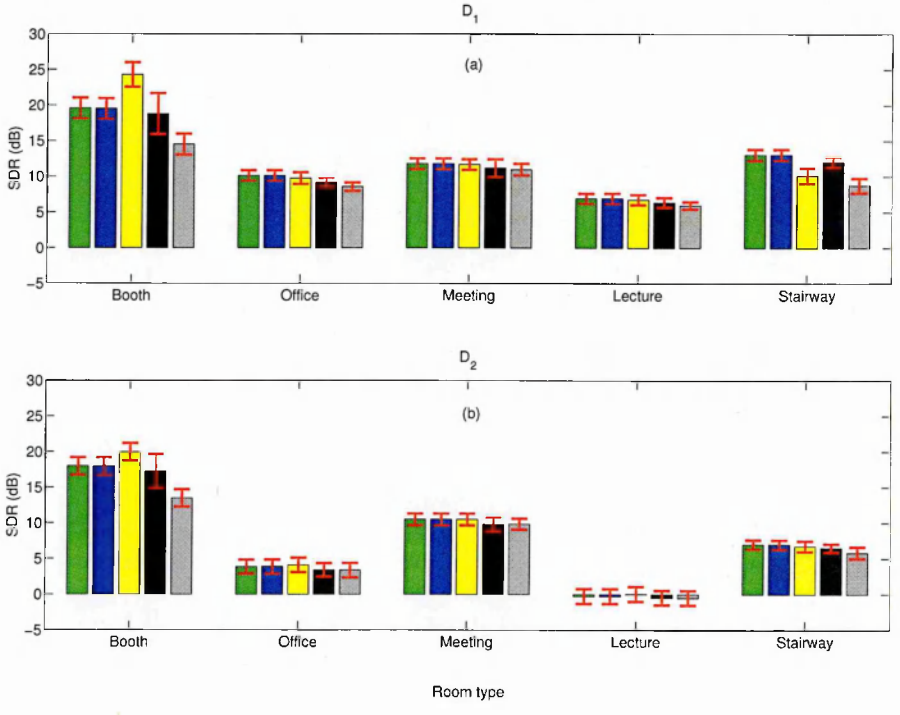


Figure 6.3: Comparison of the proposed blind dereverberation method (green bars), revised version of the proposed blind dereverberation method (blue bars), dereverberation method using the RIRs developed in Chapter 5 (yellow bars), revised version of the Jeub *et al.* method [78] (black bars), and Jeub *et al.* method [78] (gray bars) for the AIR database in terms of SDR. For each of the five rooms, two different source-microphone distances were tested, respectively $D_1 = \{0.5, 1, 1.45, 2.25, 1\}$ m and $D_2 = \{1.5, 3, 2.8, 7.1, 3\}$ m. The standard deviations are also plotted as short lines on top of the bars.

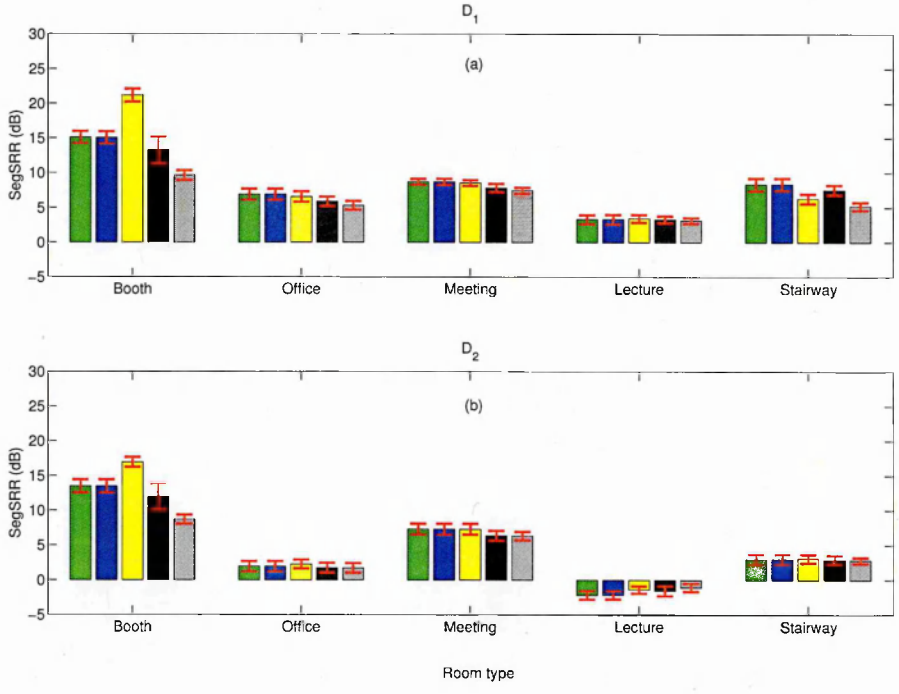


Figure 6.4: Comparison of the proposed blind dereverberation method (green bars), revised version of the proposed blind dereverberation method (blue bars), dereverberation method using the RIRs developed in Chapter 5 (yellow bars), revised version of the Jeub *et al.* method [78] (black bars), and Jeub *et al.* method [78] (gray bars) for the AIR database in terms of SegSRR. For each of the five rooms, two different source-microphone distances were tested, respectively $D_1 = \{0.5, 1, 1.45, 2.25, 1\}$ m and $D_2 = \{1.5, 3, 2.8, 7.1, 3\}$ m. The standard deviations are also plotted as short lines on top of the bars.

Figure 6.5: Block diagram showing the first scheme for the proposed joint blind dereverberation and separation algorithm. $z_1(n)$ and $z_2(n)$ are the available mixtures (microphone signals).

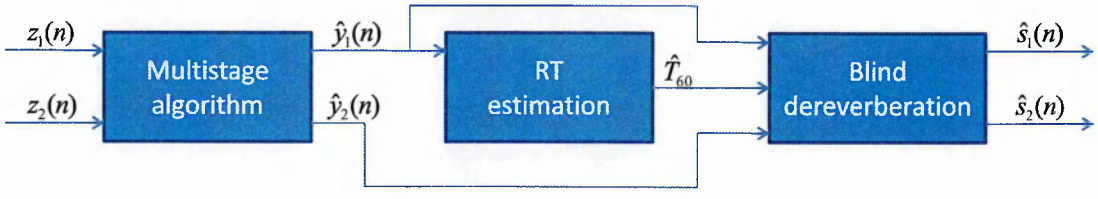


Figure 6.6: Block diagram showing the second scheme for the proposed joint blind dereverberation and separation algorithm. $z_1(n)$ and $z_2(n)$ are the available mixtures (microphone signals).

Next the separation algorithm developed in Chapter 3 (called as Multistage algorithm hereafter) is applied to the dereverberated mixtures in order to segregate the speech signals. A block diagram is given in Figure 6.5 explaining the structure of this scheme.

In the second arrangement, Multistage algorithm is applied first to the mixtures to obtain the separated speech signals. Then using the proposed blind RT estimation method, RT is estimated blindly from the separated speech followed by the frequency dependent statistical model employing the estimated RT from the previous step to estimate the spectral variance of room reverberations and then spectral subtraction mask and the smoothed mask which is used to dereverberate the separated signals. A block diagram is given in Figure 6.6 describing the second scheme.

In the third approach, a Multistage algorithm is split such that the constrained convolutive ICA method is applied first to the mixtures to obtain the estimated source signals. Next the signal obtained from the convolutive ICA is used to estimate the RT by applying the proposed blind RT estimation method followed by the dereverberation of these signals using frequency dependent statistical model. Again the frequency de-

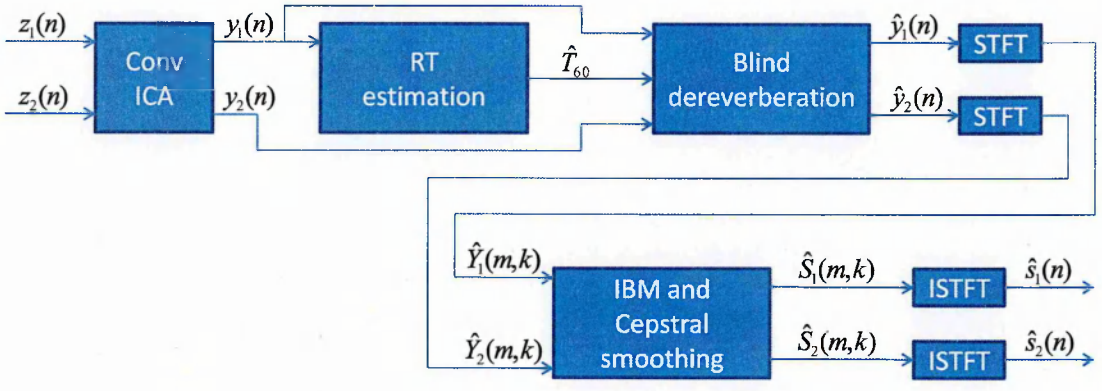


Figure 6.7: Block diagram showing the third scheme for the proposed joint blind dereverberation and separation algorithm. $z_1(n)$ and $z_2(n)$ are the available mixtures (microphone signals).

pendent statistical model is employing the RT obtained in the previous step to estimate the spectral subtraction mask followed by smoothing to achieve dereverberation. Then the T-F representation of the signals obtained in the previous step is used to estimate the IBM followed by smoothing of the estimated IBM in the cepstral domain. A block diagram is given in Figure 6.7 which is used to demonstrate the third scheme.

The performance of the proposed joint blind dereverberation and separation method has been evaluated using simulated RIRs from the image model [4] and real room recordings that were obtained in [129]. A pool of 10 different speech signals from the TIMIT database, uttered by 5 male and 5 female speakers and all sampled at 16 KHz, has been used in the experiments to generate the reverberant mixtures. A system with two inputs and two outputs is considered here in this work. The size of the room used in the case of simulated RIRs is 6.5 x 7 x 8 (m^3). The position matrices of two sources and two sensors (microphones) are set as, [1 1 3; 3 1 3], and [2 3 3; 3 3 3] respectively. Performance indices used in the evaluations are the segmental signal to reverberation ratio (SegSRR) [88], and the signal to distortion ratio (SDR) [103], as already defined in Chapter 5, in Equations (5.34) and (5.35) respectively. Notations $\Delta SegSRR$ and ΔSDR are used in the evaluations, where $\Delta SegSRR = mSegSRR_o - mSegSRR_i$ and $\Delta SDR = mSDR_o - mSDR_i$. $SegSRR_i$ and SDR_i can be obtained by replacing $\hat{s}(n)$ with an input mixture signal in (5.34) and (5.35) respectively. Similarly $SegSRR_o$

and SDR_o can be obtained by taking $\hat{s}(n)$ in (5.34) and (5.35) as the enhanced signal respectively. Note that $mSegSRR_o$, $mSegSRR_i$, $mSDR_o$, and $mSDR_i$ are the average results for fifty random tests. The performance of method proposed in this chapter is compared with that of the Multistage algorithm.

First the simulated room model [4] is used to generate the reverberant mixture signals from the pool of the clean speech signals described above, at different reverberation times, i.e., T_{60} = {200, 250, 300, 350, 400, 450, 500} ms to evaluate and compare the performance of the proposed method at different RTs. For each T_{60}, 10 anechoic signals from the pool has been used to generate different reverberant mixtures, with each consisting of two speech sources randomly picked up from the pool. In total 50 random tests have been carried out for each T_{60}, and hence in total 350 different reverberant mixtures have been used here in evaluation. Table 6.2, 6.3, and 6.4 shows for each T_{60}, the results averaged over the 50 random tests for the first, second, and third scheme of the proposed method respectively in comparison to the Multistage method.

In another set of experiments real room recordings have been used that were obtained in [129]. The real recordings were made in a reverberant room with T_{60} = 400 ms. Two omnidirectional microphones vertically placed and closely spaced are used for the recordings. Different loudspeaker positions are used to measure the room impulse responses. The room dimensions are 5.2 x 7.9 x 3.5 (m^3), and the distance between the microphones and the loudspeakers is 2 m. Further details about the recordings can be found in [129]. Clean speech signals from the pool of 10 speakers were convolved with the room impulses to generate the source signals. The average results of ΔSDR and $\Delta SegSRR$ over the 50 different random tests are given in Table 6.5, 6.6, and 6.7 for the first, second and third scheme of the proposed method respectively.

Now if the results obtained for both simulated and real data are observed in a sequence of the different schemes, it can be found that the proposed method implemented in the first scheme consistently giving better results both in terms of SDR and SegSRR than the Multistage method. For the real recordings, the proposed method in scheme 1 achieves approximately 1.5 dB gain for both SDR and SegSRR over the Multistage

Table 6.2: ΔSDR and $\Delta SegSRR$ For Simulated Data under Different $T_{60}s$

T_{60} (ms)	ΔSDR (dB)		$\Delta SegSRR$ (dB)	
	Proposed method (scheme 1)	Multistage method	Proposed method (scheme 1)	Multistage method
200	4.52	3.61	2.15	1.45
250	3.73	2.91	1.88	1.14
300	3.22	2.45	1.66	0.94
350	2.88	2.18	1.48	0.82
400	2.68	1.96	1.35	0.75
450	2.50	1.77	1.23	0.68
500	2.37	1.62	1.12	0.63

Table 6.3: ΔSDR and $\Delta SegSRR$ For Simulated Data under Different $T_{60}s$

T_{60} (ms)	ΔSDR (dB)		$\Delta SegSRR$ (dB)	
	Proposed method (scheme 2)	Multistage method	Proposed method (scheme 2)	Multistage method
200	4.49	3.61	2.06	1.45
250	3.73	2.91	1.78	1.14
300	3.20	2.45	1.55	0.94
350	2.88	2.18	1.37	0.82
400	2.63	1.96	1.22	0.75
450	2.42	1.77	1.10	0.68
500	2.27	1.62	1.01	0.63

Table 6.4: ΔSDR and $\Delta SegSRR$ For Simulated Data under Different $T_{60}s$

T_{60} (ms)	ΔSDR (dB)		$\Delta SegSRR$ (dB)	
	Proposed method (scheme 3)	Multistage method	Proposed method (scheme 3)	Multistage method
200	3.64	3.61	1.45	1.45
250	2.88	2.91	1.13	1.14
300	2.44	2.45	0.93	0.94
350	2.16	2.18	0.82	0.82
400	1.93	1.96	0.74	0.75
450	1.74	1.77	0.67	0.68
500	1.60	1.62	0.63	0.63

Table 6.5: ΔSDR and $\Delta SegSRR$ For the Real Data

Algorithm	ΔSDR (dB)	$\Delta SegSRR$ (dB)
Proposed method (scheme 1)	6.40	3.55
Multistage method	4.74	2.01

Table 6.6: ΔSDR and $\Delta SegSRR$ For the Real Data

Algorithm	ΔSDR (dB)	$\Delta SegSRR$ (dB)
Proposed method (scheme 2)	4.85	2.54
Multistage method	4.74	2.01

Table 6.7: ΔSDR and $\Delta SegSRR$ For the Real Data

Algorithm	ΔSDR (dB)	$\Delta SegSRR$ (dB)
Proposed method (scheme 3)	4.75	2.03
Multistage method	4.74	2.01

method. It is observed that in the first scheme blind dereverberation applied to the reverberant mixtures prior to separation helps in improving the separation performance. Similarly it can be found that the proposed method in the second scheme also performs better than the Multistage method for both simulated and real data. However, it can be noticed that in the second scheme of the proposed method improvement is less than the improvement achieved in the first scheme especially for real recordings. This is because in the second scheme, the separation algorithm is applied first and hence the enhancement performance is not as good as in the first scheme due to the reverberant effects in the mixture at the time of separation. The third scheme of the proposed method provides no improvement at all and the results obtained for both real and simulated data are comparable to the Multistage algorithm. Therefore, it is concluded that the proposed blind dereverberation and separation algorithm implemented in the first scheme provides better results in comparison to the implementation of the second and third scheme. Note that the proposed joint blind dereverberation and separation method has been tested based on RT estimation step employing the Gaussian decay model and it has been found that the results obtained in all the three schemes are similar to the results of the proposed joint blind dereverberation and separation method.

6.5 Summary

In this chapter a method has been developed to perform blind dereverberation and separation of convolutive speech mixtures jointly. The method has been evaluated in three different arrangements. In the first scheme, mixture signal is used to estimate RT followed by blind dereverberation and then the separation algorithm is applied to the dereverberant mixture to obtain the segregated speech signals. In the second arrangement, separation algorithm is applied first to the mixtures in order to achieve the separated speech signals. Then the obtained separated signal is used to estimate the RT blindly followed by the blind dereverberation. In the third and final scheme, the separation algorithm is divided such that the convolutive ICA is used first to obtain the estimated source signals. Then the signal obtained after convolutive ICA is used to estimate the RT followed by the blind dereverberation. Then the T-F representation

of the obtained dereverberant signals are used to estimate the IBM and finally cepstral smoothing of the IBM. As shown in the experiments that the proposed method implemented in scheme 1 performs better than scheme 2 and 3, in comparison to the related recent approach.

Chapter 7

Conclusions and Future Research

7.1 Conclusions

In this thesis the major challenging issues related to the cocktail party problem are addressed, i.e., blind separation of target speech signal from the convolutive mixtures, denoising and dereverberation, and joint blind dereverberation and separation of speech mixtures.

Firstly, the well-known problem of blind separation of speech signals is investigated. A multistage algorithm is proposed in Chapter 3 for the separation of convolutive speech mixtures using two-microphone recordings, based on the combination of ICA and IBM, together with a post-filtering process in the cepstral domain. The proposed approach consists of three major steps. A convolutive ICA algorithm [178] is first applied in order to take into account the reverberant mixing environments based on a convolutive unmixing model. Binary T-F masking is used in the second step for improving the SNR of the separated speech signal, due to its effectiveness in rejecting the energy of interference by assigning zeros to the T-F units in the masking matrix in which the energy of the interference is stronger than the target speech. The artifacts (musical noise) due to the error in the estimation of the binary mask in the segregated speech signals are further reduced by applying the cepstral smoothing technique. Compared with smoothing directly in the spectral domain, cesptral smoothing has the advantage

of preserving the harmonic structure of the separated speech signal while reducing the musical noise to a lower level by smoothing out the unwanted isolated random peaks.

The proposed method achieves considerable improvement in comparison to [178] in terms of both objective measurements using SNR and subjective listening tests, mainly due to the introduction of the binary T-F masking operation and the cepstral smoothing. The binary masking contributed mostly to the improvement of interference cancellation, and cepstral smoothing further improved the perceptual quality of the separated speech. Although the proposed method and Pedersen *et al.*'s method [129] have the similar combination structure, i.e., combination of an ICA algorithm with the IBM technique. However, the proposed algorithm directly addresses the convolutive BSS model based on the frequency-domain approach, while Pedersen *et al.*'s method is based on an instantaneous model and an instantaneous ICA algorithm, even though their algorithm has also been tested for convolutive mixtures. Second, the algorithm in [129] is iterative, which is computationally demanding. Moreover, cepstral smoothing has been introduced in the proposed method, which has the advantage of reducing the musical artifacts caused by the IBM technique.

In Chapter 4, a method is developed to deal with the effects of reflections on the target speech signal contaminated by the white Gaussian noise in a cocktail party environment. The proposed method is a one-microphone multistage algorithm. In the first step, an EMD algorithm is applied to the reverberant speech signal corrupted by white Gaussian noise to decompose it into its corresponding IMFs. Then, the IMF components with the high level of noise have been selected and denoising is applied to these selected IMFs. The denoising technique employed here is based on MMSE filtering approach called EMD-MMSE. The silence periods of the signal are detected and then the noise power spectrum is estimated by averaging the power spectra of the noisy signal. Then the MMSE estimator is applied to enhance the selected IMF components, resulting in the denoised IMF components and remaining unprocessed IMF components. In the next step, the denoised IMF components and the remaining IMF components are used to estimate the power of late reverberations as here the main focus is on late reflections which is the main cause of reducing intelligibility of target speech. It has been observed that the energy of the late reverberations is spread over the different

IMFs with different magnitudes. For this reason, spectral subtraction is applied to each IMF component according to the energy of the late reverberations present in the corresponding IMF components. Finally, the enhanced signal is reconstructed from the processed IMF components. The experimental results are provided which clearly show that using spectral subtraction for the IMF components of the noisy reverberant speech offers better denoising and dereverberation in comparison to the related method that directly uses the full-band noisy reverberant speech.

In Chapter 5, an algorithm is developed to treat the room reflections only by targeting at the late as well as the early reflections. The proposed method has two steps. In the first step a frequency dependent statistical model of the decay rate of the late reverberations is used to estimate the spectral variance of the late reverberations, and then the mask is estimated containing the spectral subtraction gain functions in the T-F domain. In order to remove the processing artifacts (musical noise) due to the error in the estimation of the mask, a smoothing function is applied to the mask in the T-F domain to filter out the artifacts. Finally, the smoothed gain function is applied to the reverberant speech to reduce the late reverberations. In the second step of the proposed method, a Wiener filtering approach is applied to reduce the early reflections. This step of the algorithm exploits the low coherence of the sound field between the different microphones (sensors) to estimate the power spectral density of the direct speech and to remove all non-coherent signal parts while keeping the coherent parts unaffected, as only the direct speech shows a high coherence among sensors. As a result the early reverberations are attenuated. It has been shown in the experimental results that the proposed algorithm offers considerably higher dereverberation performance as compared with a related recent approach using the frequency independent model.

In Chapter 6, an algorithm is presented in which the separation performance of the method proposed in Chapter 3 has been improved by incorporating the dereverberation technique developed for late reverberation in Chapter 5, with an additional step of estimating the RT blindly from the reverberant signal and hence the developed algorithm operates in a blind manner. The developed method has been employed in three different ways. Firstly, the available mixture signals are used to estimate blindly the RT based on a ML method and statistical modelling of the sound decay rate of the re-

verberant speech, followed by the dereverberation of the mixture signals to suppress the late reflections using the method based on the frequency depenedent statistical model as described in Chapter 5. Then, the separation algorithm proposed in Chapter 3 is applied to these resultant mixtures so that the source speech signals can be obtained. Secondly, the separation algorithm is applied first to the mixtures to segregate the speech signals, followed by the blind estimation of RT from the separated speech signal. Then, the dereverberation is employed to the segregated speech signals to suppress the late reflections. In the third scheme, the multistage separation algorithm proposed in Chapter 3 is split such that the convolutive ICA is first applied to the mixtures to obtain the estimated source signals. Then, the signal obtained from the convolutive ICA is used to estimate the RT followed by the blind dereverberation of the signals obtained from convolutive ICA. Then, the T-F representation of dereverberant signals are used to estimate the IBM followed by cepstral smoothing to enhance the separated speech signals. The evaluation results show that the proposed algorithm further enhances the separation performance of the multistage separation algorithm developed in Chapter 3 of the thesis.

7.2 Future Research

This dissertation suggests different directions for future research. An obvious one is the extension of the algorithm developed in Chapter 3 to the underdetermined cases. Currently this algorithm is working efficiently for the determined scenario, however its extension can offer research in the direction that is envisaged to have some potential. Similarly in Chapter 4 the method developed is based on single microphone dereverberation system. Futher research might be conducted to investigate the potentials of this method for multi-microphone system. Also the developed method is only treating the late reverberations, hence some method can be incorporated to deal with the early reflections also.

In Chapter 5 the proposed method is based on the fact that the acoustic impulse response has an exponential decay and hence the spectral variance estimator for late reverberations is using such decays. Despite the fact that this assumption is true for

many enclosed spaces, generalization will make it more interesting. For example in some cases there are coupled rooms (an enclosed space connected together using some opening), here the exponential decay rate that exhibits in each room is different and hence the total decay consists of a sum of exponential decays [161].

Another interesting idea is about the estimation procedure of RT proposed in Chapter 6, in which a statistical model based approach is adopted for estimating the RT. Currently, the proposed method is locating the free decay regions first in the reverberant speech and then employ the statistical model based ML approach to these regions to estimate RT. It can be extended in future such that the RT can be estimated from the reverberant speech without locating the free decay regions first.

References

[1] Y. Ainhoren, S. Engelberg, and S. Friedman. The cocktail party problem. *IEEE Instrumentation and Measurement Magazine*, 2008.

[2] A. Aissa-El-Bey, K. Abed-Meraim, and Y. Grenier. Blind separation of underdetermined convolutive mixtures using their time-frequency representation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1540–1550, July 2007.

[3] J. Allen, D. Berkley, and J. Blauert. Multimicrophone signal processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America*, 62:912–915, 1977.

[4] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of Acoustical Society of America*, 65, 1979.

[5] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Wang. Multichannel blind deconvolution and equalization using the natural gradient. In *IEEE Workshop on Signal Processing*, pages 101–104, 1997.

[6] S. Araki, S. Makino, H. Sawada, and R. Mukai. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica. In *Proceedings of 5th International Conference on Independent Component Analysis and Blind Signal Separation*, pages 898–905, Granada, Spain, 2004.

[7] S. Araki, S. Makino, H. Sawada, and R. Mukai. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency

mask. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 81–84, USA, 2005.

[8] S. Araki, R. Mukai, S. Makino, and H. Saruwatari. The fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. *IEEE Transactions on Speech and Audio Processing*, 11:109–116, 2003.

[9] S. Araki, H. Sawada, R. Mukai, and S. Makino. Underdetermined blind sparse source separation for arbitrarily arranged multiple sources. *EURASIP Journal of Applied Signal Processing*, 87:1833–1847, 2007.

[10] B. Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, (12):35–50, 1992.

[11] C. Avendano and H. Hermansky. Study on the dereverberation of speech based on temporal envelope filtering. In *Proceedings of 4th International Conference on Spoken Language Processing*, volume 2, pages 889–892, Philadelphia, PA, 1996.

[12] A. D. Back and A. C. Tosi. Blind deconvolution of signals using a complex recurrent network. In *IEEE Workshop Neural Networks Signal Processing*, pages 565–574, 1994.

[13] D. Bees, M. Blostein, and P. Kabal. Reverberant speech enhancement using cepstral processing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 977–980, Toronto, ON, Canada, 1991.

[14] J. Benesty, M. M. Sondhi, and Y. A. Huang. *Springer Handbook of Speech Processing*. Springer-Verlag, New York, 2007.

[15] T. Blumensath and M. Davies. Compressed sensing and source separation. In *International Conference on Independent Component Analysis and Blind Source Separation*, 2007.

[16] P. Bofill and M. Zibulevsky. Underdetrmined blind source separation using sparse representations. *Signal Processing*, pages 2353–2362, 2001.

[17] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.

[18] A. O. Boudraa and J. C. Cexus. Denoising via empirical mode decomposition. In *Proceedings of the IEEE International Symposium on Control, Communications, and Signal Processing*, March 2006.

[19] A. O. Boudraa, J. C. Cexus, S. Benramdane, and A. Beghdadi. Noise filtering using empirical mode decomposition. In *Proceedings of the IEEE International Symposium on Signal Processing and its Applications*, pages 1–4, February 2007.

[20] A. O. Boudraa, J. C. Cexus, and Z. Saidi. Emd-based signal noise reduction. *International Journal of Signal Processing*, 1(1):33–37, 2004.

[21] A. S. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, MA, 1990.

[22] P. J. Brockwell and R. A. Davis. *Time series: theory and methods.* New York: Springer, 1991.

[23] A. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple talker condition. *Acoustica*, (86):117–128, 2000.

[24] G. J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech Language*, 8(4):297–336, 1994.

[25] H. Buchner, R. Aichner, and W. Kellermann. *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, chapter Blind source separation for convolutive mixtures: A unified treatment, In Y.Huang and J. Benesty (eds.), pages 255–293. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004.

[26] E. Cands and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problem*, 23(3):969–985, 2006.

[27] E. J. Cands. Compressive sampling. In *Proceedings of the International Congress of Mathematics*, Madrid, Spain, 2006.

[28] E. J. Cands and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, March 2008.

[29] N. Chatlani and J. J. Soraghan. Speech enhancement using adaptive empirical mode decomposition. In *Proceedings of the 16th IEEE International Conference on Digital Signal Processing*, pages 1–6, 2009.

[30] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

[31] W. T. Chu. Comparison of reverberation measurements using schroeder's impulse method and decay curve averaging method. *Journal of the Acoustical Society of America*, 63:1444–1450, 1978.

[32] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing.* Wiley Press, 2002.

[33] A. Cichocki, R. Zdunek, and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 621–624, Toulouse, France, 2006.

[34] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141177, 2001.

[35] M. P. Cooke. Computational auditory scene analysis in listeners and machines. In *Tutorial at NIPS2002*, December 2002.

[36] T. J. Cox, F. Li, and P. Darlington. Extracting room reverberation time from speech using artificial neural networks. *Journal of the Audio Engineering Society*, 49:219–230, 2001.

[37] M. Davies and N. Mitianoudis. A simple mixture model for sparse overcomplete ica. In *IEE Proceedings Vision, Image, and Signal Processing*, volume 151, pages 35–43, August 2004.

[38] M. Delcroix, T. Hikichi, and M. Miyoshi. Precise dereverberation using multichannel linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):430–440, February 2007.

[39] L. Deng and D. O'Shaughnessy. *Speech Processing: A Dynamic and Optimization-Oriented Approach.* Signal Processing and Communications. Marcel Dekker, New York, 2003.

[40] H. Dillon. *Hearing aids.* New York: Thieme, 2001.

[41] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[42] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.

[43] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrica*, 81:425–455, 1994.

[44] S. Douglas, H. Sawada, and S. Makino. Natural gradient multichannel blind deconvolution and speech separation using causal fir filters. *IEEE Transactions on Speech and Audio Processing*, 13(1):92–104, January 2005.

[45] S. C. Douglas, M. Gupta, H. Sawada, and S. Makino. Spatio-temporal fastica algorithms for the blind separation of convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1511–1520, 2007.

[46] S. C. Douglas and X. Sun. Convolutive blind separation of speech mixtures using the natural gradient. *Speech Communication*, 39:65–78, December 2002.

[47] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

[48] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2), 1985.

[49] A. S. Feng and R. Ratnam. Neural basis of hearing in real-world situations. *Annual Review of Psychology*, (51):699–725, 2000.

[50] C. Fevotte and S. Godsill. A bayesian approach for blind separation of sparse sources. *IEEE Transactionss on Speech and Audio Processing*, 2005.

[51] D. FitzGerald, M. Cranitch, and E. Coyle. Shifted non-negative matrix factorization for sound source separation. In *IEEE International Workshop on Statistical Signal Processing*, pages 1132–1137, 2005.

[52] D. FitzGerald, M. Cranitch, and E. Coyle. Sound source separation using shifted non-negative tensor factorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 653–656, 2006.

[53] J. Flanagan and R. Lummis. Signal processing to reduce multipath distortion in small rooms. *Journal of the Acoustical Society of America*, 47:1475–1481, 1970.

[54] P. Flandrin, P. Goncalves, and G. Rilling. Detrending and denoising with empirical mode decomposition. In *Proceedings of the 12th European Signal Processing Conference (EUSIPCO'04)*, pages 1581–1584, Vienna, Austria, September 2004.

[55] P. Flandrin, G. Rilling, and P. Goncalves. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114, February 2004.

[56] K. Furuya and A. Kataoka. Robust speech dereverberation using multichannel blind deconvolution with spectral subtarction. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1579–1591, July 2007.

[57] N. D. Gaubitch, E. A. P. Habets, and P. A. Naylor. Multimicrophone speech dereverberation using spatiotemporal and spectral processing. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 3222–3225, Seattle, WA, May 2008.

[58] B. Gillespie, H. Malvar, and D. Florencio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3701–3704, Salt Lake City, UT, 2001.

[59] B. Gygi, G. R. Kidd, and C. S. Watson. Spectral-temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America*, 115(3):1252–1265, 2004.

[60] E. Habets, N. Gaubitch, and P. Naylor. Temporal selective dereverberation of noisy speech using one microphone. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4577–4580, Las Vegas, NV, April 2008.

[61] E. A. P. Habets. Single and multi-microphone speech dereverberation using spectral enhancement. *Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, The Netherlands*, June 2007.

[62] E. A. P. Habets, S. Gannot, and I. Cohen. Late reverberant spectral variance estimation based on a statistical model. *IEEE Signal Processing Letters*, 16:770–773, September 2009.

[63] D. E. Hall. *Musical Acoustics.* Brooks Cole, 3rd edition edition.

[64] S. Han, J. Cui, and P. Li. Post-processing for frequency-domain blind source separation in hearing aids. In *Proceedings of 7th International Conference on Information, Communications and Signal Processing, ICICS*, pages 356–360, 2009.

[65] S. Harding, J. Barker, and G. J. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:58–67, 2006.

[66] W. M. Hartmann. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter Listening in a room and the precedence effect, pages 191–210. Erlbaum, New York, 1999.

[67] O. Hazrati, K. Kokkinakis, and P. C. Loizou. A blind subband-based dereverberation algorithm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4714–4717, Texas, USA, March 2010.

[68] Z. He, S. Xie, S. Ding, and A. Cichocki. Convolutive blind source separation in the frequency domain based on sparse representation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1551–1563, 2007.

[69] P. G. Hoel. *Elementary Statistics.* New York: Wiley, 4th edition, 1976.

[70] G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15:1135–1150, 2004.

[71] N. E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and hilbert spectrum for nonlinear and nonstationary time series analysis. In *Proceedings of The Royal Society of London*, volume 454, pages 903–995, 1998.

[72] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing. a Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

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

[74] M. H. Ichir and A. M. Djafari. Hidden markov models for wavelet based blind source separation. *IEEE Transactions on Image Processing*, 15:1887–1899, July 2006.

[75] International Organization for Standardization (ISO), Geneva. *Acoustics- Measurements of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters*, 1997.

[76] T. Jan, W. Wang, and D. L. Wang. A multistage approach for blind separation of convolutive speech mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1713–1716, Taiwan, April 2009.

[77] T. Jan, W. Wang, and D. L. Wang. A multistage approach to blind separation of convolutive speech mixtures. *EURASIP Journal Speech Communication*, 53:524–539, April 2011.

[78] M. Jeub, M. Schafer, T. Esch, and P. Vary. Model-based dereverberation preserving binaural cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:1732–1745, September 2010.

[79] M. Jeub, M. Schafer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings of International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, 2009.

[80] A. Jourjine, S. Rickard, and O. Ylmaz. Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988, Turkey, June 2000.

[81] K. Khaldi, M. T. H. Alouane, and A. O. Boudraa. A new emd denoising approach dedicated to voiced speech signals. In *Proceedings of the IEEE International Conference on Signals, Circuits, and Systems*, 2008.

[82] K. Khaldi, M. T. H. Alouane, and A. O. Boudraa. Speech denoising by adaptive weighted average filtering in the emd framework. In *Proceedings of the IEEE International Conference on Signals, Circuits, and Systems*, pages 1–5, 2008.

[83] K. Khaldi, A. O. Boudraa, A. Bouchikhi, and M. T. Alouane. Speech enhancement via emd. *Eurasip Journal on Advances in Signal Processing*, 2008:1–8, March 2008.

[84] B. Kingsbury and N. Morgan. Recognizing reverberant speech with rasta-plp. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1259–1262, 1997.

[85] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):534–545, May 2009.

[86] J. Kocinski. Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms. *EURASIP Journal Speech Communication*, 50:2937, 2008.

[87] S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace Distribution and*

Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Birkhauser, Boston, 2001.

[88] P. Krishnamoorthy and S. R. Mahadeva Prasanna. Reverberant speech enhancement by temporal and spectral processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2):253–266, February 2009.

[89] H. Kuttruff. *Room Acoustics.* Elsevier Science Publishers Ltd., Lindin, 3rd edition, 1991.

[90] R. H. Lambert and A. J. Bell. Blind separation of multiple speakers in a multipath environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 423–426, April 1997.

[91] T. Langhans and H. W. Strube. Speech enhancement by nonlinear multiband envelope filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 156–159, Paris, France, 1982.

[92] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: on the uniqueness of nmf. *Computational Intelligence and Neuroscience*, 2008.

[93] K. Lebart. Speech dereverberation applied to automatic speech recognition and hearing aids. *Ph.D. dissertation, Univ. de Rennes, Rennes, France*, 1999.

[94] K. Lebart, J. Boucher, and P. N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Journal of Acta Acoustica*, 87:359–366, 2001.

[95] D. D. Lee and H. S. Seung. Learning of the parts of object by non-negative matrix factorization. *Nature*, 401(10):788–791, 1999.

[96] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing*, pages 556–562. MIT Press, 2001.

[97] T. W. Lee. *Independent Component Analysis: Theory and Applications.* Kluwer Academic Publishers, September 1998.

[98] T. W. Lee, A. J. Bell, and R. Orglmeister. Blind source separation of real world signals. In *IEEE International Conference on Neural Networks*, pages 2129–2135, June 1997.

[99] H. W. Lollmann and P. Vary. Estimation of the reverberation time in noisy environments. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Washington, USA, September 2008.

[100] H. W. Lollmann, E. Yilmaz, M. Jeub, and P. Vary. An improved algorithm for blind reverberation time estimation. In *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010.

[101] N. Madhu, C. Breithaupt, and R. Martin. Temporal smoothing of spectral masks in the cepstral domain for speech separation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 45–48, Las Vegas, USA, 2008.

[102] S. Makino, H. Sawada, R. Mukai, and S. Araki. Blind source separation of convolutive mixtures of speech in frequency domain. *IEICE Transactionss on Fundamentals of Electronics, Communications and Computer Sciences*, E88-A(7):1640–1655, July 2005.

[103] M. I. Mandel, R. J. Weiss, and P. W. Ellis. Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:382–394, February 2010.

[104] K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. In *International Conference on Independent Component Analysis*, pages 722–727, San Diego, CA, USA, December 2001.

[105] R. Mazur and A. Mertins. An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:117–126, 2009.

[106] I. McCowan and H. Bourlard. Microphone array post-filter based on noise field co-

herence. *IEEE Transactions on Speech and Audio Processing*, 11:709–716, November 2003.

[107] N. Mitianondis and M. Davies. Audio source separation: solutions and problems. *International Journal of Adaptive Control and Signal Processing*, pages 1–6, 2002.

[108] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 36(2):145–152, 1988.

[109] J. Mourjopoulos. On the variation and invertibility of room impulse response functions. *Journal of Sound Vibrations*, 102:217–228, 1985.

[110] J. Mourjopoulos and J. Hammond. Modelling and enhancement of reverberant speech using an envelope convolution method. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 1144–1147, Boston, MA, 1983.

[111] R. Mukai, H. Sawada, S. Araki, and S. Makino. Frequency domain blind source separation for many speech signals. In *International Conference on Independent Component Analysis*, pages 461–469, September 2004.

[112] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, October 2001.

[113] J. Nannariello and F. Fricke. The prediction of reverberation time using neural network analysis. *Applied Acoustics*, 58:305–325, 1999.

[114] S. M. Naqvi, Y. Zhang, and J. A. Chambers. Multimodal blind source separation for moving sources. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 125–128, Taiwan, April 2009.

[115] P. A. Naylor and N. D. Gaubitch. *Speech dereverberation.* Springer, 2010.

[116] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):34–43, 2007.

[117] S. Neely and J. Allen. Invertibility of a room impulse response. *Journal of the Acoustical Society of America*, 66:165–169, 1979.

[118] F. Nesta, M. Omologo, and P. Svaizer. Separating short signals in highly reverberant environment by a recursive frequency-domain bss. In *Proceedings Hands-Free Speech Communication and Microphone Arrays, HSCMA*, pages 232–235, Trento, Italy, 2008.

[119] F. Nesta, T. S. Wada, and B H. Juang. Coherent spectral estimation for a robust solution of the permutation problem. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 105–108, 2009.

[120] R. M. Nickel and A. N. Iyer. A novel approach to automated source separation in multispeaker environments. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 629–632, 2006.

[121] R. K. Olsson and L. K. Hansen. Blind separation of more sources than sensors in convolutive mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 657–660, 2006.

[122] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing.* Prentice Hall, New Jersey, 1975.

[123] A. Oppenhiem, R. schafer, and J. T. G. Stockham. Nonlinear filtering of multiplied and convolved signals. In *Proceedings of IEEE*, volume 56, pages 1264–1291, 1968.

[124] D. O'Shaughnessy. *Speech Communications-Human and Machine.* Institute of electrical and electronic engineers, New York, 2nd edition edition, 2000.

[125] F. S. Pacheco and R. Seara. Spectral subtraction for reverberation reduction applied to automatic speech recognition. In *Proceedings of 6th International Telecommunication Symposium*, pages 795–800, Brazil, September 2006.

[126] L. Parra and C. Spence. Convolutive blind separation of non stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8:320–327, May 2000.

[127] R. M. Parry and I. Essa. Incorporating phase information for source separation via spectrogram factorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 661–664, Honolulu, Hawaii, USA, April 2007.

[128] B. A. Pearlmutter and A. M. Zador. Monaural source separation using spectral cues. In *International Conference on Independent Component Analysis*, pages 478–485, 2004.

[129] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems. Two-microphone separation of speech mixtures. *IEEE Transactions on Neural Networks*, 19:475–492, March 2008.

[130] T. Petsatodis, C. Boukis, F. Talantzis, Z. Tan, and R. Prasad. Convex combination of multiple statistical models with application to VAD. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2314–2327, 2011.

[131] V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 1994.

[132] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications.* Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[133] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality.* Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[134] M. H. Radfar and R. M. Dansereau. Single channel speech separation using soft mask filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:2299–2310, November 2007.

[135] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan. Performance evaluation of three features for model-based single channel speech separation problem. In *Interspeech*, pages 2610–2613, Pittsburgh, PA, USA, September 2006.

[136] K. Rahbar and J. P. Reilly. A frequency domain method for blind source separation of convolutive audio mixtures. *IEEE Transactions on Speech and Audio Processing*, 13(5):832–844, 2005.

[137] R. Ratnam, D. L. Jones, and W. D. O Brien. Fast algorithms for blind estimation of reverberation time. *IEEE Signal Processing Letters*, 11(6):537–540, June 2004.

[138] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OBrien, C. R. Lansing, and A. S. Feng. Blind estimation of reverberation time. *Journal of the Acoustical Society of America*, 114(5):2877–2892, 2003.

[139] V. G. Reju, S. N. Koh, and I. Y. Soon. Underdetermined convolutive blind source separation via time-frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:101–116, 2010.

[140] G. Rilling, P. Flandrin, and P. Goncalves. On empirical mode decomposition and its algorithms. In *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03*, Italy, June 2003.

[141] G. Rilling, P. Flandrin, and P. Goncalves. Empirical mode decomposition, fractional gaussian noise and hurst exponent estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 489–492, March 2005.

[142] G. F. Rodrigues and H. C. Yehia. Limitations of the spectrum masking technique for blind source separation. In *Proceedings of the 8th Independent Component Analysis and Signal Separation*, pages 621–628, 2009.

[143] N. Roman, D. L. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114:2236–2252, 2003.

[144] W. C. Sabine. Collected papers on acoustics. 1922.

[145] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of dominant target sources using ica and time-frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2165–2173, 2006.

[146] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1592–1604, 2007.

[147] H. Sawada, R. Mukai, S. Araki, and S. Makino. Polar coordinate based nonlinear function for frequency-domain blind source separation. *IEICE Transactionss Fundamentals*, E86-A(3):590–596, March 2003.

[148] H. Sawada, R. Mukai, S. Araki, and S. Makino. A robust and precise method for solving the permutation problem of frequency domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12:530–538, September 2004.

[149] P. Scalart and J. V. Filho. Speech enhancement based on a priori signal to noise estimation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 629–632, Atlanta, USA, May 1996.

[150] M. N. Schmidt and M. Morup. Nonnegative matrix factor 2d deconvolution for blind single channel source separation. In *6th International Conference on Independent Component Analysis and Blind Signal Separation*, pages 700–707, Charleston, SC, USA, 2006.

[151] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Interspeech*, 2006.

[152] L. Schobben and W. Sommen. A frequency domain blind signal separation method based on decorrelation. *IEEE Transactions on Signal Processing*, 50(8):1855–1865, August 2002.

[153] M. R. Schroeder. New method of measuring reverberation time. *Journal of the Acoustical Society of America*, 37:409–412, 1965.

[154] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, pages 21–34, 1998.

[155] P. Smaragdis. Non-negative matrix factor deconvolution, extraction of multiple sound sources from monophonic inputs. In *5th International Conference on*

Independent Component Analysis and Blind Signal Separation, pages 494–499, Granada, Spain, 2004.

[156] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, 2007.

[157] P. Smaragdis and J. C. Brown. Nonnegative matrix factorization for polyphonic music transcription. In *IEEE International Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[158] I. Y. Soon, S. N. Koh, and C. K. Yeo. Noisy speech enhancement using discrete cosine transform. *EURASIP Journal Speech Communication*, 24:249–257, 1998.

[159] V. C. Soon, L. Tong, Y. F. Huang, and R. Liu. A robust method for wideband signal separation. In *IEEE International Symposium Circuits Systems*, pages 703–706, May 1993.

[160] S. Subramaniam, A. Petropulu, and C. Wendt. Cepstrum based deconvolution for speech dereverberation. *IEEE Transactions on Speech and Audio Processing*, 4(5):392–396, 1996.

[161] J. E. Summers, R. R. Torres, and Y. Shimizu. Statistical-acoustics models of energy decay in systems of coupled rooms and their relation to geometrical acoustics. *Journal of the Acoustical Society of America*, 116(2):958–969, 2004.

[162] Y. Tahara and T. Miyajima. A new approach to optimum reverberation time characteristics. *Applied Acoustics*, 54:113–129, 1998.

[163] K. Todros and J. Tabrikian. Blind separation of non stationary and non gaussian independent sources. In *IEEE Convention of Electrical and Electronics*, Israel, 2004.

[164] M. Tohyama, R. Lyon, and T. Koike. Source waveform recovery in a reverberant space by cepstrum dereverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 157–160, Minneapolis, MN, 1993.

[165] P. Vary and R. Martin. *Digital Speech Transmission. Enhancement, Coding and Error Concealment.* Chichester, U.K. : Wiley, 2006.

[166] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), March 2007.

[167] B. Wang and M. D. Plumbley. Musical audio stream separation by non-negative matrix factorization. In *DMRN Summer Conference*, Glasgow, 2005.

[168] D. L. Wang. *Speech Separation by Humans and Machines*, chapter On ideal binary mask as the computational goal of auditory scene analysis, pages 181–297. Kluwer Academic, Norwell MA, 2005.

[169] D. L. Wang. Time-frequency masking for speech separation and its potential for hearing aid design. *Trends in Amplification*, 12:332–353, 2008.

[170] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10:684–697, 1999.

[171] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10:684–697, May 1999.

[172] D. L. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* Wiley/IEEE Press, Hoboken NJ, 2006.

[173] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner. Speech intelligibility in background noise with ideal binary time-frequency masking. *Journal of Acoustical Society of America*, 125:2336–2347, 2009.

[174] W. Wang. Squared euclidean distance based convolutive non-negative matrix factorization with multiplicative learning rules for audio pattern separation. In *IEEE International Symposium on Signal Processing and Information Technology*, 2007.

[175] W. Wang. Available: http://personal.ee.surrey.ac.uk/Personal/W.Wang/demondata.html, 2010.

[176] W. Wang, A. Cichocki, and J. A. Chambers. A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance. *IEEE Transactions On Signal Processing*, 57:2858–2864, July 2009.

[177] W. Wang, Y. Luo, S. Sanei, and J. A. Chambers. Note onset detection via non-negative factorization of magnitude spectrum. *EURASIP Journal on Advances in Signal Processing*, pages 447–452, 2008.

[178] W. Wang, S. Sanei, and J. A. Chambers. Penalty function-based joint diagnolization approach for convolutive blind separation of nonstationary sources. *IEEE Transactions on Signal Processing*, 53:1654–1669, May 2005.

[179] M. Wu and D. L. Wang. A two-stage algorithm for one microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):774–784, 2006.

[180] Z. Wu and N. E. Huang. A study of the characteristics of white noise using the empirical mode decomposition method. In *Royal Society of London Proceedings Series A*, volume 460, pages 1597–1611, 2004.

[181] Z. Wu and N. E. Huang. Ensemble empirical mode decomposition: a noise assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(1):1–41, 2009.

[182] Z. Wu and N. E. Huang. On the filtering properties of the empirical mode decomposition. *Advances in Adaptive Data Analysis*, 2:397–414, 2010.

[183] N. Xiang. Evaluation of reverberation times using a nonlinear regression approach. *Journal of the Acoustical Society of America*, 98:2112–2121, 1995.

[184] T. Xu and W. Wang. A compressed sensing approach for underdetermined blind audio source separation with sparse representations. In *IEEE International Workshop on Statistical Signal Processing*, pages 493–496, Cardiff, UK, 2009.

[185] T. Xu and W. Wang. A block-based compressed sensing method for underdetermined blind speech separation incorporating binary mask. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Texas, USA, March 2010.

[186] B. Yegnanarayana and P. S. Murthy. Enhancement of reverberant speech using lp residual signal. *IEEE Transactions on Speech and Audio Processing*, 8(3):267–281, 2000.

[187] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin. Processing of reverberant speech for time-delay estimation. *IEEE Transactions on Speech and Audio Processing*, 13(5):1110–1118, November 2005.

[188] T. Yoshioka, T. Hikichi, and M. Miyoshi. Dereverberation by using time-variant nature of speech production system. *Eurasip Journal on Advance Signal Processing*, 2007(10.1155/2007/65698), June 2007.

[189] T. Yoshioka, T. Nakatani, and M. Miyoshi. Fast algorithm for conditional separation and dereverberation. In *Proceedings of the 17th European Signal Processing Conference*, pages 1432–1436, 2009.

[190] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2578–2581, New York, 1988.

[191] M. Zibulevsky and P. Bofill. Underdetermined blind source separation using sparse representations. *IEEE Transactions on Signal Processing*, 81(11):2353–2362, 2001.

[192] M. Zibulevsky and B. A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.