Robust Single-Channel Speech Enhancement and Speaker Localization in Adverse Environments

by

Saeed Mosayyebpour
B.Sc., Amirkabir University of Technology, 2007, 2009
M.Sc., Amirkabir University of Technology, 2009

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

Robust Single-Channel Speech Enhancement and Speaker Localization in Adverse Environments

by

Saeed Mosayyebpour
B.Sc., Amirkabir University of Technology, 2007, 2009
M.Sc., Amirkabir University of Technology, 2009

Supervisory Committee

Dr. T. Aaron Gulliver, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. Morteza Esmaeili, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical and Computer Engineering)

Dr. George Tzanetakis, Outside Member
(Department of Computer Science)

Supervisory Committee

Dr. T. Aaron Gulliver, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. Morteza Esmaeili, Co-Supervisor
(Department of Electrical and Computer Engineering)

Dr. Wu-Sheng Lu, Departmental Member
(Department of Electrical and Computer Engineering)

Dr. George Tzanetakis, Outside Member
(Department of Computer Science)

ABSTRACT

In speech communication systems such as voice-controlled systems, hands-free mobile telephones and hearing aids, the received signals are degraded by room reverberation and background noise. This degradation can reduce the perceived quality and intelligibility of the speech, and decrease the performance of speech enhancement and source localization. These problems are difficult to solve due to the colored and non-stationary nature of the speech signals, and features of the Room Impulse Response (RIR) such as its long duration and non-minimum phase. In this dissertation, we focus on two topics of speech enhancement and speaker localization in noisy reverberant environments.

A two-stage speech enhancement method is presented to suppress both early and late reverberation in noisy speech using only one microphone. It is shown that this method works well even in highly reverberant rooms. Experiments under different acoustic conditions confirm that the proposed blind method is superior in terms of

reducing early and late reverberation effects and noise compared to other well known single-microphone techniques in the literature.

Time Difference Of Arrival (TDOA)-based methods usually provide the most accurate source localization in adverse conditions. The key issue for these methods is to accurately estimate the TDOA using the smallest number of microphones. Two robust Time Delay Estimation (TDE) methods are proposed which use the information from only two microphones. One method is based on adaptive inverse filtering which provides superior performance even in highly reverberant and moderately noisy conditions. It also has negligible failure estimation which makes it a reliable method in realistic environments. This method has high computational complexity due to the estimation in the first stage for the first microphone. As a result, it can not be applied in time-varying environments and real-time applications. Our second method improves this problem by introducing two effective preprocessing stages for the conventional Cross Correlation (CC)-based methods. The results obtained in different noisy reverberant conditions including a real and time-varying environment demonstrate that the proposed methods are superior compared to the conventional TDE methods.

Contents

List of Tables

List of Figures

ACRONYMS

AIF Adaptive Inverse Filtering

AIR Aachen Impulse Response

ASR Automatic Speech Recognition

BSD Bark Spectral Distortion

CRLB Cramer-Rao Lower Bound

CS Compressed Sensing

DFT Discrete Fourier Transform

ELRR Late Reverberation energy Ratio

EVD Eigen Value Decomposition

FFT Fast Fourier Transform

FIR Finite Impulse Response

GCC Generalized Cross-Correlation

GMM Gaussian Mixture Models

LP Linear Predictive

LPC Linear Predictive Coding

ISTFT Inverse Short-Time Fourier Transform

ML Maximum Likelihood

MOS Mean Opinion Score

MSE Mean Square Error

MTF Modulation Transfer Function

PESQ Perceptual Evaluation of Speech Quality

PHAT Phase Transform

PSD Power Spectral Density

RIR Room Impulse Response

RT60 Reverberation Time

TDE Time Delay Estimation

TDOA Time Difference Of Arrival

SCOT Smoothed Coherence Transform

SegSIR Segmental Signal-to-Interference Ratio

SegSNR Segmental Signal-to-Noise Ratio

SNR Signal to Noise Ratio

SRR Signal to Reverberation Ratio

STD Standard Deviation

STFT Short Time Fourier Transform

STPSD Short Time Power Spectral Density

ACKNOWLEDGEMENTS

"In the name of Allah, the most Gracious, the most Merciful."

I would never have been able to finish my dissertation without the guidance of my committee members, help from my family and wife.

My first and sincere appreciation goes to Prof. Aaron Gulliver, my supervisor for all I have learned from him and for his continuous help and support in all stages of my Ph.D. study. He always encouraged me to move forward and progress. He is the real meaning of a supportive supervisor as he has supported me not only by providing the research assistantship over almost three years, but also academically and emotionally through the rough road of my study in Canada. I want to sincerely thank him from the bottom of my heart for all he did for me. Also I would like to express my deepest gratitude and respect to Prof. Morteza Esmaeili as my co-supervisor whose advice and insight was invaluable to me.

I would like to thank Dr. Ivan J. Tashev from Microsoft Research for his willingness to serve as my external examiner. I am deeply indebted to Dr George Tzanetakis and Prof. Wu-Sheng Lu from for their great efforts and significant amount of time to serve on my Ph.D. supervisory committee. I will not forget Dr. Tzanetakis for his indispensable advice, help and support on different aspect of my study. With Prof. Lu I had probably the most useful classes I have ever taken and I do not think I will even realize how much it helped me until I start working.

In addition, I would like to thank Dr. Hamid Sheikhzadeh and Dr. Hamidreza Amindavar from Amirkabir University for their exceptional knowledge and generous help.

I would like to express my deepest gratitude for the constant support, understanding, inspiration and unconditional love that I have received from my family.

Finally, and most importantly, I would like to thank my wife Mahdis. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock.

Saeed Mosayyebpour, 2014

DEDICATION

To my parents and my lovely wife, Mahdis.

Chapter 1

Introduction

1.1 Speech Source Signal

In general, wideband speech covering the frequency range 0.3-8 kHz has a more pleasant quality compared to narrowband speech which covers the range 0.3-4 kHz [1]. This dissertation considers wideband speech with a sampling frequency of 16 kHz. The speech signal has colored and non-stationary characteristics, making problems such as speech enhancement and localization more challenging. The analysis of the speech signal is typically done on a block-by-block basis (here 32 ms). A speech signal, $s[n]$, can be modeled as an excitation signal, $e[n]$, convolved with a vocal tract filter, $h_s[n]$ [2]. In frequency domain, this can be written as

$$S(z) = E(z)H_s(z) \tag{1.1}$$

where $S(z)$, $E(z)$, and $H_s(z)$ are the z-transform of $s[n]$, $e[n]$, and $h_s[n]$. The vocal tract filter is usually modeled as a linear system that is assumed to be time-varying such that over short time intervals it can be described by the all-pole transfer function [2]

$$H_s(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.2}$$

where G, and p, are the gain and number of poles for the all-pole transfer function. The signals are related by a difference equation of the form [2]

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Ge[n] \tag{1.3}$$

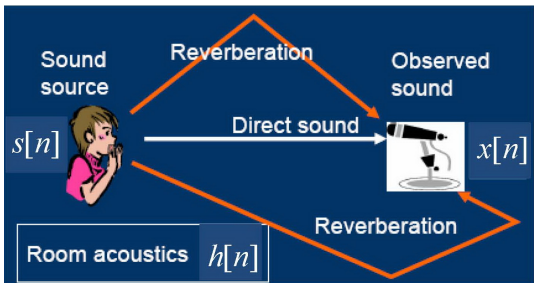

Figure 1.1: The speech signal received by a microphone in a room.

Using standard Linear Predictive (LP) analysis, a set of prediction coefficients a_k that minimize the mean-squared prediction error between $s[n]$ and a predicted signal can be obtained [2].

1.2 Reverberation in Enclosed Spaces

Signals recorded with a distant microphone in an enclosed room usually contain reverberation artifacts caused by reflections from walls, floors, and ceilings. In the context of this work, reverberation is due to multi-path propagation of the speech signal from its source to one or more microphones. This leads to spectral colouration causing a deterioration of the signal quality and intelligibility in many communication environments such as hands-free telephony and audio-conferencing. This can seriously degrade applications such as automatic speech recognition, speech separation and source localization. These detrimental effects are magnified when the speaker to microphone distance is increased.

In addition, the received signal is distorted by additive noise. The main difference between the noise and reverberation is that the reverberation is dependent on the speech signal whereas the noise can be assumed to be independent from this signal. Thus the problem of reverberation is more challenging than the problem of additive noise.

Figure 1.1 shows the received speech signal at the microphone, $x[n]$, which is composed of the reverberant speech signal, $z[n]$, and the background noise, $\nu[n]$, i.e.

$$z[n] = s[n] \star h[n] \tag{1.4}$$

$$x[n] = z[n] + \nu[n] \tag{1.5}$$

where $\star$ denotes convolution. $s[n]$ is the clean speech, and $h[n]$ is the Room Impulse Response (RIR). The impulse response of an acoustic channel is usually very long and has nonminimum phase, making the problems given above even more difficult.

The reverberation time quantifies the severity of the reverberation in a room, and is denoted by RT60 . This is usually defined as the time for the sound pressure to be attenuated by 60 dB after the source is switched off. The RIR is usually modeled by a finite impulse response (FIR) filter whose length is approximately RT60 $\times f_s$ where f_s is the sampling frequency (here 16 kHz). Reverberation is related to the surface absorption coefficient α_i, $i = 1, \ldots, 6$, where i denotes one of the room surfaces. This coefficient which determines how much sound is absorbed (and thus reflected) from room surfaces. This coefficient is a function of the incident angle, frequency, and material properties. In practice, it is averaged over the possible incidence angles. The reverberation time is related to the absorption coefficient through Sabines equation [10]

$$\text{RT60} = 0.163\frac{V}{A} \tag{1.6}$$

where V is the room volume, S_i is the reflection surface area, and A is the total absorption surface area given by $A = \sum_i \alpha_i S_i$.

The perception of reverberation is mainly based on a two-dimensional perceptual space. The two components are *coloration* and *echo* [10]. While echoes smear the speech spectra and reduce the intelligibility and quality of the speech signals, coloration distorts the speech spectrum [10]. Coloration which results from the non-flat frequency response of the early reflections (reflections that arrive shortly after the direct sound). The echoes are directly related to the reverberation time. Furthermore, the late reverberation components (reflections that arrive after the early reverberation), increase as RT60 is increased.

1.3 Scope and Dissertation Outline

This dissertation considers several new techniques aimed to address the problems of single-channel speech enhancement and speaker localization in adverse conditions such as high reverberation and additive background noise. For the first problem, the goal is to effectively suppress the effects of both early and late reverberation in noisy speech using a signal from one microphone. For the second problem, the goal is to accurately localize the speaker position in a highly reverberant room with

additive background noise using the a small number of microphones. These goals are very challenging yet the problems are significant. Here we briefly mention the main contributions of this dissertation for both speech enhancement and source localization. For speech enhancement, we propose a two-stage method using the inverse filtering to reduce the early reverberation in the first stage and spectral subtraction to reduce the noise and the residual reverberation in the second stage. Our contributions to speech enhancement are listed below.

- We propose an adaptive gradient-ascent algorithm for the input LP residual of a reverberant speech signal based on skewness instead of the commonly used metric (kurtosis).

- We optimize the algorithm for implementation. This includes an effective algorithm for estimating the expected value of the feedback function, and an efficient procedure for filter initialization, which can be used with very high reverberation times (above 2 s).

- A denoising algorithm is presented which is superior to other well-known denoising methods in noisy reverberant environments. Several denoising methods have been proposed [36]-[39] that perform well under noisy conditions. However, most perform poorly when both noise and reverberation is present, especially when the noise is non-stationary and speech-like (babble noise). This is largely because estimation of the short time power spectral density (STPSD) of the noise is greatly affected by the reverberation, particularly with babble noise. To solve this problem, for each frequency-bin in a time frame, statistical noise estimation is used to obtain the optimal spectral weighting based on the estimated Signal to Noise Ratio (SNR). This provides more robust denoising in reverberant conditions.

- A late reverberation reduction method is proposed which is more effective than the spectral subtraction of Wu and Wang [19]. This is because a better weight function has been used to estimate the STPSD of the late components. Then the spectral weight for filtering has been modified using the *a priori* Signal to Reverberation Ratio (SRR) to calculate the *a posteriori* SRR, decision-directed estimator, and changing the power of the spectral weight based on the SRR.

- A new method is proposed to reduce the effects of the pre-echo components remaining after inverse filtering. These components are one of the most serious

problems in speech enhancement because they are not a natural phenomenon to which the ear is accustomed.

For speaker localization, we propose two new techniques for Time Delay Estimation (TDE) and the contributions are listed below.

- A novel technique for TDE based on adaptive inverse filtering is proposed. This method uses the inverse filtering algorithm to estimate the inverse filter of the channels in order to accurately estimate the Time Difference Of Arrival (TDOA).

- Two preprocessing stages for TDE method are introduced, namely all-pass processing and spectral subtraction. It is shown that with these preprocessing stages, the performance of the TDE method is improved.

The dissertation is organized as follows.

- Chapter 2 presents a solution to the problem of single-microphone speech enhancement in a noisy reverberant room. This chapter consists of 5 sections. In the first section, a brief review of existing single-microphone speech enhancement methods is provided, and the main challenges and unsolved problems are given. The next three sections present the steps of the proposed solution. Performance results which demonstrate the effectiveness of the proposed method in highly reverberant rooms noise are provided in the last section.

- Chapter 3 present a solution to the problem of speaker localization in a reverberant room. The three main categories of techniques to solve the problem of source localization are introduced. TDOA-based methods are the most effective solutions for this problem. Accurate and robust TDE is the key to the effectiveness of the localization in this category. So this chapter mostly devotes to the problem of TDE and this problem in reverberant noisy conditions is investigated. The most common TDE methods in the literature are reviewed and the main challenges to be solved are presented. Then, in Section 3.1.1, our novel and the most accurate TDE method based on adaptive inverse filtering is thoroughly presented. In Section 3.1.2, we introduce another method based on two novel preprocessing for TDE application. The results shown in Section 3.1.4 demonstrate the effectiveness of our methods compared with the conventional techniques in the literature.

- Chapter 4 outlines some future works and the plan for ongoing research. Eight main ideas are presented to extend and improve the existing methods to solve the problem of both speech enhancement and speaker localization in a noisy reverberant room.
- A summary of our research is provided in Chapter 5.

1.4 Publications

1.4.1 Journal Publications

- S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "**Single-Microphone LP Residual Skewness-based Approach for Inverse Filtering of Room Impulse Response**," IEEE Trans. Audio, Speech and Lang. Process., vol. 20, pp. 1617–1632, July 2012.
- S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "**Single-Microphone Early and Late Reverberation Suppression in Noisy Speech**," IEEE Trans. Audio, Speech, Lang. Process.,vol. 21, no. 2, pp. 322–335, Feb. 2013.
- S. Mosayyebpour, A. Keshavarz, M. Biguesh, T. A. Gulliver, and M. Esmaeili "**Speech-Model based Accurate Blind Reverberation Time Estimation Using an LPC Filter**," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 6, pp. 1884–1893, Aug. 2012.
- S. Mosayyebpour, et al "**Single Source Time Delay Estimation using Two Microphones in a Noisy Reverberant Environment**," IEEE Trans. Audio, Speech, Lang. Process., 2014 (submitted).
- S. Mosayyebpour, et al "**Time Delay Estimation based on Logarithm Phase Difference in Reverberant and Time Varying Environments**," IEEE Signal Process. Letters, 2014 (submitted).

1.4.2 Conference Publications

- S. Mosayyebpour, H. Lohrasbipeydeh, M. Esmaeili, and T. A. Gulliver, "**Time Delay Estimation via Minimum-Phase and All-Pass Component Pro-**

cessing," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.(ICASSP), Vancouver, BC, pp. 4285–4289, May 2013.

- S. Mosayyebpour, T. A. Gulliver, and M. Esmaeili, "**Single-Microphone Speech Enhancement by Skewness Maximization and Spectral Subtraction**," International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1–4, Sep. 2012.

- S. Mosayyebpour, A. Sayyadiyan, M. Zareian, and A. Shahbazi, "**Single Channel Inverse Filtering of Room Impulse Response by Maximizing Skewness of LP Residual**," IEEE Int. Conf. on Signal Acquisition and Process. (ICSAP), pp. 130–134, Feb. 2010.

- S. Mosayyebpour, A. Sayyadiyan, E. Soltan Mohammadi, A. Shahbazi, and A. Keshavarz, "**Time Delay Estimation using One Microphone Inverse Filtering in a Highly Reverberant Room**," Proc. IEEE Int. Conf. on Signal Acquisition and Process. (ICSAP), pp. 140–144, Feb. 2010.

Chapter 2

Single-Channel Speech Enhancement in a Noisy Reverberant Room

Speech enhancement in a noisy reverberant environment is a difficult problem because (i) speech signals are colored and nonstationary, (ii) noise signals can change dramatically over time, and (iii) the impulse response of an acoustic channel is usually very long and has nonminimum phase. When multiple microphones are available, spatial processing can be used to improve the performance of speech enhancement techniques. However many speech communication systems are equipped with only a single microphone. As a consequence, a number of single microphone speech enhancement techniques have been developed.

There has been significant research on single microphone additive noise suppression algorithms, e.g. [4]. If the noise is negligible, the speech enhancement task is just speech dereverberation. Bees et al. [5] employed a cepstrum based method to estimate the Room Impulse Response (RIR), and used a least squares technique for inversion. Satisfactory results were only obtained for minimum phase or mixed phase responses with a few zeros outside the unit circle in the z-plane, which restricts the use of this algorithm in real conditions. Similarly, Kumar and Stern [6] built on recent developments that represent reverberation in the cepstral feature domain as a filtering operation. They formulated a maximum likelihood objective function to obtain an inverse reverberation filter. However, this method can only improve the Automatic Speech Recognition (ASR) for moderate reverberation times. Unoki et

al. [7] proposed the power envelope inverse filtering method, which is based on the Modulation Transfer Function (MTF), to recover the average envelope modulation spectrum of the original speech. However, this method has limited applicability due to the assumptions which do not necessarily match the features of real speech (real speech signals were not considered), and reverberation (a simple exponential model was employed for the RIR). Nakatani et al. [8] have shown that it is possible to accurately estimate the dereverberation filter for a Reverberation Time (RT60) up to 1 s. However, the method in [8] requires that the RIR remains constant for a considerable time duration.

Several researchers have considered only late reverberation suppression by assuming the early and late reverberant speech components are independent. The late reflection component is suppressed in the Short-Time Fourier Transform (STFT) domain using so-called spectral enhancement methods. This is achieved by estimating the Short-Time Power Spectral Density (STPSD) of the late reverberant speech component in order to perform magnitude subtraction without phase correction. Thus the main challenge is to estimate the STPSD of the late reverberant speech component from the received signal. More recently, a variety of techniques have been proposed to estimate the STPSD of the late reverberant speech component [9]-[15].

Spectral subtraction is a commonly employed technique for dereverberation. It can be used in real-time applications, and results show a reduction in both additive noise and late reverberation. However, artifacts such as musical noise are introduced due to the nonlinear filtering, and *a priori* knowledge of the RIR (i.e. the reverberation time), is usually required. Yegnanarayana and Murthy [16] proposed an LP residual based approach which identifies and manipulates the residual signal according to the regions of reverberant speech, namely, high Signal to Reverberation Ratio (SRR), low SRR, and reverberant signal only. This temporal domain method mainly enhances the speech specified features in the high SRR regions. In [17], the authors effectively combined a modified LP residual based approach (to enhance reverberant speech in the high SRR regions), with spectral subtraction to reduce late reverberation. In [18], a method was proposed which makes use of the complex cepstrum and LP residual signal to deconvolve the reverberant speech signal.

To date, most single microphone dereverberation methods have been designed to reduce the effects due mostly to late reverberation. However, the early reverberation frequency response is rarely flat, so it distorts the speech spectrum and reduces speech quality. Since joint suppression of both early and late reverberation is quite challeng-

ing, few (single-microphone) two-stage algorithms have appeared in the literature. Wu and Wang [19] proposed an inverse filtering method which maximizes the kurtosis of the LP residual to reduce the early reverberation, followed by spectral subtraction to reduce late reverberation. However, the inverse filtering to reduce early reverberation effects is only effective when the reverberation time is in the range 0.2-0.4 s. For high reverberation times, the kurtosis based objective function for adaptive inverse filtering has many saddle points (along with the maximum points), and convergence is usually to one of them, leading to an inaccurate filter estimate [28]. Moreover, their spectral subtraction tends to produce annoying musical noise, particularly at high reverberation intensities. They also did not consider noisy environments. A similar approach is described in [20] where temporal averaging to combat early reverberation is combined with spectral subtraction.

In a real environment, the reverberant speech signals are usually contaminated with nonstationary additive background noise. This can greatly deteriorate the performance of dereverberation techniques. Some single-microphone methods take the presence of noise into account, and they typically employ spectral subtraction for noise reduction. Habets et al. [20] used a statistical model for applying spectral subtraction to reduce both the reverberation and noise. However, reverberation time estimation in noisy conditions is required which is a non-trivial problem. Similarly, joint suppression of late reverberation and additive background noise was achieved in [14] using a generalized spectral subtraction rule with Maximum Likelihood (ML) estimation of the reverberation time. Attias et al. [21] presented a unified probabilistic framework for denoising and dereverberation, but their method is not effective for long reverberation times. The long-term correlation in the Discrete Fourier Transform (DFT) domain was exploited in [22] to suppress only late reverberation and noise. In [23], a method was proposed for reducing only the late reverberation of speech signals in noisy environments using the amplitude of the clean speech signal. This signal was obtained using an adaptive estimator that minimizes the Mean Square Error (MSE) under signal presence uncertainty. Finally, an ML based method was proposed in [24] for noise suppression and dereverberation. However, it requires that the Power Spectral Density (PSD) of the noise is known.

From the above discussion, it can be concluded that the joint suppression of early and late reverberation in noisy conditions, especially with long reverberation times and using only one microphone, is a very challenging yet significant problem. A two-stage speech enhancement method is proposed to reduce the both early and

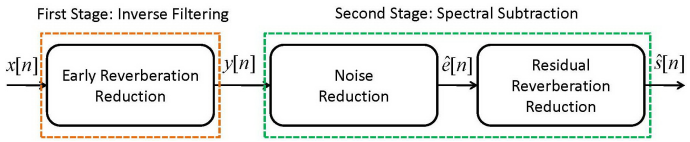

Figure 2.1: Block diagram of the proposed two-stage method for speech signal enhancement in noisy reverberant environments.

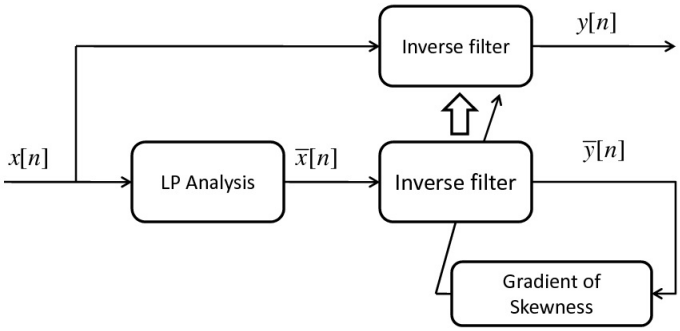

Figure 2.2: Block diagram of the inverse filtering method for the first stage of speech enhancement.

late reverberation effects in noisy speech [25]. A block diagram of the two-stage speech enhancement method is shown in Fig. 2.1. In the first stage, a blind inverse filtering method [28] is used to reduce the early reverberation effects. Then spectral subtraction is used to reduce both the noise and the residual reverberation effects [25]-[26]. In the following sections, each stage of the proposed method is described.

2.1 Inverse Filtering for Early Reverberation Suppression

Generally, methods based on inverse filtering provide better dereverberation and greatly mitigate early reverberation as long as the RIR is time-invariant. However, current single-microphone inverse filtering methods are sensitive to noise, and they perform poorly in highly reverberant rooms. Therefore, a blind inverse filtering method is presented here which works even in highly reverberant rooms and is robust to low to moderate additive background noise.

A block diagram of the inverse filtering technique is shown in Fig. 2.2, where $x[n]$ is the reverberant speech received by the microphone and $\mathbf{h}^{(r)}$ is the FIR inverse filter of length L in the r-th iteration. The LP residual signal $\overline{x}[n]$ is calculated from the

reverberant speech using an Linear Predictive Coding (LPC) filter of order 10 with a frame size of 32 ms. The signal after inverse filtering is given by

$$\overline{y}_n = (\mathbf{h}^{(r)})^T \overline{\mathbf{x}}[n], \tag{2.1}$$

where

$$\mathbf{h}^{(r)} = \left[h_0^{(r)}, h_1^{(r)}, \ldots, h_{L-1}^{(r)}\right]^T, \tag{2.2}$$

and $\overline{\mathbf{x}}[n]$ is a vector of length L containing elements n to $n-L+1$ of $\overline{x}[n]$. $\mathbf{h}$ is estimated recursively to maximize the skewness, denoted by $\Psi^{(s)}(\overline{y}_n) = \frac{E\{\bar{y}_n^3\}}{E^{\frac{3}{2}}\{\bar{y}_n^2\}}$, using an adaptive gradient-ascent algorithm. The filter update rule in the time domain is given by [28]

$$\mathbf{h}^{(r+1)} = \mathbf{h}^{(r)} + \mu \nabla_{\Psi^{(s)}(\mathbf{h}^r)}, \tag{2.3}$$

$$\nabla_{\Psi^{(s)}(\mathbf{h}^r)} \approx 3\left(\frac{\bar{y}^2 E\{\bar{y}^2\} - \bar{y}E\{\bar{y}^3\}}{E^{\frac{5}{2}}\{\bar{y}^2\}}\right)\bar{\mathbf{x}} = g\bar{\mathbf{x}} \tag{2.4}$$

where g is the feedback function. μ is the step-size controlling the learning rate which is set to 3×10^{-9}.

As a direct time domain implementation may have slow or no convergence, a frequency domain implementation of the adaptive filter is used [28]. In this formulation, the LP residual of the reverberant speech signal $\bar{x}[n]$ is segmented into blocks of length L. The blocks are increased to $2L$ samples by zero-padding, and a Fast Fourier Transform (FFT) of length $2L$ is computed for each block. The feedback function g is segmented into blocks of length $2L$ with L samples overlapping, and an FFT of length $2L$ is computed for each block. Denote the number of blocks by T. The filter update in the frequency domain is then

$$\mathbf{H}'^{(r+1)} = \mathbf{H}^{(r)} + \frac{\mu}{T}\sum_{i=1}^{T} \mathbf{G}_i \bar{\mathbf{X}}_i^*, \tag{2.5}$$

$$\mathbf{H}^{(r+1)} = \frac{\mathbf{H}'^{(r+1)}}{|\mathbf{H}'^{(r+1)}|}, \tag{2.6}$$

where $\mathbf{H}^{(r)}$ is the FFT of the inverse filter $\mathbf{h}$ in the rth iteration. $\mathbf{G}_i$ and $\bar{\mathbf{X}}_i$ denote, respectively, the FFT of g and $\bar{x}$ for the ith block. The superscript $*$ denotes complex

conjugate. The inverse filter is initialized with a simple all-pass filter

$$\mathbf{H}^{(0)} = [1\ 1\ 1 \ldots 1]^T. \tag{2.7}$$

Equation (2.6) ensures that the inverse filter is normalized. This is necessary to keep the algorithm numerically stable since an increasing $\bar{y}$ increases $\Psi(\bar{y}_n)$ without improving the inverse filter estimation, in which case the norm of $\mathbf{h}^{(r)}$ grows rapidly [28]. Our results show that a step size of $\mu = 3 \times 10^{-9}$ requires approximately 300 iteration for convergence.

As the RIR length is proportional to the Reverberation Time (RT60)[1], the inverse filter length L should be chosen accordingly. The length should be as short as possible to limit the computational complexity. Suitable inverse filter lengths for different reverberation times based on our extensive experimental results are given in Table 2.1 [28]. This table can be used when the reverberation time is known or has been

Table 2.1: Inverse Filter Lengths for Different RT60 Values

RT60 (ms)	150-500	600-1100	1200-4000
L (sample)	2000	4000	6000

estimated, e.g. using our approach in [29]. This table is not precise for all RIRs which might have different room dimension and different speaker-microphone positions. The most reliable solution especially when the reverberation time is unknown, is exploiting a characteristic of good inverse filters, namely a dominant peak that exponentially decays in reverse time [25].

2.2 Background Noise Reduction

The inverse-filtered speech signal can be expressed as

$$y[n] = e[n] + \nu'[n], \tag{2.8}$$

$$e[n] = s[n] \star h_{eq}[n], \tag{2.9}$$

where $h_{eq}[n]$ is the equalized impulse response, $s[n]$ is the clean speech signal and $\nu'[n]$ is additive noise. $\star$ denotes convolution. A block diagram of the spectral subtraction method for noise and late reverberation reduction is shown in Fig. 2.3. This method

[1]The length of the RIR is approximately equal to RT60 $\times f_s$ (sampling frequency).

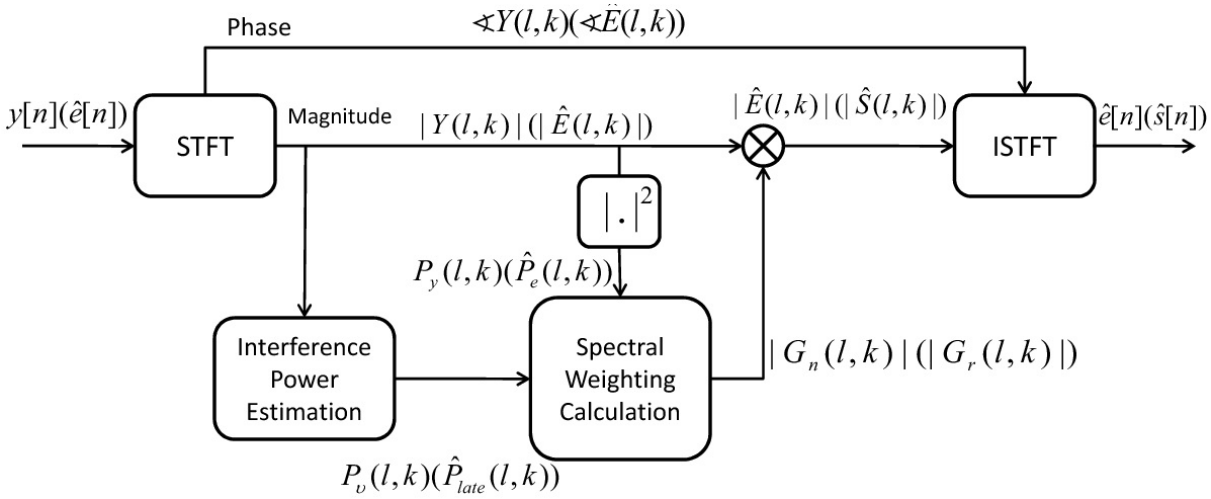

Figure 2.3: Block diagram of the spectral subtraction method for noise reduction (symbols without parenthesis) and late reverberation suppression (symbols with parenthesis).

is based on modifying the short-time spectral magnitude of the input signal by multiplying it with the spectral weighting obtained from the noise or late reverberation Power Spectral Density (PSD).

Since the analysis is in the time-frequency domain, the input speech signal is transformed using a Short-Time Fourier Transform (STFT) giving

$$Y(l,k) = \sum_{n=0}^{K-1} y[n+lR]u[n]e^{-i\frac{2\pi k}{K}n}, \tag{2.10}$$

where $i = \sqrt{-1}$, $l = 0, 1, \ldots$ is the time frame index, $k = 0, 1, \ldots, K-1$ is the frequency-bin index, $u[n]$ is a Hamming window of size K (here 32 ms), and R is the frame rate which is the number of samples between two successive frames (here 16 ms).

It can be assumed that $e[n]$ and $\nu'[n]$ are statistically independent so the PSD of $y[n]$ is equal to the sum of the PSDs of $e[n]$ and $\nu'[n]$. Let $P_\nu(l,k)$ and $P_y(l,k)$ denote the estimated STPSD of the noise and inverse-filtered signal, respectively. $P_\nu(l,k)$ can be estimated using minimum statistics [33]-[35]. The STPSD of the inverse-filtered speech signal is obtained as

$$P_y(l,k) = |Y(l,k)|^2, \tag{2.11}$$

where $|.|$ denotes magnitude. Then the PSD of the noise signal $\bar{P}_\nu(l)$ and the inverse-filtered speech signal $\bar{P}_y(l)$ are

$$\bar{P}_y(l) = \sum_{k=0}^{K-1} P_y(l,k), \tag{2.12}$$

$$\bar{P}_\nu(l) = \sum_{k=0}^{K-1} P_\nu(l,k). \tag{2.13}$$

The optimal spectral weighting can be calculated as follows

$$G_n(l,k) = \begin{cases} \min(\rho V(l,k), 1) & \text{for } V(l,k) \geq \frac{1}{o(l,k)+\rho} \\ 1 - o(l,k)V(l,k) & \text{otherwise} \end{cases} \tag{2.14}$$

where $V(l,k)$ is defined as

$$V(l,k) = \sqrt{\frac{P_\nu(l,k)}{P_y(l,k)+\varepsilon_y}} \tag{2.15}$$

ε_y is set to a small value (e.g. 1), when $P_y(l,k)$ is zero to avoid infinite values for $V(l,k)$ and is zero elsewhere. ρ is the noise floor parameter which is set to 0.1. $o(l,k)$ in (3.42) is the subtraction factor which depends on the SNR and is given by

$$o(l,k) = \begin{cases} \sqrt{1+(o_{\max}-1)\frac{\min\left(\max\left(10\log_{10}\frac{\bar{P}_y(l)}{\bar{P}_\nu(l)+\varepsilon_\nu}, SNR_{\text{maxo}}\right), SNR_{\text{mino}}\right)-SNR_{\text{mino}}}{SNR_{\text{maxo}}-SNR_{\text{mino}}}} \\ \qquad \text{for } \bar{P}_\nu(l) > 0 \text{ and } k = 0,\ldots,K-1 \\ \\ 1 \qquad \text{for } \bar{P}_\nu(l) = 0 \text{ and } k = 0,\ldots,K-1 \end{cases} \tag{2.16}$$

where o_{max} is the maximum subtraction factor value which is set to 3. $SNR_{\text{maxo}} = -5$ dB and $SNR_{\text{mino}} = 20$ dB are the maximum and minimum SNR values for the subtraction factor [36]. ε_ν is set to a small value (e.g. 1), when $\bar{P}_\nu(l)$ is zero to avoid infinite values and is zero elsewhere.

The amplitude of the STFT of the inverse-filtered speech signal, as shown in Fig. 8, is then

$$|\widehat{E}(l,k)| = |Y(l,k)|\ G_n(l,k) \tag{2.17}$$

Finally, $\hat{e}[n]$ is obtained from this modified amplitude and the original phase using an Inverse Short-Time Fourier Transform (ISTFT) via the overlap-add method

$$\hat{e}[n] = \sum_l \sum_{k=0}^{K-1} \widehat{E}(l,k)\bar{u}(n-lR)e^{i\frac{2\pi}{K}(n-lR)k}, \tag{2.18}$$

where $\bar{u}(n)$ is a synthesis window that is biorthogonal to the analysis window $u(n)$ [10].

2.3 Residual Reverberation Reduction

The signal after spectral subtraction for noise reduction can be approximated by

$$\hat{e}[n] \approx s[n] \star h_{eq}[n]. \tag{2.19}$$

The equalized impulse response is a delayed impulse-like function that can be modeled as

$$h_{eq}[n] = a_1\delta[n-n_1] + a_2\delta[n-n_2] + ... + a_d\delta[n-n_d] + ... + a_N\delta[n-n_{N_{imp}}] \tag{2.20}$$

where N_{imp} is the length of the impulse response, and a_i is the amplitude of the reflection arriving after a delay of n_i samples. The direct signal has amplitude a_d (maximum value) and delay n_d. The replicas arriving before the direct signal (n_i for $i < d$) are called pre-echoes and those arriving after the direct signal (n_i for $i > d$) are called late impulse components. The pre-echoes and the direct signal are called early impulse components. As these components are assumed to be uncorrelated with the late impulse components, the late reverberation can be mitigated using spectral subtraction.

2.3.1 Reduction of Late Impulse Effects

Fig. 2.3 shows that the spectral subtraction involves calculating the spectral weights followed by multiplication with the STFT of the input signal. In order to calculate the optimal weights, the STPSD of the late impulse response components must be estimated. The estimation method is given below.

Estimation of the STPSD of the Late Impulse Response Components

The STPSD of $\hat{e}[n]$ can be expressed as [17]-[19]

$$P_{\hat{e}}(l,k) \approx P_{early}(l,k) + P_{late}(l,k), \tag{2.21}$$

where $P_{early}(l,k)$ and $P_{late}(l,k)$ are the STPSDs of the early and late impulse response components, respectively. Generally, the STPSD of the late components can be approximated as a smoothed and shifted version of the STPSD of the inverse filtered

speech [19]

$$\widehat{P}_{late}(l,k) = \gamma w[l-D] * |\widehat{E}(l,k)|^2. \tag{2.22}$$

where γ is a scale factor denoting the relative strength of the late impulse components (set to 0.32), and $w[n]$ is a weight (smoothing) function which is delayed by D samples. The short-time speech spectrum is obtained with a Hamming window with a frame length of 16 ms and a frame shift of 8 ms. Assuming a 50 ms delay between the early and late impulses and considering the frame shift of 8 ms for FFT analysis, the delay D in (2.22) is set to 7.

Weight function: The weight function $w(n)$ was previously considered to be a fixed Rayleigh distribution that provides a reasonable match to the shape of the equalized impulse response [19]. However, setting $w(n)$ to a fixed function which does not depend on the RIR can be inaccurate and thus unsuitable for the equalized impulse response $h_{eq}[n]$. It is better to utilize a weight function which is based on the equalized impulse response to estimate the STPSD of the late components. Our algorithm provides a weight function which depends on the input speech signal $\hat{e}[n]$, and hence on $h_{eq}[n]$.

The weight function is used to approximate the late components through a weighted delayed version of $\hat{e}[n]$. Considering the D frames around each frame as the desired signal and the frame shift of 8 ms, the duration of the weight function is limited to N samples where

$$N \leq \frac{\text{RT60 (ms)}}{8} - D. \tag{2.23}$$

This is because the duration of the RIR and thus the equalized impulse response is approximated by RT60, therefore when using block based processing with a frame shift of 8 ms, the number of previous blocks incorporated in the current frame should be less than RT60 (ms)/8. In addition, the D frames around each frame are considered to be the desired signal and so should not be included. For high reverberation times, the index of the direct component n_d is higher, so N should be chosen much less than the upper bound in (2.23) so that the STPSD of the late impulse components is not overestimated. Based on our extensive experimental results, it was found that $N = 18$ provides good dereverberation performance for a range of reverberation conditions.

In contrast to the fixed weight function in [19] which is unrelated to the input speech signal, our algorithm generates a weight function by averaging the correlation of the input speech signal spectra in different frequency bins. The weight function

values are then

$$w[n] = \frac{|w'[n]|}{\sum_i |w'[i]|}, \quad n = 1, 2, \ldots, N \tag{2.24}$$

where

$$w'[n] = \frac{1}{K(L_f - n - D)} \sum_{k=1}^{K} \sum_{l=n+D+1}^{L_f} \frac{\widehat{E}(l,k)\widehat{E}^*(l-n-D,k)}{|\widehat{E}(l-n-D,k)|^2}, \tag{2.25}$$

L_f and K refer to the number of time frames and frequency bins, respectively, and $|.|$ denotes absolute value. Note that $w'[n]$ is a complex function. This weight function is similar to that introduced in [40].

Spectral Subtraction

The enhanced speech signal is obtained by subtracting the estimated STPSD of the late impulse response components from the input speech signal. The magnitude of the enhanced speech spectra is acquired by filtering in the frequency domain which gives

$$|\widehat{S}(l,k)| = |\widehat{E}(l,k)|G_{r1}(l,k), \tag{2.26}$$

where $G_{r1}(l,k)$ is the spectral weight for filtering given by

$$G_{r1}(l,k) = \left(\frac{|\widehat{E}(l,k)|^2 - \widehat{P}_{late}(l,k)}{|\widehat{E}(l,k)|^2} \right)^{\kappa} \tag{2.27}$$

$$= 1 - \frac{1}{(\zeta_1(l,k))^{\kappa}}, \tag{2.28}$$

with

$$\zeta_1(l,k) = \frac{|\widehat{E}(l,k)|^2}{\widehat{P}_{late}(l,k)}. \tag{2.29}$$

Thus $G_{r1}(l,k)$ depends on an estimate of the *a posteriori* Signal to Reverberation Ratio (SRR) given by $\zeta_1(l,k)$. The parameter κ can be fixed for all frames and frequency bins at a nominal value of 0.5. Note that increasing κ can further reduce the residual late impulses, but it can also introduce undesirable distortion. This distortion is related to the SRR of the speech frame, so κ can be increased in low SRR regions that are mainly reverberation, but kept small when the frame is mainly speech (high SRR). In order to keep the proposed method simple, we first obtain the

enhanced speech with a fixed value of $\kappa = 0.5$ and directly use the resulting enhanced speech signal $\widehat{S}(l,k)$ to determine if speech is present. The ratio of the power of $\widehat{S}(l,k)$ and the input signal $\widehat{E}(l,k)$ can be used as an indicator of the presence of speech in the current frame l

$$\lambda(l) = \frac{\sum_{k=0}^{K-1} |\widehat{S}(l,k)|^2}{\sum_{k=0}^{K-1} |\widehat{E}(l,k)|^2} \qquad 0 \leq \lambda(l) \leq 1. \tag{2.30}$$

If the frame is mainly speech (high SRR), $\lambda(l) \approx 1$. On the other hand, late reverberation reduction will strongly attenuate the input signal in low SRR regions or during speech pauses so that $\lambda(l) \approx 0$. Since κ should be chosen based on the SRR, we use the following decision-directed estimator for determining $\kappa(l)$ in each frame

$$\kappa(l) = \alpha_\kappa \kappa(l-1) + (1-\alpha_\kappa)((1-\lambda(l))(\kappa_{\max} - \kappa_{\min}) + \kappa_{\min}), \tag{2.31}$$

where α_κ is the forgetting factor set to 0.9, and $\kappa_{\max}$ and $\kappa_{\min}$ are the maximum and minimum values of $\kappa(l)$ set to 1 and 0.5, respectively.

As overestimation of the STPSD of the late impulse components may produce values of $\widehat{S}(l,k)$ which are very small or even negative, so the enhanced speech spectra should be limited using a threshold [19]. In addition, spectral subtraction creates small, isolated peaks in the spectrum which occur randomly in time and frequency and sound like frequency tones that change randomly from frame to frame. Thus the resulting speech signal suffers from musical noise [36]. This common problem with spectral subtraction for noise or reverberation reduction has been addressed in the literature. We employ two modifications which have recently been introduced [14].

The first modification for limiting musical noise is to use the *a priori* SRR $\xi_1(l,k)$ to calculate the *a posteriori* SRR

$$\zeta_1(l,k) = 1 + \xi_1(l,k). \tag{2.32}$$

The modified spectral weight for filtering is then

$$G_{r1}(l,k) = 1 - \frac{1}{\sqrt{1+\xi_1(l,k)}}. \tag{2.33}$$

Since the first 50 ms of reverberant speech is perceived as part of the direct speech signal [19], the enhanced speech spectra is equal to the input speech spectra during

this time. Thus the STPSD of the late impulse components for the first D frames is considered to be zero

$$\widehat{P}_{late}(l,k) = 0 \qquad \text{for } 1 \leq l \leq D, \tag{2.34}$$

and the *a priori* SRR is estimated using a decision-directed approach as in (2.11)

$$\xi_1(l,k) = \begin{cases} \beta\xi_1(l-1,k) + (1-\beta)(\max\{\zeta_1(l,k) - 1, \varepsilon\}) \\ \qquad\qquad\qquad\qquad \text{for} \quad l \geq D+3 \\ |\widehat{S}(l,k)|^2 / \widehat{P}_{late}(l,k) \\ \qquad\qquad\qquad\qquad \text{for} \quad D < l < D+3 \end{cases} \tag{2.35}$$

Three frames are added (giving $D+3$), to avoid infinity values in the *a priori* SRR for frames close to the first D frames, which have zero STPSD for the late impulse components. β is the forgetting factor set to 0.5, and ε is the *a priori* SRR threshold set to 0.0663.

The second modification to avoid musical noise is the use of a spectral floor, which confines the enhanced speech spectra above a threshold $\varsigma|\widehat{E}(l,k)|$, where ς is the spectral floor factor which is set to 0.02. Therefore we have

$$|\widehat{S}(l,k)| = \max\{|\widehat{E}(l,k)|G_{r1}(l,k), \varsigma|\widehat{E}(l,k)|\}. \tag{2.36}$$

The enhanced speech signal $\hat{s}[n]$ is calculated using the enhanced magnitude spectrum $|\widehat{S}(l,k)|$ and the original phase. This phase is obtained from the phase of the input speech signal $\hat{e}[n]$ and is used to obtain the enhanced speech signal by using the overlap-add technique followed by an ISTFT (as described in Section 2.2).

2.3.2 Reduction of the Pre-echo Effects

Inverse filtering can produce pre-echo components which introduce annoying temporal characteristics which deteriorate the speech quality. Thus speech enhancement using inverse filtering as the first-stage should incorporate an effective algorithm to reduce the pre-echo effects, especially in high reverberation environments. In this section, we propose a simple spectral subtraction based algorithm to deal with this problem.

Estimation of the STPSD of the Pre-echo Components

Assuming that the STPSD of $\hat{s}[n]$, denoted by $P_{\hat{s}}(l,k)$, is an estimate of the STPSD of the early impulse response components of $\hat{e}[n]$, denoted by $P_{early}(l,k)$, we have

$$P_{\hat{s}}(l,k) \approx P_{early}(l,k) = P_{direct}(l,k) + P_{preecho}(l,k), \tag{2.37}$$

where $P_{direct}(l,k)$ and $P_{preecho}(l,k)$ are the STPSD of the direct path and pre-echo components of $\hat{e}[n]$, respectively. Similarly, the STPSD of the pre-echo components can be approximated as a smoothed and shifted version of the STPSD of the enhanced speech signal which is given by

$$P_{preecho}(l,k) = \gamma \sum_{i=0}^{N-1} w(i)\widehat{S}(l+i+D,k), \tag{2.38}$$

where the parameters are the same as those in (2.22). The weight function is obtained using (2.24).

Spectral Subtraction

The final speech signal is obtained by subtracting the estimated STPSD of the pre-echo components from the enhanced speech signal $\hat{s}[n]$. The magnitude of the final speech spectra is obtained by a filtering operation in the frequency domain given by

$$|\widetilde{S}(l,k)| = |\widehat{S}(l,k)|G_{r2}(l,k), \tag{2.39}$$

where the spectral weight for filtering is

$$G_{r2}(l,k) = \left(\frac{|\widehat{S}(l,k)|^2 - \widehat{P}_{preecho}(l,k)}{|\widehat{S}(l,k)|^2} \right)^{0.5} \tag{2.40}$$

$$= 1 - \frac{1}{(\zeta_2(l,k))^{0.5}} \tag{2.41}$$

with

$$\zeta_2(l,k) = \frac{|\widehat{S}(l,k)|^2}{\widehat{P}_{preecho}(l,k)} = 1 + \xi_2(l,k). \tag{2.42}$$

$\xi_2(l,k)$ is the *a priori* SRR. As before, the STPSD of the pre-echo components for the last D frames is considered to be zero

$$P_{preecho}(l,k) = 0 \qquad \text{for } L_f - D \leq l \leq L_f,$$

where L_f is the number of speech frames. The *a priori* SRR is estimated using a decision-directed approach as

$$\xi_2(l,k) = \begin{cases} \beta\xi_2(l+1,k) + (1-\beta)(\max\{\zeta_2(l,k) - 1, \varepsilon\}) \\ \qquad \text{for} \quad l \leq L_f - D - 3 \\ \\ |\widetilde{S}(l,k)|^2 / \widehat{P}_{preecho}(l,k) \\ \qquad \text{for} \quad L_f - D - 3 < l < L_f - D \end{cases} \tag{2.43}$$

The final enhanced speech is then given by

$$|\widetilde{S}(l,k)| = \max\{|\widehat{S}(l,k)|G_{r2}(l,k), \varsigma|\widehat{S}(l,k)|\}, \tag{2.44}$$

where the parameters are the same as those defined in (2.35) and (2.36).

Reducing the residual reverberation effects, namely the pre-echo components, by spectral subtraction after reduction of the late-impulse effects may introduce undesirable distortion due to overestimation of $P_{preecho}(l,k)$, especially when the reverberation time is not high. To limit this distortion, we use some simple criteria to ensure that spectral subtraction is not used a second time. The normalized cross correlation ϕ_j is used as a measure of the similarity between signal frames

$$\phi_{l,j} = \frac{\sum_{k=1}^{K} \widehat{S}(l,k)\widehat{S}^*(l+j,k)}{\sqrt{\sum_{k=1}^{K} |\widehat{S}(l,k)|^2 \sum_{k=1}^{K} |\widehat{S}(l+j,k)|^2}}. \tag{2.45}$$

The energy for each frame is defined as

$$E_l = \frac{1}{K} \sum_{k=1}^{K} |\widehat{S}(l,k)|^2. \tag{2.46}$$

There are two cases when $|\widehat{S}(l,k)|$ is kept unchanged. First, it is not changed when

$\phi_{l,D+1} \geq \phi_{thr}$ and $|E_{l+D+1} - E_l| < E_{thr}$, i.e.

$$|\widetilde{S}(l,k)| = |\widehat{S}(l,k)| \text{ if } \phi_{l+D+1} \geq \phi_{thr} \text{ and } |E_{l+D+1} - E_l| < E_{thr}$$

where $\phi_{thr} = 0.1 - 0.4$[2] and $E_{thr} = 2$ are the thresholds for frame similarity and frame energy difference, respectively. These conditions are typically satisfied when there are long, frequent speech components (voiced segments), as a result of prolongated phonemes. Second, $|\widehat{S}(l,k)|$ is kept unchanged when the frame energy is less than an energy floor E_{min} so that

$$|\widetilde{S}(l,k)| = |\widehat{S}(l,k)| \qquad \text{if} \quad E_l < E_{min}. \tag{2.47}$$

The energy floor is set to $E_{min} = 0.06$. After calculating $|\widetilde{S}(l,k)|$, the final speech signal is obtained using this spectrum and the original phase by applying the overlap-add technique followed by an ISTFT.

In contrast to noisy conditions, the phase of the strong spectral components is greatly distorted in reverberant environments [19]. Thus, in this case phase correction is as important as magnitude correction. Although the second stage of the proposed method cannot compensate for the phase distortion (mainly caused by reverberation), the first stage provides this compensation. However, the two-stage method in [19], as with other single-microphone methods, cannot compensate for this distortion in highly reverberant conditions. As a result, the speech enhancement is much better with the proposed approach, as will be shown in the next section.

2.4 Performance Results

In this section, we evaluate our proposed method (prop) and compare it with the technique in [19] (Wu) and the temporal and spectral processing method presented in [17] (LP). This is done using 20 s segments of clean speech (for four male and four female speakers), from the TIMIT database which are sampled at 16 kHz. The simulated RIRs are constructed using the image method [3]. The speech signals are assumed to have been received by an omnidirectional microphone placed in a rectangular room with dimensions $[5 \times 4 \times 6]$ (m). All six wall surfaces of the room are assumed to have the same reflection coefficient. We first examine the performance

[2]For low reverberation times it is better to use a lower value, e.g. 0.1, to limit the possibility of distortion.

of our method in reverberant environments free from noise. Then, our denoising algorithm is evaluated in reverberant environments. Finally, our method is evaluated in noisy conditions, including real recorded noise, with the reverberation intensity fixed at a sufficiently high level, i.e., a reverberation time of RT60 = 1 s and a speaker-microphone distance of $d = 2$ m.

Four measures are used to evaluate the performance. The Segmental Signal-to-Interference Ratio (SegSIR) is a measure of the distortion caused by interference (reverberation and noise) in the time domain, and hence is a good indicator of the effectiveness of speech enhancement methods [10]. The difference between the clean speech signal of the direct path $s_d[n] = \alpha_d s[n - n_d]$ (see (2.20)), and the enhanced speech signal $\tilde{s}[n]$ can be expressed as [10]

$$\text{SegSIR} = \frac{1}{L_b} \sum_{l=0}^{L_b-1} \left(10 \log_{10} \left(\frac{\sum_{n=lR}^{lR+N-1} s_d^2[n]}{\sum_{n=lR}^{lR+N-1} (s_d[n] - \tilde{s}[n])^2} \right) \right), \tag{2.48}$$

where L_b is the number of blocks. Bark Spectral Distortion (BSD) is a perceptual-domain measure of the reduction in colouration and the effects of late reverberation [10]. The BSD is calculated using three steps: critical-band filtering, equal loudness pre-emphasis and phon-to-sone conversion, and is defined as [10]

$$\text{BSD} = \frac{1}{L_b} \sum_{l=0}^{L_b-1} \left(10 \log_{10} \left(\frac{\sum_{k_b=1}^{K_b} (L_{s_d}(l, k_b) - L_{\tilde{s}}(l, k_b))^2}{\sum_{k_b=1}^{K_b} (L_{s_d}(l, k_b))^2} \right) \right), \tag{2.49}$$

where L_{s_d} and $L_{\tilde{s}}$ are the Bark spectra of the direct signal $s_d[n]$ and the enhanced signal $\tilde{s}[n]$, respectively, and k_b is a Bark frequency bin. In order to evaluate the reduction in only colouration caused by early reverberation, we employ segmental LP residual kurtosis, which is a commonly used measure [20] and is given by

$$\text{SegKurt} = \frac{1}{L_b} \sum_{l=0}^{L_b-1} \frac{E\{\bar{\tilde{s}}_l[n]^4\}}{E\{\bar{\tilde{s}}_l[n]^2\}^2}, \tag{2.50}$$

where $\bar{\tilde{s}}_l[n]$ is the LP residual signal of the lth frame of $\tilde{s}[n]$ and $E\{.\}$ denotes expectation. We also consider the Perceptual Evaluation of Speech Quality (PESQ) [10], which employs a perceptual model to assess the quality of a processed speech signal. The PESQ is a recognized estimator for the Mean Opinion Score (MOS) [10]. These four measures are applied on 32 ms frames with a 50% overlap. Finally, subjec-

tive listening tests were performed following the guidelines described in [10]. Twenty listeners were asked to give a score between one and five to evaluate the enhanced speech quality [1 = bad, 2 = poor, 3 = fair, 4 = good, and 5 = excellent]. They were instructed to rate the reduction in distortion caused by reverberation and noise and the overall speech quality. The individual ratings, averaged over all listeners, constitutes the widely used MOS [10]. The original clean speech samples (four females and four males with an average duration of 4 s), were considered as the reference speech signals with a score of 5, while the speech samples under the worst conditions have a score of 1.

2.4.1 Speech Dereverberation in Different Environments

We evaluate the dereverberation methods with two sets of RIRs. One set has a speaker-microphone distance of 2 m and a reverberation time from 200 to 1200 ms, while the other has a speaker-microphone distance of 4 m with the same reverberation times. The results averaged over the 8 utterances are shown in Figs. 2.4-2.7 for the four measures, where "rev", "inv", "Wu", "LP" and "prop" indicate the calculated values for the reverberant speech signals, the inverse-filtered speech signals using the our inverse filtering method presented in Section 2.1, and the processed speech signals using the two-stage method proposed by Wu and Wang [19], the two-stage method proposed in [17] and the proposed two-stage method[3]. The upper plots denote a speaker-microphone distance of $d = 2$ m, and the lower ones a distance of $d = 4$ m.

The SegSIR values in Fig. 2.4 show a significant reduction in reverberation distortion using the proposed two-stage method compared to inverse filtering and the two other methods. The difference between the first-stage method (inverse filtering) and the two-stage method (inverse filtering with spectral subtraction) verifies that the proposed spectral subtraction can effectively reduce the distortion remaining after inverse filtering. The effectiveness of the proposed method compared to that of Wu and Wang is very evident with larger speaker-microphone distances ($d = 4$ m). This is because in this case, the distortion is dominated by early reverberation effects, and the inverse filtering method presented in Section 2.1 is superior in reducing these effects. Fig. 2.5 shows that the BSD is greatly reduced by both inverse filtering and the proposed two-stage method, compared to the approach by Wu and Wang and the

[3]In the noise free case, the spectral subtraction algorithm for denoising described in Section 2.2 is not employed.

spectral-temporal processing method [17], particularly when $d = 4$ m (lower plot). The reduction in early reverberation distortion is clearly evident in Fig. 2.6, but the Wu and Wang method is only effective for low reverberation times (RT60 $\leq$ 400 ms) [19]. Note, however, that our inverse filtering method is much better even in this region. Spectral-temporal processing provides a slightly higher LP residual kurtosis than the original reverberant speech signal, as it is not able to deal with the problem of early reverberation. The PESQ results in Fig. 2.7 indicate that the speech quality is improved by the inverse filtering and proposed methods compared to the others. The improvement provided by our second stage compared to the first stage indicates the effectiveness of the spectral subtraction method in reducing the late reverberation effects while introducing negligible audible artifacts. Finally, the mean opinion score (MOS) results in Fig. 2.4.1 also confirm that our proposed method provides superior speech quality for all reverberation times considered.

We also conducted experiments using four measured binaural RIRs from the Aachen Impulse Response (AIR) database [41]: 1) office, RT60 = 0.66 s, $d = 3$ m; 2) meeting room, RT60 = 0.67 s, $d = 2.8$ m; 3) lecture room, RT60 = 1.23 s, $d = 8.68$ m; 4) stairway, and RT60 = 1.95 s, $d = 3$ m, with an azimuth angle of 30^o. All RIRs were measured without a dummy head using only the right channel [41]. The average of the four objective measures for the 8 utterances are shown in Figs. 2.9-2.12. Fig. 2.9 indicates that the proposed two stage method "prop" successfully decreases the reverberation effects in all four room types, and provides better performance than the other methods in all cases. Fig. 2.10 shows that the colouration and late reverberation effects are mitigated using our inverse filtering method "inv" and the proposed method "prop". Thus our inverse filtering is effective in real situations. Fig. 2.11 (lower plot) demonstrates that these methods can deal with the problem of early reverberation while the Wu and Wang method "Wu" and spectral-temporal processing "LP" provides little improvement. Finally, Fig. 2.12 confirms that better speech quality is obtained using our proposed method in real environments compared to the other methods.

To further illustrate the performance of the proposed dereverberation method, the speech enhancement for a female speaker obtained from the TIMIT database is shown in Fig. 2.15. The reverberant speech is constructed by convolving the clean speech signal with the RIR for RT60 = 1 s and $d = 2$ m, which is shown in Fig. 2.13 (a). Fig. 2.15 shows that the reverberation smears the harmonic structure and temporal properties of the speech signal so that the silent gaps between words are

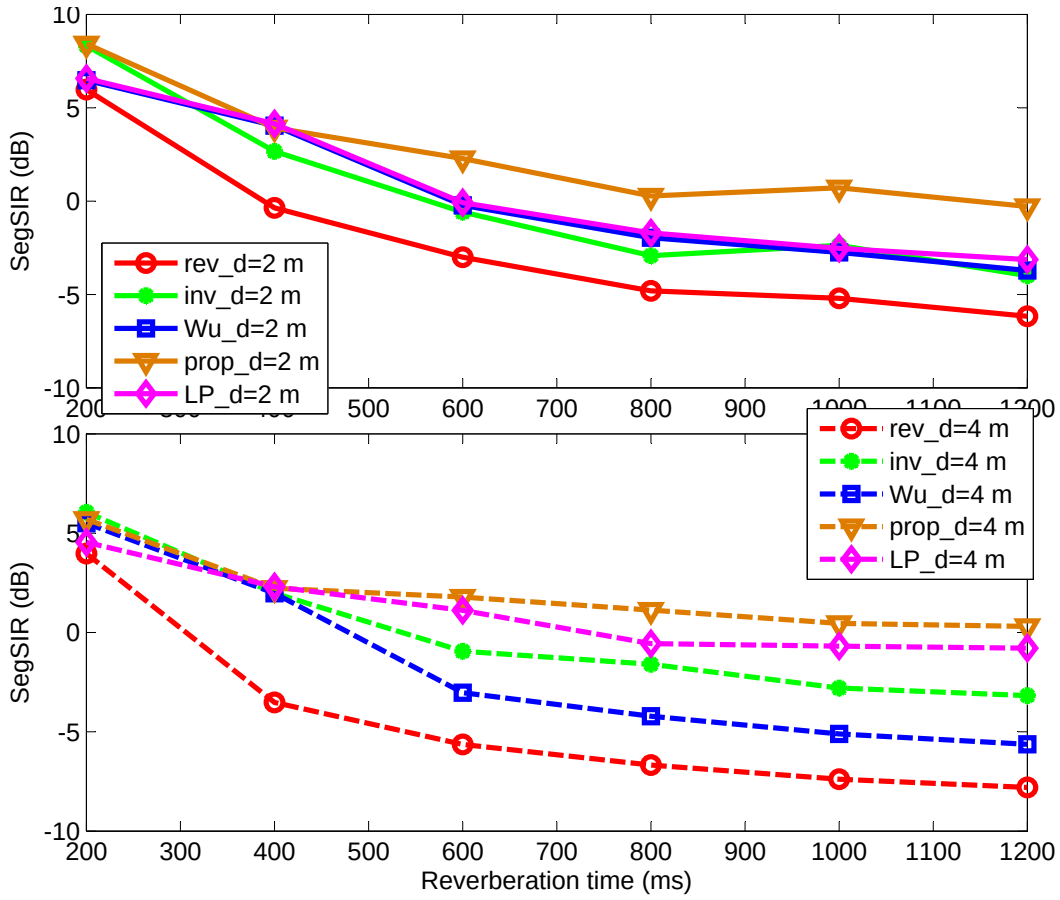

Figure 2.4: SegSIR for different reverberation times, $d = 2$ m (upper plot) and $d = 4$ m (lower plot). "rev", "inv", "Wu", "LP" and "prop" represent the SegSIR for the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

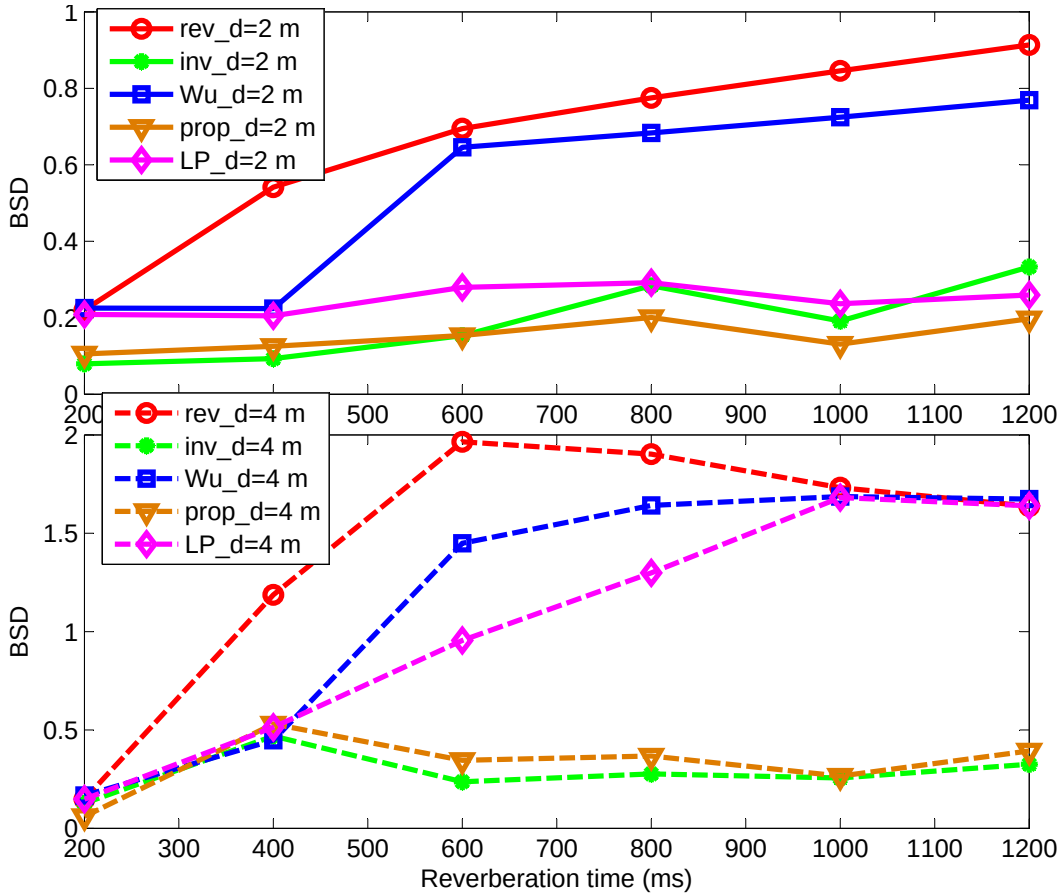

Figure 2.5: BSD for different reverberation times, $d = 2$ m (upper plot) and $d = 4$ m (lower plot). "rev", "inv", "Wu", "LP" and "prop" represent the BSD for the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

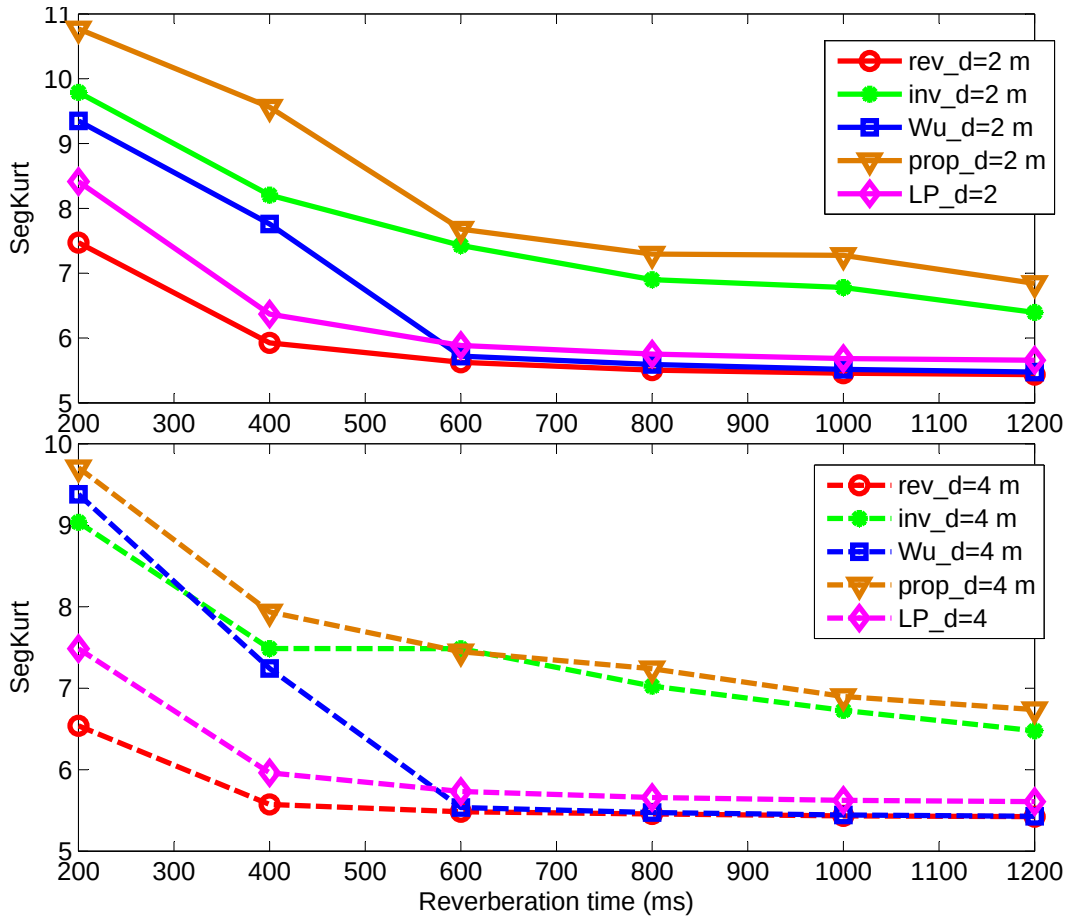

Figure 2.6: LP residual kurtosis for different reverberation times, $d = 2$ m (upper plot) and $d = 4$ m (lower plot). "rev", "inv", "Wu", "LP" and "prop" represent the values for the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

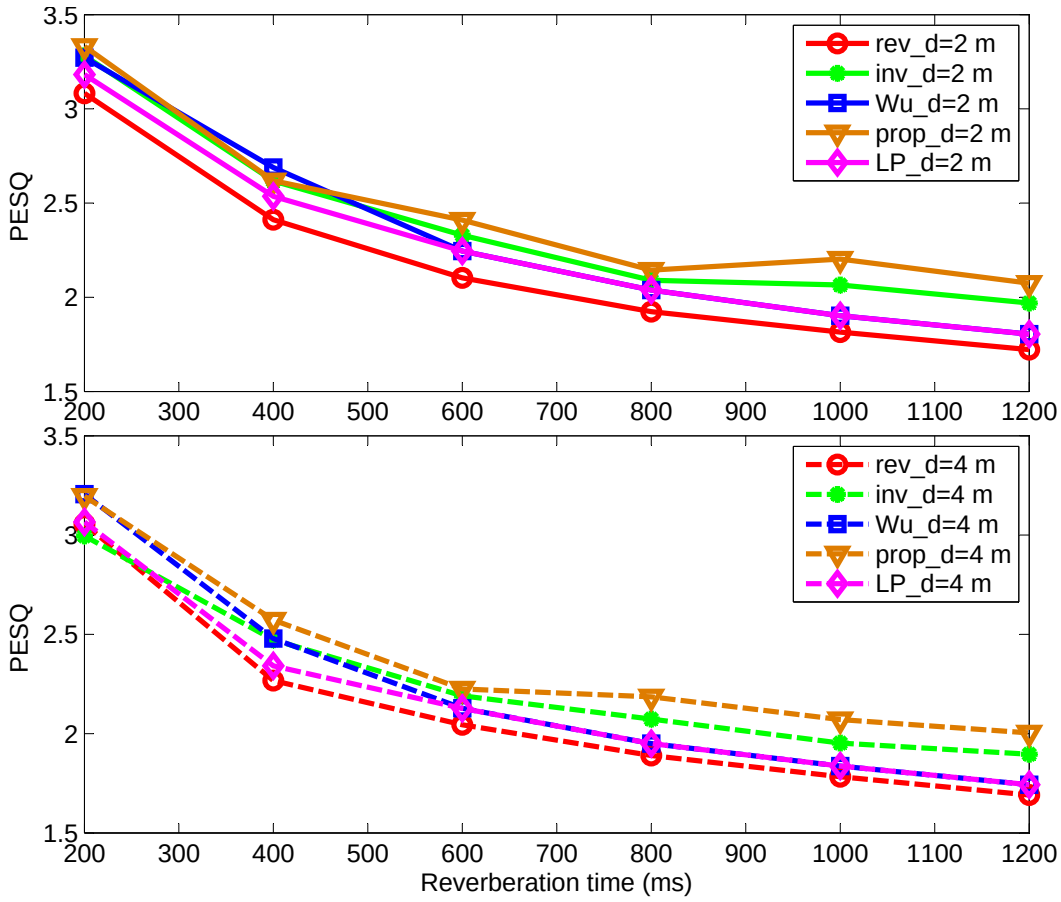

Figure 2.7: PESQ for different reverberation times, $d = 2$ m (upper plot) and $d = 4$ m (lower plot). "rev", "inv", "Wu", "LP" and "prop" represent the PESQ for the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

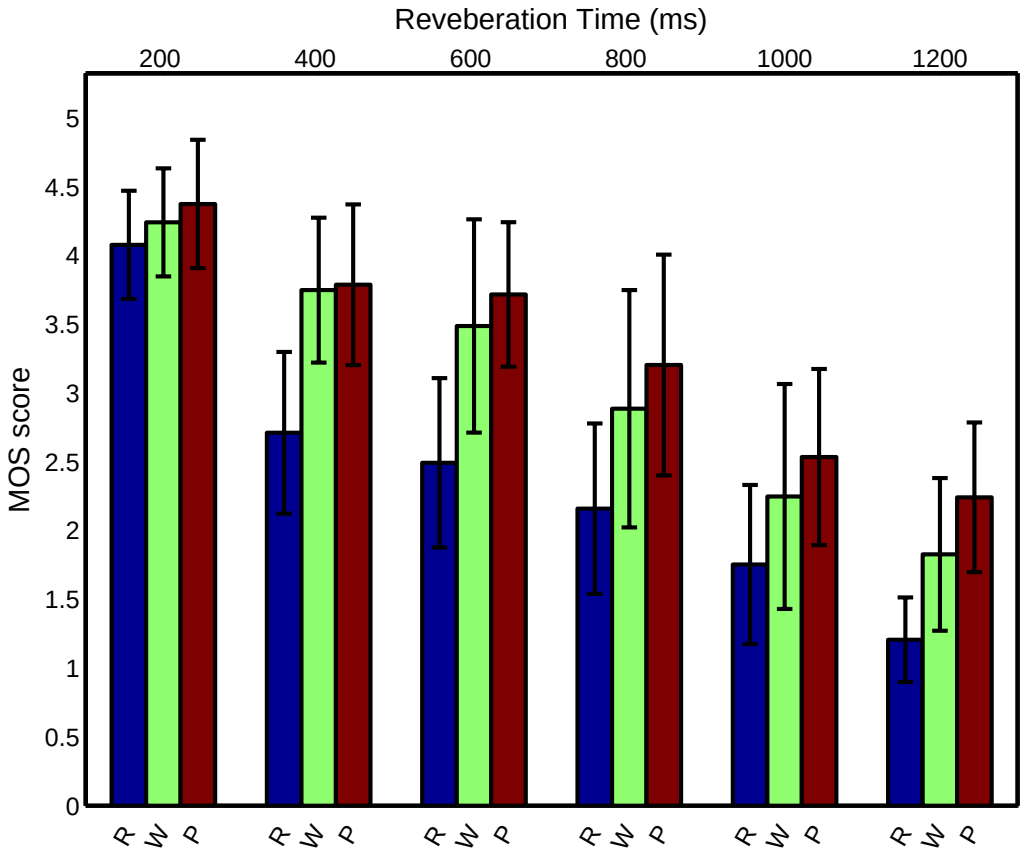

Figure 2.8: MOS scores for different reverberant noise free environments with RT60 = 1 s and $d = 2$ m. "R", "W", and "P" represent the scores for the reverberant speech, and the enhanced speech obtained using the Wu and Wang method [19] and the proposed method, respectively. The variances are indicated by the vertical lines.

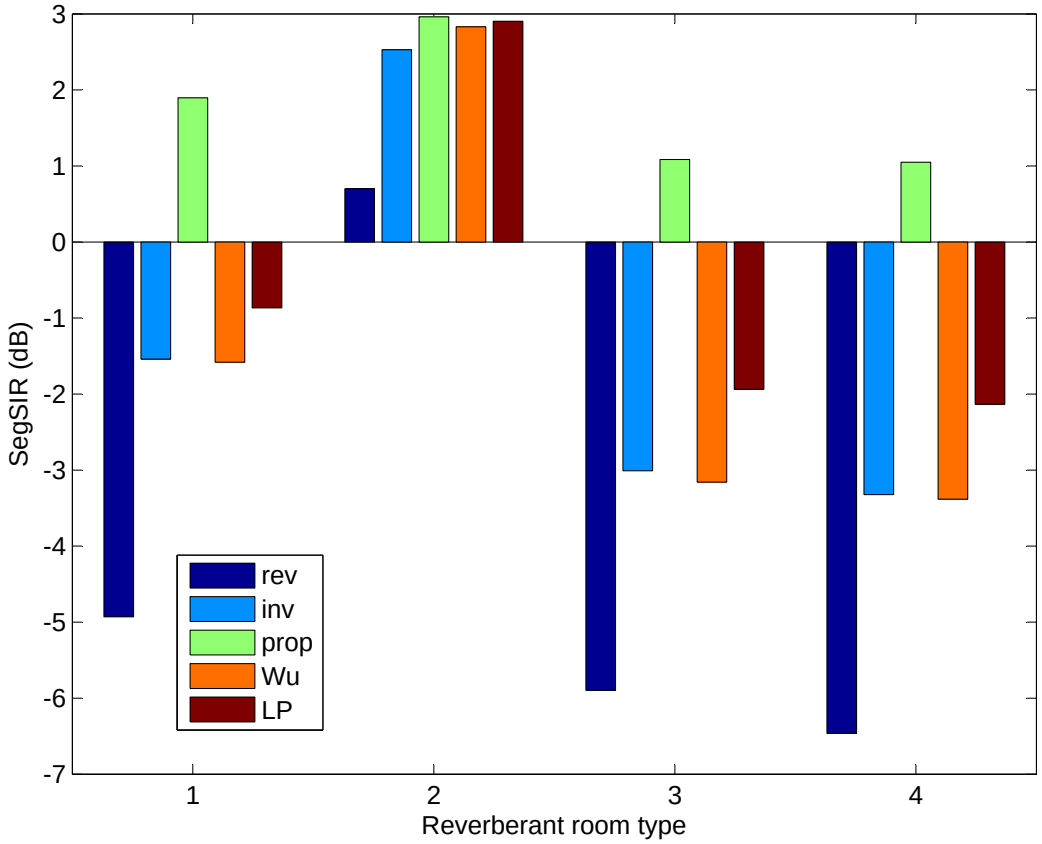

Figure 2.9: SegSIR for four real reverberant environments. "rev", "inv", "Wu", "LP" and "prop" represent the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

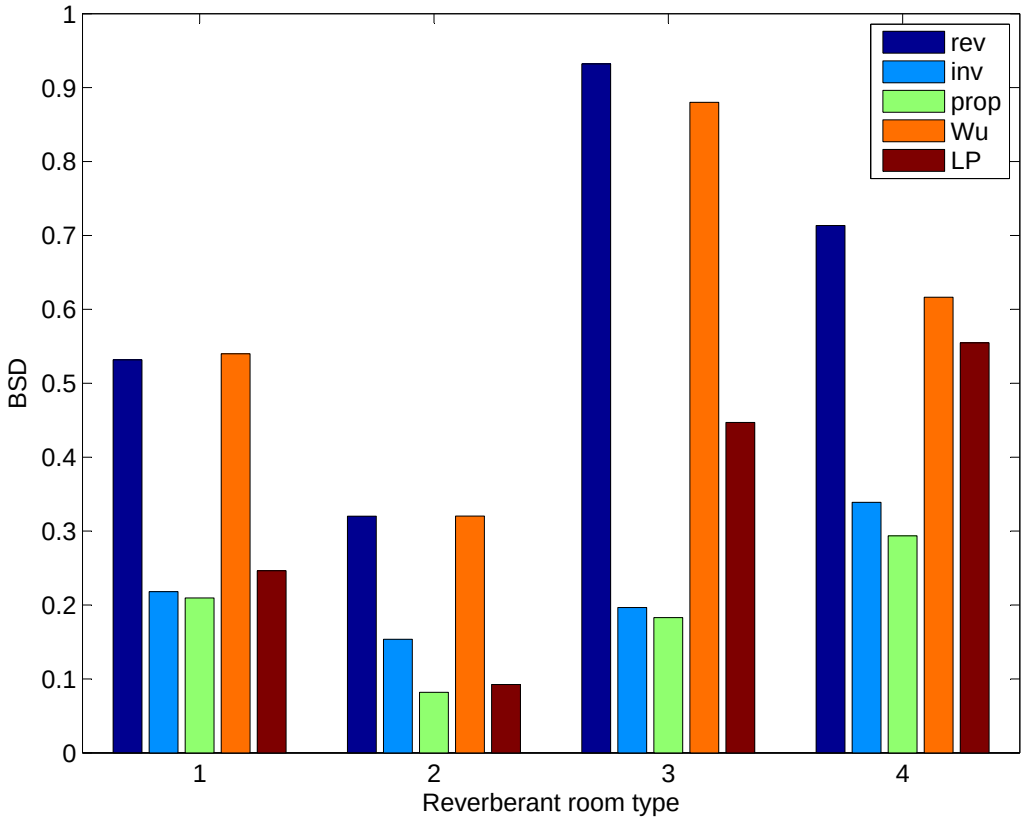

Figure 2.10: BSD for four real reverberant environments. "rev", "inv", "Wu", "LP" and "prop" represent the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], and the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage speech enhancement method.

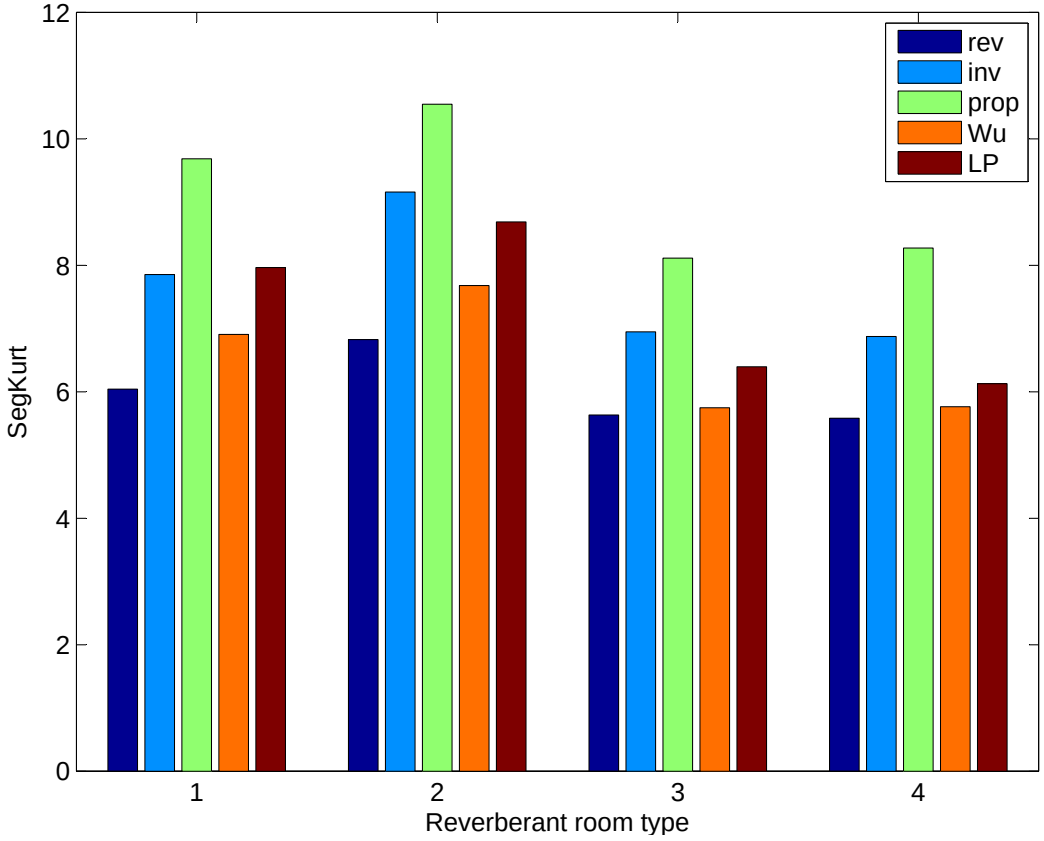

Figure 2.11: LP residual kurtosis for four real reverberant environments. "rev", "inv", "Wu", "LP" and "prop" represent the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

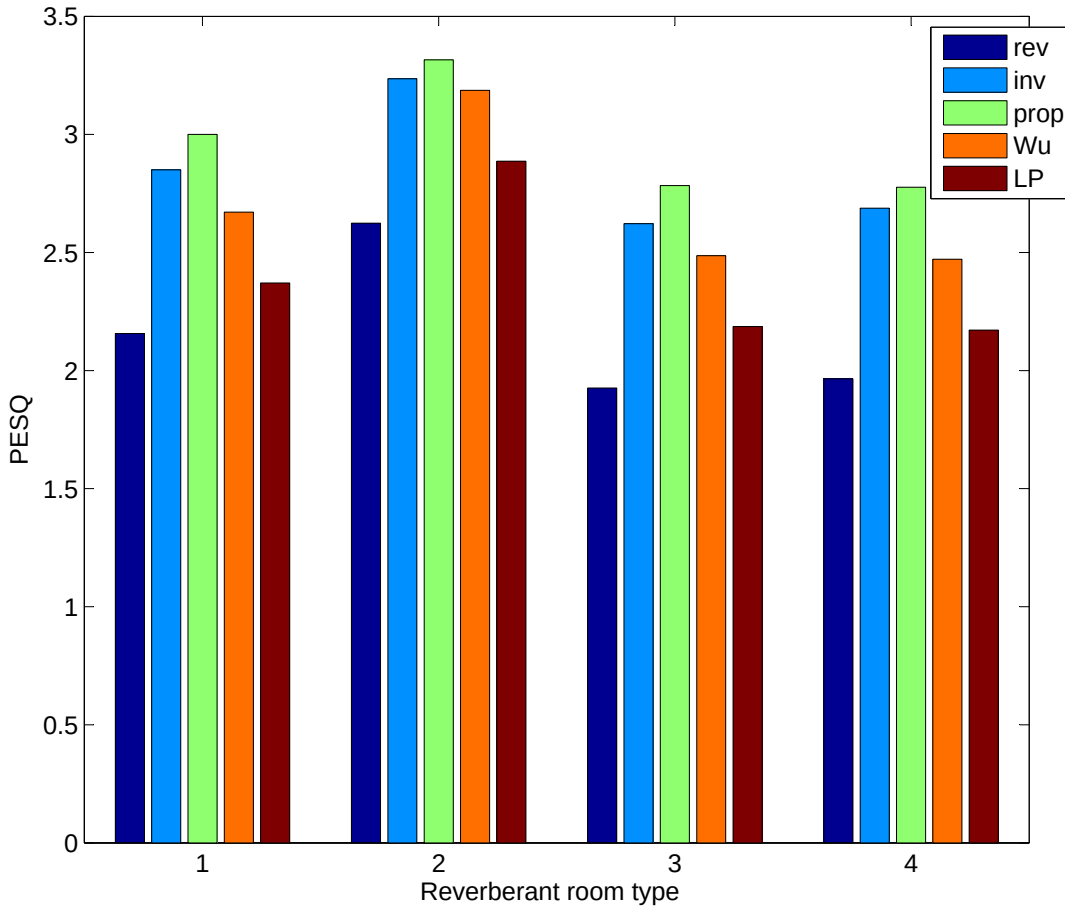

Figure 2.12: PESQ for four real reverberant environments. "rev", "inv", "Wu", "LP" and "prop" represent the reverberant speech, the inverse-filtered speech using the inverse filtering method proposed in [27]-[28], the processed speech using the Wu and Wang method, the method in [17] and the proposed two-stage method.

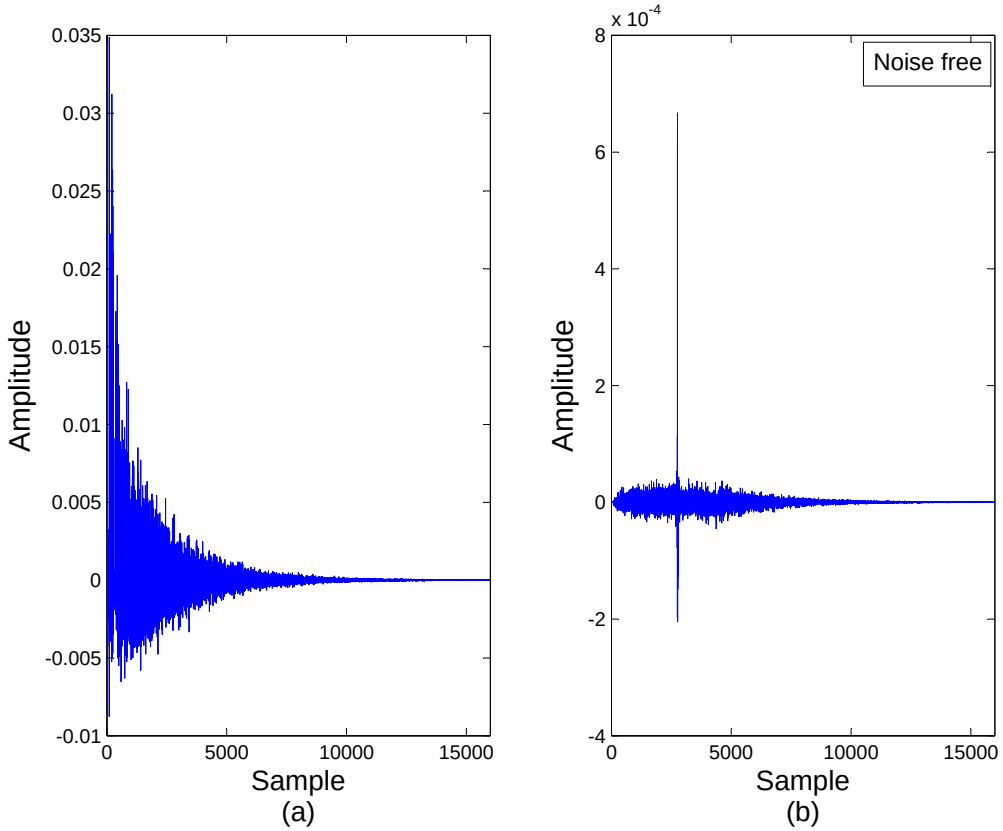

Figure 2.13: (a) RIR with RT60 = 1000 ms and d = 2 m. (b) Equalized impulse response using the inverse filtering method proposed in [27]-[28].

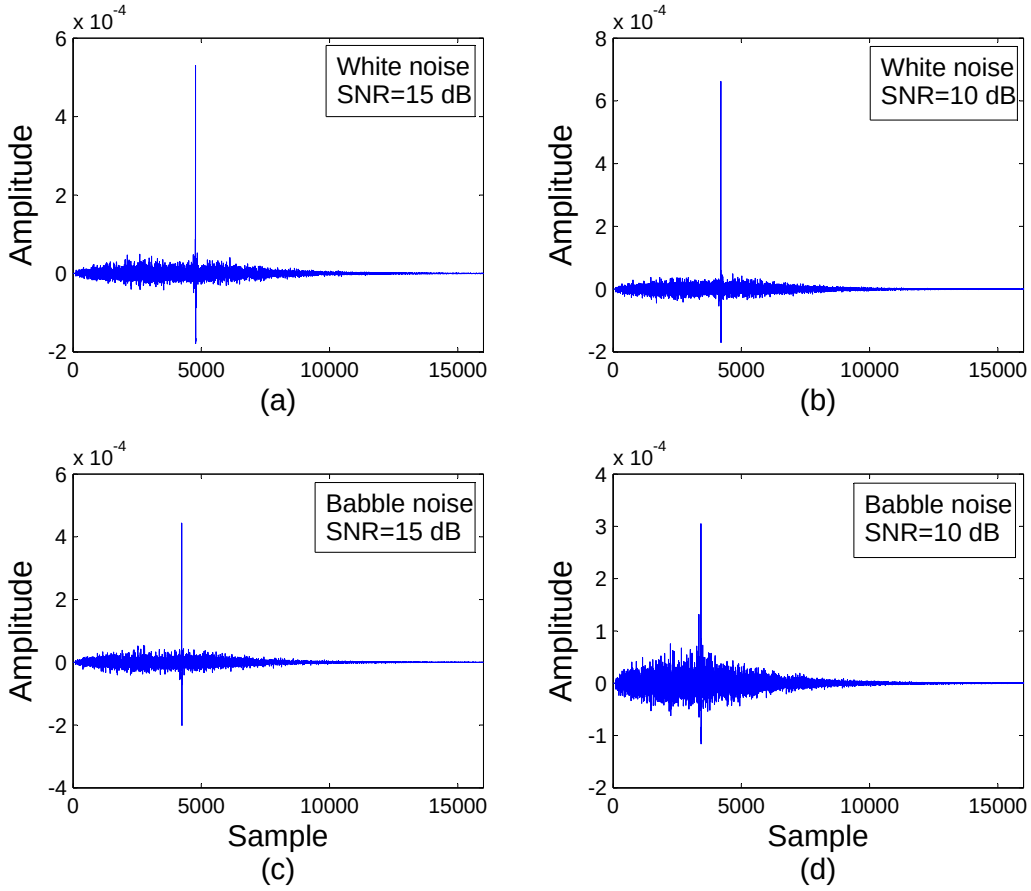

Figure 2.14: Equalized impulse response for RT60 = 1000 ms and $d = 2$ m in different noisy conditions. The upper two are for while Gaussian noise and the lower two are for babble noise. The left two are for SNR=15 dB and the right two are for SNR=10 dB.

filled. In order to show the effectiveness of our proposed pre-echo reduction algorithm, the processed speech is shown using our two-stage method with and without this pre-echo reduction in Figs. 2.15 (g-h) and 2.15 (i-j), respectively. It is clear that without the reduction algorithm, the processed speech contains pre-echo effects which also fill the silent gaps. However, these effects are largely reduced when the reduction algorithm is used. This clearly demonstrates the advantage and effectiveness of our algorithm to reduce the pre-echo effects. Comparing the results of our method in Fig. 2.15 (i-j) with the reverberant speech in Fig. 2.15 (c-d), the reverberation effects are greatly reduced and the harmonic structure is largely restored, with decreased smearing. The corresponding results for the Wu and Wang method given in Fig. 2.15 (e-f) show little improvement in the harmonic structure. This is because using only spectral subtraction for inverse filtering is ineffective for long reverberation times [19]. Thus our two-stage dereverberation method significantly outperforms their approach.

2.4.2 Speech Denoising in Reverberant Environment

In this section, we evaluate the performance of our denoising algorithm in different noisy reverberant conditions. The reverberant speech signals are obtained by convolving the clean speech with the RIR (RT60 = 1 s, $d = 2$ m and $d = 0.5$ m). These signals are then added to two types of additive noise: 1) white computer generated Gaussian noise, and 2) recorded babble noise[4]. The Signal to Noise Ratio (SNR) was varied from -5 to 40 dB. Note that only the denoising algorithm is used in this section in order to evaluate additive noise reduction performance. The denoising algorithm presented in Section 2.2 is compared with the algorithms in [36]-[39]. The Segmental Signal-to-Noise Ratio (SegSNR) as defined in [42] (p. 45, eq. 2.12) and the PESQ are used to evaluate the performance of the denoising algorithms. The SegSNR is clamped to between 35 dB and -10 dB as suggested in [43]. The results averaged over 8 utterances (four male and four female speakers) are shown in Figs. 2.16-2.17 where “noisy”, “prop”, “Berouti”, “Cohen”, “Gusta”, and “Kamath” denote the values for the noisy reverberant speech signal, and the processed speech signal using our denoising algorithm and methods in [36], [37], [38], and [39], respectively.

Fig. 2.16 shows that the proposed denoising algorithm has higher SegSNR values compared with the other algorithms. This demonstrates the effectiveness of the algorithm in reverberant conditions. It is clear that the Berouti algorithm has the

[4][online]. Available: http://www.ee.columbia.edu/ dpwe/sounds/noise/babble.wav

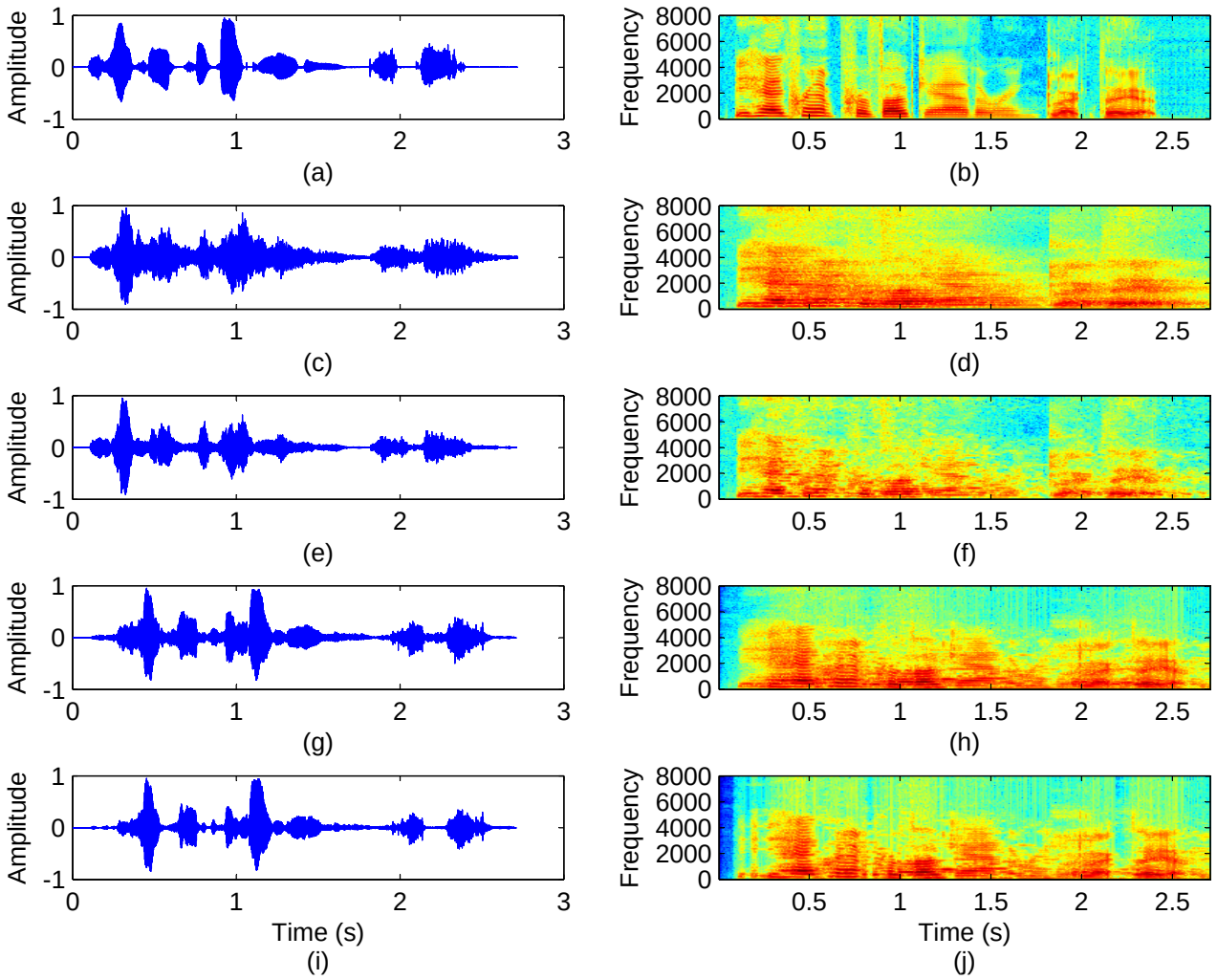

Figure 2.15: Speech signals for RT60 = 1000 ms, $d = 2$ m and SNR = ∞, (a) clean speech, (b) spectrogram of the clean speech, (c) reverberant speech, (d) spectrogram of the reverberant speech, (e) speech processed using the Wu and Wang method, (f) spectrogram of the processed speech using the Wu and Wang method, (g) speech processed using the proposed algorithm with out pre-echoes effect reduction, (h) spectrogram of the processed speech using the proposed algorithm with out pre-echoes effect reduction, (i) speech processed using the proposed algorithm, and (j) spectrogram of the processed speech using the proposed algorithm.

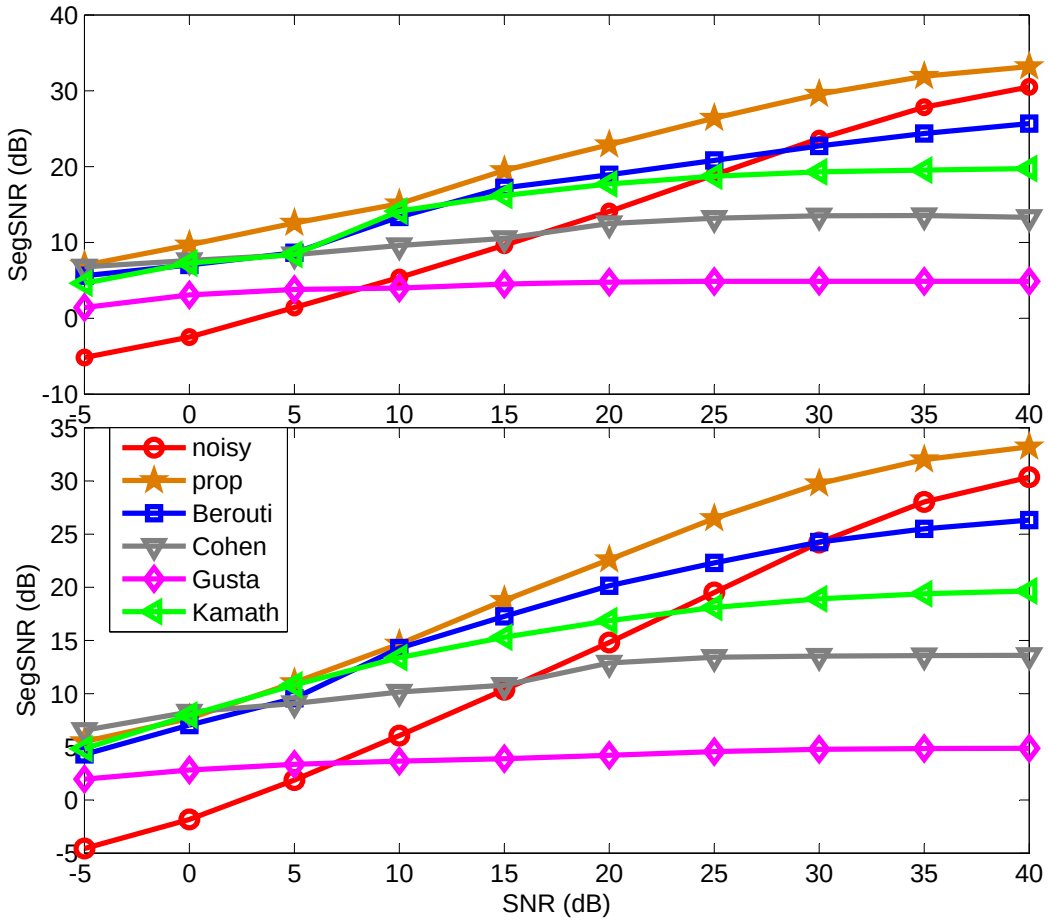

Figure 2.16: SegSNR for different noise conditions with RT60 = 1000 ms $d = 2$ m and $d = 0.5$ m. "noisy", "prop", "Berouti", "Cohen", "Gusta"and "Kamath" represent the SegSNR for the noisy reverberant speech, and the processed speech using our denoising algorithm and the methods in [36], [37], [38], and [39], respectively. The upper plot corresponds to white noise, and the lower to babble noise.

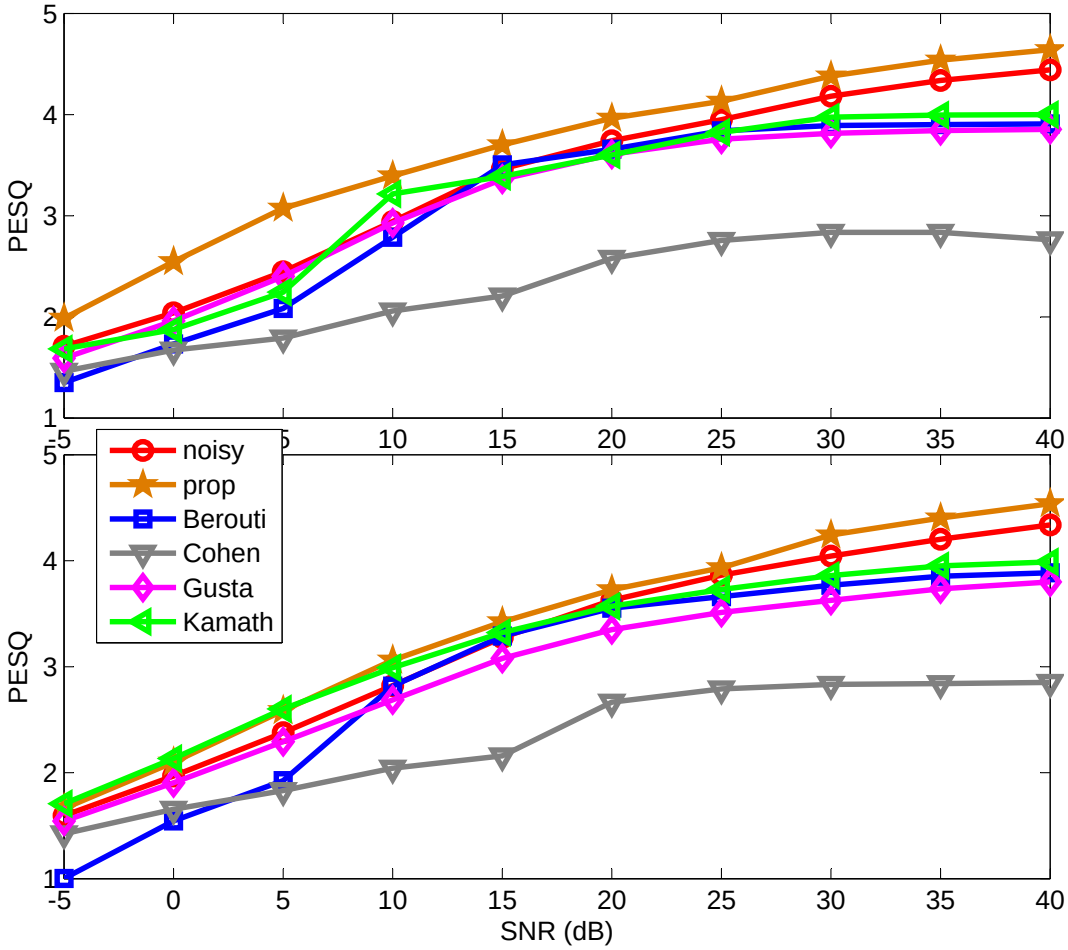

Figure 2.17: PESQ evaluations in different noise conditions with RT60 = 1000 ms $d = 2$ m and $d = 0.5$ m. "noisy", "prop", "Berouti", "Cohen", "Gusta"and "Kamath" represent the PESQ values of the noisy reverberant speech, the processed speech using our denoising algorithm, the one using the method in [36], [37], [38], and [39], respectively. The upper plot corresponds to white noise, and the lower is related to babble noise.

next best performance. This is not surprising as the proposed algorithm is based on the work of Berouti et al. [36]. Fig. 2.17 confirms that better speech quality is obtained with the proposed algorithm in reverberant environments compared to the other methods. Our extensive simulation results in different reverberant conditions show similar results. In addition, audio demonstrations can be found at: `https://sites.google.com/site/derverberation/experimental-data` for noisy reverberant conditions with SNR = 25 dB and an RIR for RT60 = 1000 ms and $d = 0.5$ m. Based on our experiments, it can be concluded that our denoising algorithm under reverberant conditions is superior to the other methods, especially when the noise level is high and the additive noise is nonstationary (babble noise). To further illustrate the performance in low SNR conditions with additive babble noise, the denoising results for a female speaker from the TIMIT database are shown in Fig. 2.18. A clean speech signal was convolved with the RIR (with RT60 = 1000 ms and $d = 0.5$ m), to produce the reverberant speech. This is shown in Fig. 2.18 (a) (speech waveform) and Fig. 2.18 (b) (speech spectrogram). The reverberant speech signal was added to babble noise with SNR = 5 dB resulting in Fig. 2.18 (c-d). The processed speech using the Berouti et al. algorithm [36] is shown in Fig. 2.18 (e-f), and using our denoising algorithm is shown in Fig. 2.18 (g-h). It is clear from this figure that the Berouti et al. algorithm distorts the harmonic structure of the speech signal in this very noisy environment. Conversely, the proposed algorithm successfully suppresses the additive noise while having only a minor effect on the signal when compared with the original reverberant speech.

2.4.3 Reverberant Speech Enhancement in Noisy Conditions

In this section, we evaluate the performance of our speech enhancement method with both noise and reverberation. The noisy reverberant speech signals were produced by convolving the clean speech with the RIR (RT60 = 1 s and $d = 2$ m), and then noise was added. As before, the additive noise is white computer generated Gaussian noise and recorded babble noise. The SNR was varied from -5 to 40 dB. In order to fairly compare our method with that of Wu and Wang [19] and spectral-temporal processing [17], their algorithms were modified by adding the denoising algorithm in Section 2.2 prior to spectral subtraction. The speech signal was convolved with the RIR and then the white or babble noise was added. The inverse filters were obtained using the algorithm proposed in Section 2.1. As an example, the equalized impulse

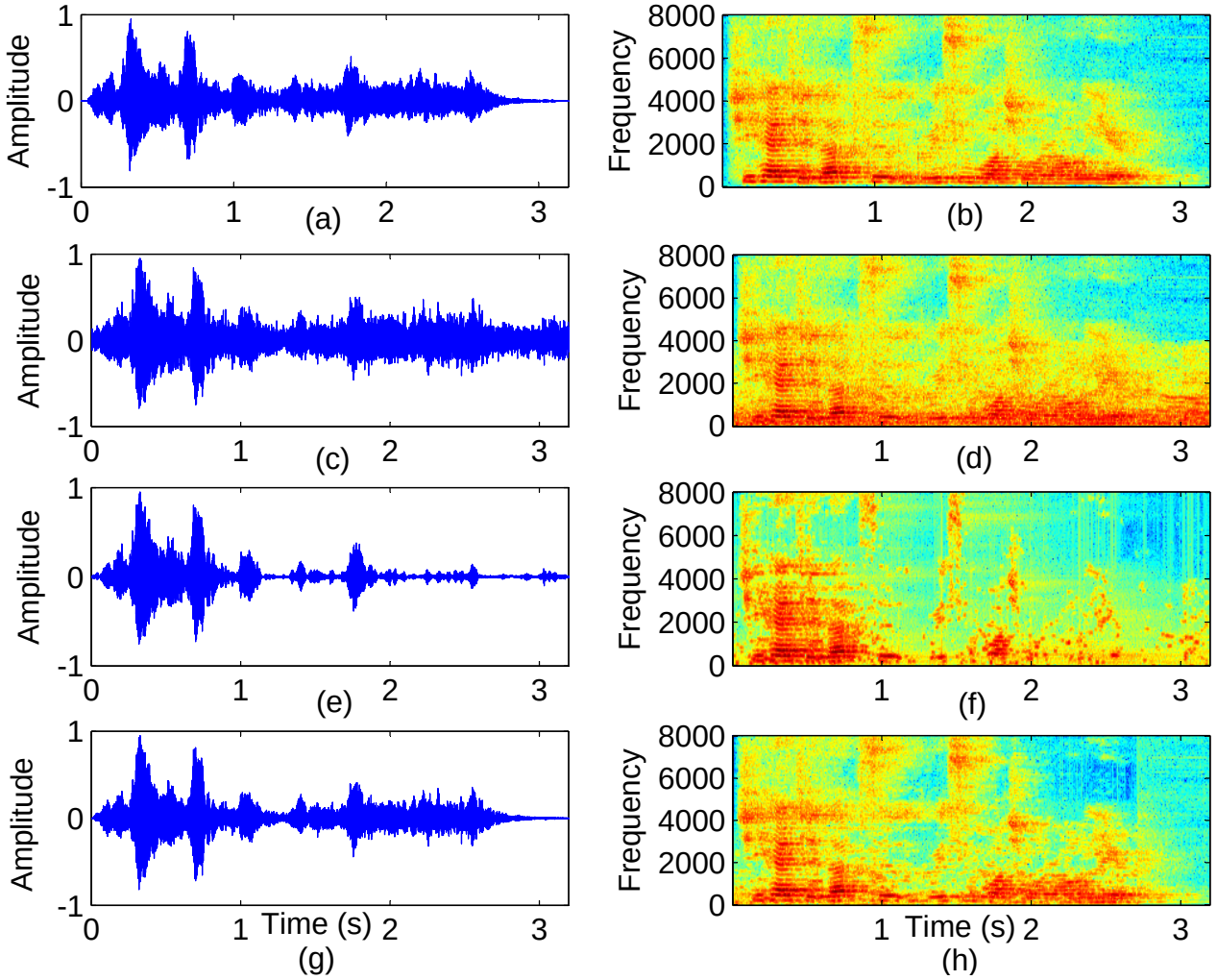

Figure 2.18: Speech signals for RT60 = 1000 ms, $d = 0.5$ m and SNR = 5 dB (babble noise). Reverberant speech (a), spectrogram of the reverberant speech (b), reverberant speech added to babble noise (c), and spectrogram of the noisy reverberant speech (d). Denoising results: speech processed using the Berouti algorithm [36] (e), spectrogram of the processed speech using the Berouti algorithm (f), speech processed using the proposed algorithm (g), and spectrogram of the processed speech using the proposed algorithm (h).

responses are shown for two SNR values 10 dB and 15 dB in Fig. 2.14. These results indicate that the inverse filter can be accurately estimated in low-to-moderate noise conditions (SNR $\geq$ 10 dB).

The four measures discussed previously were used to evaluate the speech enhancement methods. The results averaged over 8 utterances (four male and four female speakers) are shown in Figs. 2.19-2.22 where "received", "Wu", "LP" and "prop" denote the values for the noisy reverberant speech signal, and the processed speech signal using the Wu and Wang method, the spectral-temporal processing algorithm, and the proposed two-stage method. The upper plots correspond to white noise, and the lower to babble noise. The SegSIR improvement in Fig. 2.19 shows that our two-stage method significantly reduces the distortion caused by noise and reverberation. In addition, compared to the other two methods, the performance of our method is better for SNR $\geq$ 10 dB, largely because the first-stage inverse filtering is more effective under these conditions. The BSD in Fig. 2.20 also shows that the proposed method can deal with both noise and reverberation, and reduce the colouration and reverberation tail effects under noisy conditions, whereas the other two method perform poorly, especially for SNR $\geq$ 10 dB. This again shows the usefulness of our skewness-based inverse filtering method as the first stage and the efficiency of the proposed spectral subtraction method in dealing with reverberation in noise. The SegKurt in Fig. 2.21 shows the effectiveness of the methods in removing colouration effects in noisy reverberant conditions. The proposed method is again superior to the other two methods for SNR $\geq$ 10 dB since the inverse filtering method proposed in [27]-[28] works well under these conditions. The most commonly employed perceptual measure, PESQ, is shown in Fig. 2.22. This clearly illustrates the effectiveness of our approach compared with the other methods. Finally, the MOS evaluation for different reverberant noisy conditions (with both white Gaussian noise and babble noise), with an SNR from 5 to 35 dB are shown in Figs. 2.23 and 2.24. So it can be seen from the figures that the proposed method has higher scores in noisy reverberant conditions than the Wu and Wang method. An audio demonstration can be found at `https://sites.google.com/site/derverberation/experimental-data`.

2.4.4 Conclusions

In this chapter, a two-stage single-microphone speech enhancement method was proposed which employs inverse filtering and spectral subtraction. The inverse filtering

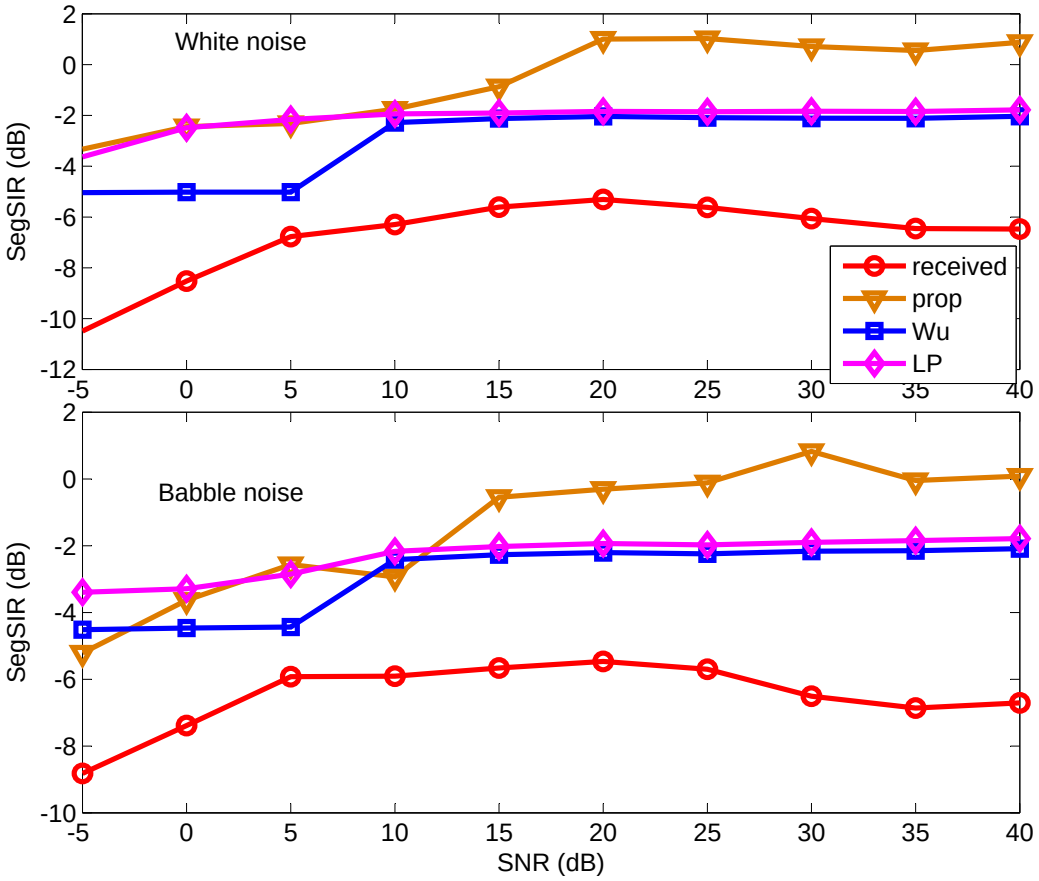

Figure 2.19: SegSIR for different noisy conditions with RT60 = 1000 ms and $d = 2$ m. "received", "Wu", "LP" and "prop" represent the SegSIR of the received speech, and the processed speech using the Wu and Wang method, the spectral-temporal processing method [17], and the proposed method. The upper plot corresponds to white noise, and the lower corresponds to babble noise.

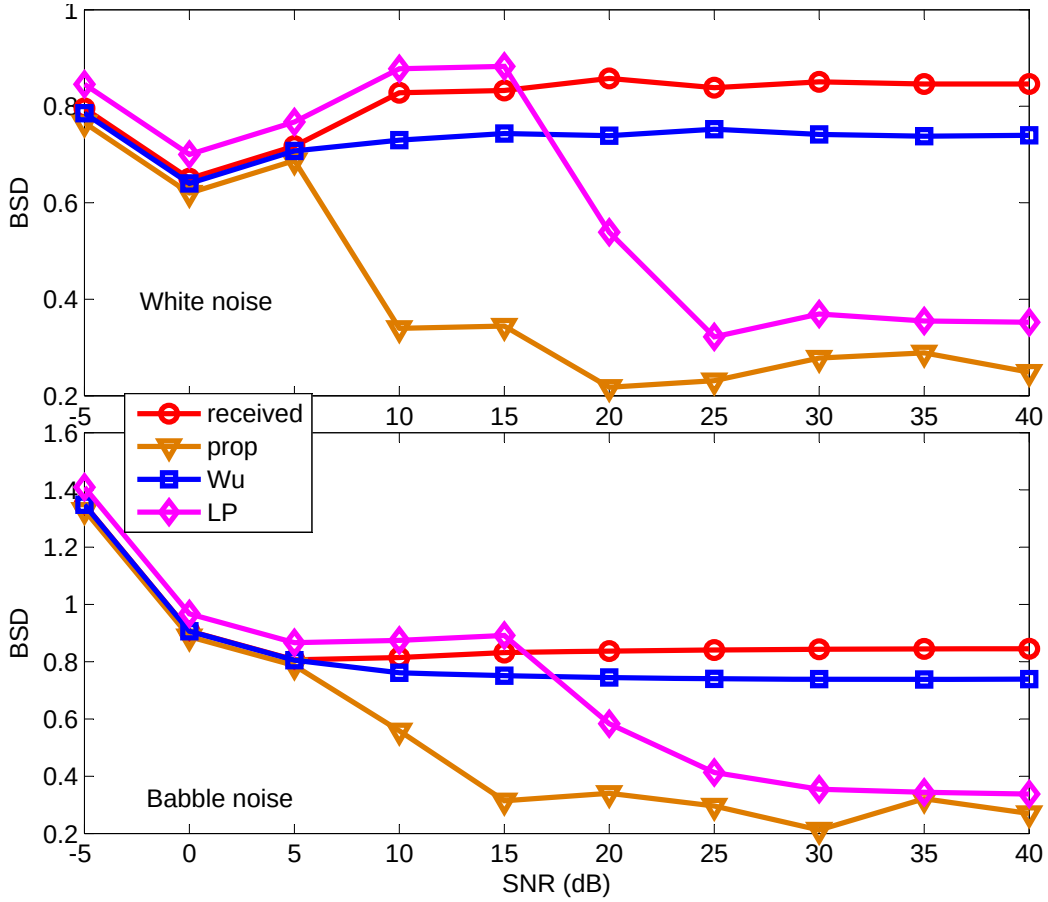

Figure 2.20: BSD for different noisy conditions with R$T60$ = 1000 ms and d = 2 m. "received", "Wu", "LP" and "prop" represent the BSD of the received speech, and the processed speech using the Wu and Wang method, the spectral-temporal processing method [17], and the proposed method. The upper plot corresponds to white noise, and the lower corresponds to babble noise.

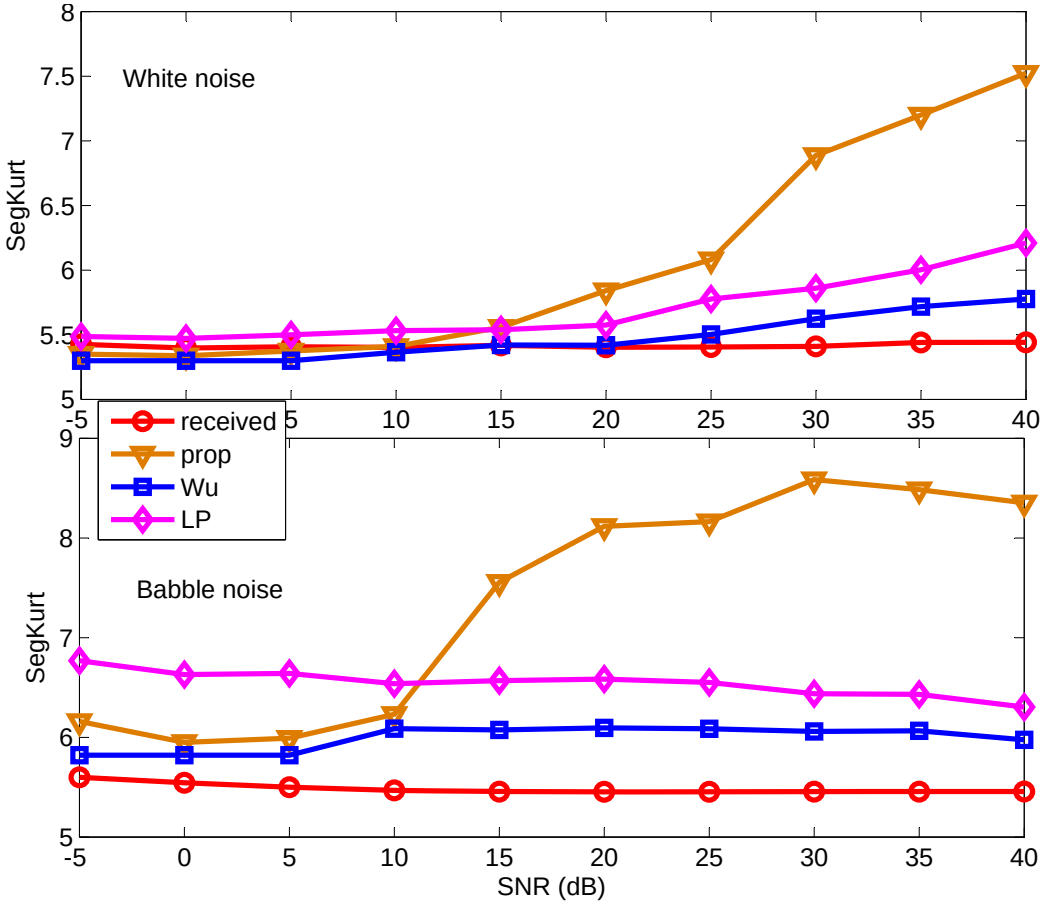

Figure 2.21: LP residual kurtosis in different noisy conditions with RT60 = 1000 ms and $d = 2$ m. "received", "Wu", "LP" and "prop" represent the LP residual kurtosis of the received speech, and the processed speech using the Wu and Wang method, the spectral-temporal processing method [17], and the proposed method. The upper plot corresponds to white noise, and the lower corresponds to babble noise.

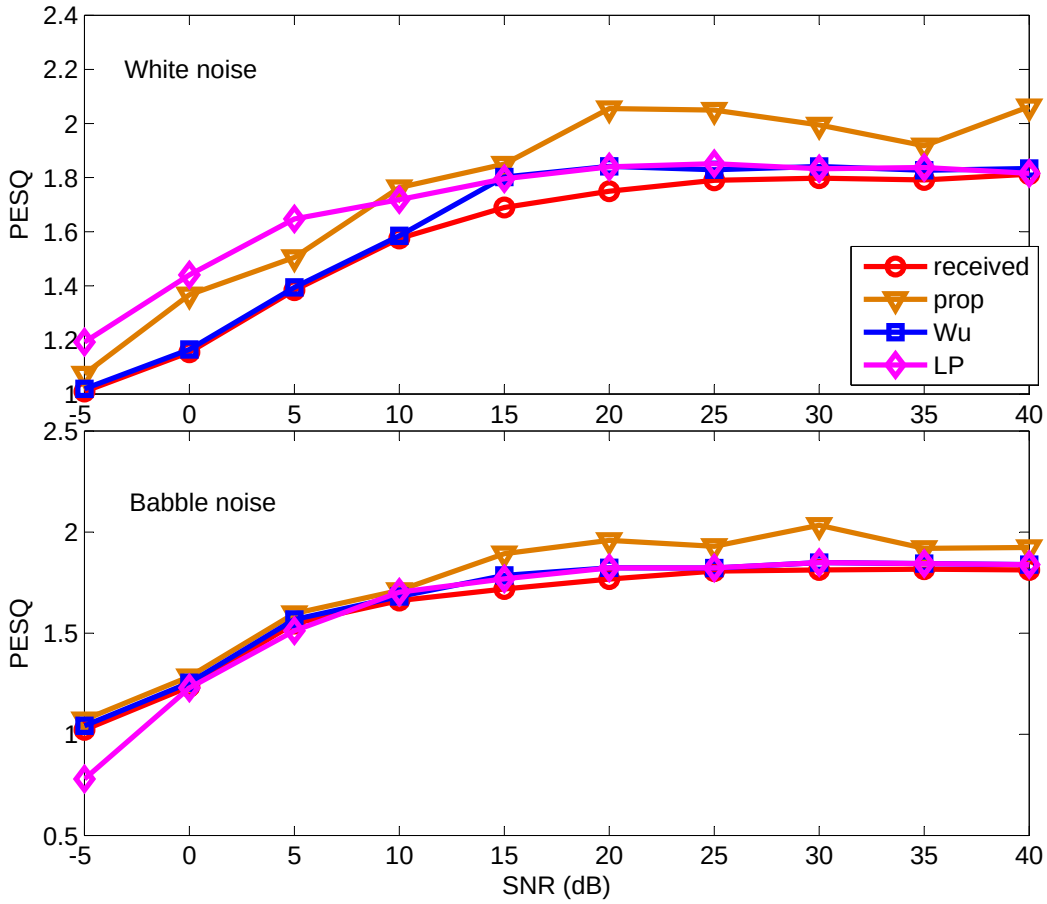

Figure 2.22: PESQ in different noisy conditions with RT60 = 1000 ms and $d = 2$ m. "received", "Wu", "LP" and "prop" represent the PESQ of the received speech, and the processed speech using the Wu and Wang method, the spectral-temporal processing method [17], and the proposed method. The upper plot corresponds to white noise, and the lower corresponds to babble noise.

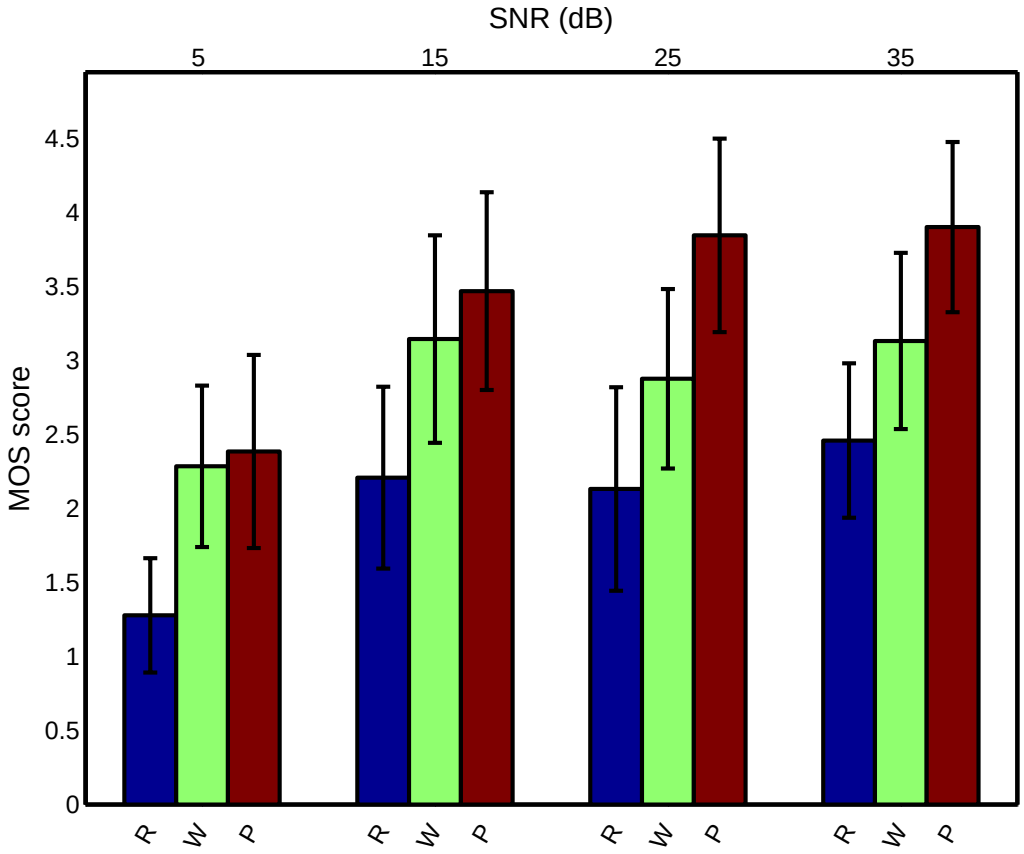

Figure 2.23: MOS scores for different noisy reverberant environments with RT60 = 1 s and $d = 2$ m and white Gaussian noise. "R", "W", and "P" represent the scores for the reverberant speech, and the enhanced speech obtained using the Wu and Wang method [19] and the proposed method, respectively. The variances are indicated by the vertical lines.

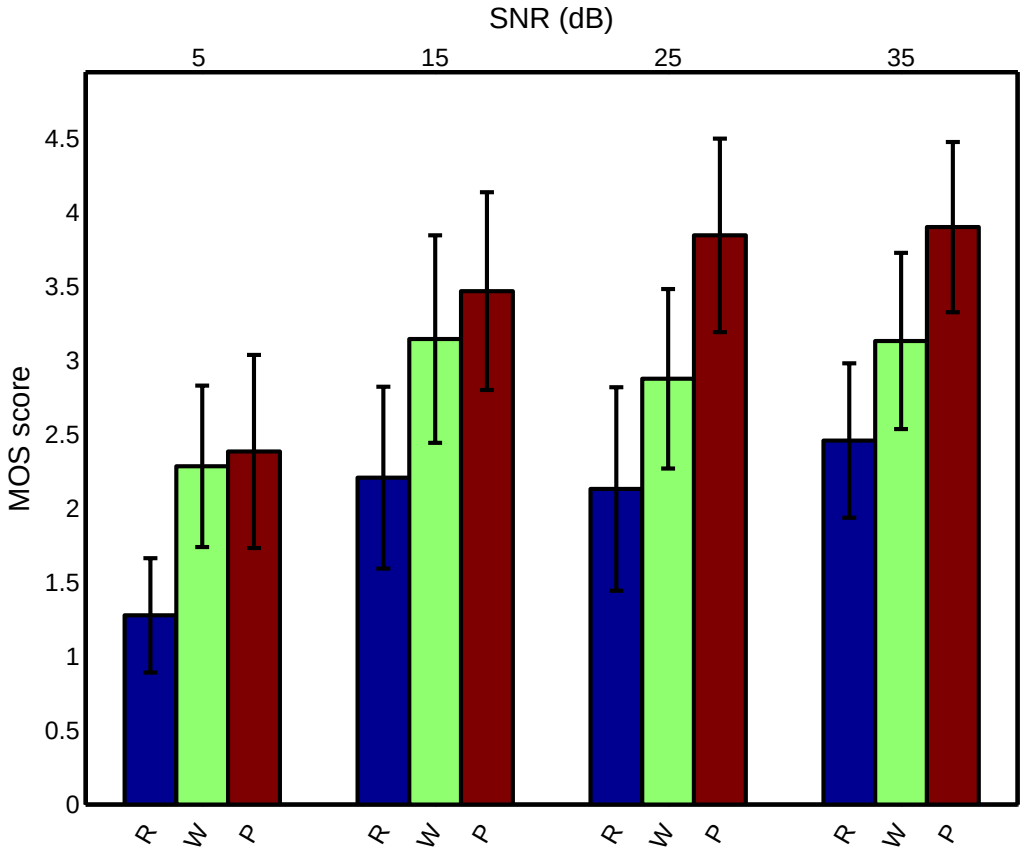

Figure 2.24: MOS scores for different noisy reverberant environments with RT60 = 1 s and $d = 2$ m and babble noise. "R", "W", and "P" represent the scores for the reverberant speech, and the enhanced speech obtained using the Wu and Wang method [19] and the proposed method, respectively. The variances are indicated by the vertical lines.

reduces the early reverberation even in highly reverberant rooms (RT60> 1 s) with low to moderate additive background noise. A noise suppression method based on spectral subtraction was used to reduce the additive background noise in reverberant conditions. Finally, the residual reverberation effects were reduced by a method based on spectral subtraction which provides fewer residual artifacts than other methods. Consequently, the proposed method has several significant advantages over other single microphone methods:

a) the speech enhancement method is **blind**.

b) both **early and late reverberation** is reduced.

c) speech enhancement is achieved even in **very high reverberation conditions**, i.e., a high reverberation time with a low Direct to Reverberation Ratio (DRR).

d) it is robust to **additive background noise**.

e) only **minor artifacts** are introduced, including musical noise.

Chapter 3

Speaker Localization in a Noisy Reverberant Room

Localization of acoustic sources using microphone arrays is an important task in many applications of practical interest. Typical examples can be found in videoconferencing, multimedia, surveillance, hance-free talking systems. The three broad strategies to deal with this problem are: 1) steered response power of a beamformer [44]-[45], 2) high resolution spectrum estimation [46]-[47], and 3) Time Difference Of Arrival (TDOA) estimation [48]-[49].

In a steered beamformer, the microphone array is steered to various locations to search for a peak in the output power. The delay-and-sum beamformer will add appropriate time shifts to the received signals to compensate for the propagation delays. Once these signals are time-aligned, they are added to create a single, enhanced output signal. While beamforming has been extensively used in speech-array applications for voice capture, it has rarely been applied to the speaker localization problem. The task of computing the steered output power for an appropriate dense set of candidate locations is computationally complex and highly dependent on the spectral content of the source signal.

The second category of source location algorithms are based on high resolution spectrum estimation. In this case the spatio-spectral correction matrix is derived from the signals received at the microphones. This matrix is derived using an ensemble average of the signals over the intervals in which the noise and the sources are assumed to be stationary, thus the estimation parameters are assumed to be constant. These high resolution methods are designed for narrowband stationary signals, and hence

are difficult to apply in the case of wideband non-stationary signals like speech.

Methods based on TDOA estimation involve a two-step process, and are more suitable for speech source localization than the previous two approaches. In the first step, the relative TDOA between pairs of microphone signals is estimated. Then in the second step, the source position is obtained from the estimated TDOAs according to some strategy (e.g. geometrical triangulation). Accurate and robust Time Delay Estimation (TDE) is the key to the effectiveness of the localization in this category. Thus in this chapter, the TDE problem is considered to provide robust techniques to estimate the TDOA in a noisy reverberant room.

3.1 Time Delay Estimation in a Noisy Reverberant Room

The goal of Time Delay Estimation (TDE) methods is to estimate the relative Time Difference of Arrival (TDOA) between spatially separated microphones. The estimated time delay information is useful in determining the location of a speaker in a room and tracking the source of a signal. It is also used for applications such as speech enhancement, automatic camera tracking in video-conferencing, and microphone array beam steering. This is a difficult problem since the input signal typically has colored and non-stationary characteristics, and additive noise is present. The presence of reverberation is particularly detrimental because of its long duration and the nonminimum phase nature of the Room Impulse Response (RIR).

The most popular methods for TDE are based on locating the peak in the cross-correlation of the signals received by a pair of microphones [50]. To better deal with noise and reverberation, a number of Generalized Cross-Correlation (GCC) algorithms have been proposed which employ weighting functions [51]. These include constant weighting, the Smoothed Coherence Transform (SCOT), the Phase Transform (PHAT), and Maximum-Likelihood (ML) processing [51]. GCC methods work well in moderate background noise, but can fail even under moderate reverberation, which is common in typical acoustic environments. To improve the performance in noise and reverberation, improved TDE algorithms based on the GCC have been proposed e.g. [54] and [30]. For example, cepstral prefiltering has been employed which is based on the estimation and subtraction of the minimum-phase component of the channel cepstrum from the total cepstrum of each microphone signal [54] . However,

noise remains a problem when the Signal to Noise Ratio (SNR) is low, and most TDE methods perform poorly in the presence of significant reverberation. The failure to work well under reverberant conditions is due to the use of an ideal signal model consisting of single propagation paths from the source to the microphones.

Unlike the methods described above, the adaptive Eigenvalue Decomposition (EVD) technique [56] models the reverberation explicitly. The RIRs are blindly estimated using the covariance matrix of the signals. Although this method provides improved robustness to reverberation, the zeros of the two RIRs can be close, especially in high reverberation conditions, which leads to an ill-conditioned system that is difficult to identify [57]. To deal with this problem, the EVD method has been extended to employ frequency-domain block-processing and multichannel techniques [57]. More than two microphones can be used to provide additional (redundant) information and improve performance. However, in some applications only two microphones are available to estimate the TDOA. Thus it is of significant practical importance to develop robust and efficient techniques which utilize the information from only two microphones for TDE applications. Consequently, in this research we consider the TDE problem using only two microphones receiving a signal from a single source. In the case of multiple sources, approaches such as blind source separation-based localization can be used [58]-[59]. These methods typically rely on the sparseness of the speech or assume mutually uncorrelated sources.

From the above discussion, TDE using only two microphones when the SNR is low and the reverberation is high remains a very challenging problem. In this thesis, two TDE methods are presented using the signals received by two microphones from a single source. The first is based on Adaptive Inverse Filtering (AIF), and employs an estimated inverse filter of the RIR. This makes it very robust to reverberation, but also computationally demanding due to the adaptive estimation of the inverse filter for at least one microphone. In addition, it performs poorly in very low SNR conditions and when the input signals have a symmetric pdf. This motivates us to propose another method to resolve these problems. The second method is based on the GCC and employs two preprocessing stages, namely all-pass processing and spectral subtraction, to improve the performance in noisy reverberant conditions. This results in a robust technique with lower computational complexity that performs better than existing two microphone-based techniques in time-varying and real noise reverberant conditions.

The contributions of the first proposed method based on AIF are as follows.

- In our previous work [31], the TDOA was estimated using the inverse filter of the RIRs which were estimated separately using the approach in Section 2.1. In this work, the estimated inverse filter of the first microphone is used as the initial filter in estimating the inverse filter of the second microphone. This increases the speed of estimation and requires less input data. Then the TDE method in [31] is further improved by estimating the required delay to be used with the LP residual signals.

- An algorithm is developed to estimate the inverse filter of the second microphone directly using the estimated inverse filter of the first microphone. This significantly decreases the computational complexity. A general TDE method is proposed based on this algorithm which performs well in a variety of conditions with a negligible number of TDOA estimation failures. It can be used for any input signal which has an asymmetric pdf.

To the best of our knowledge, no other TDE method based on the signals from only two microphones provides performance similar to the proposed approach in reverberant conditions in terms of the accuracy and number of TDOA estimation failures.

The contributions of the second method based on the GCC are as follows.

- All-pass processing is proven to improve the performance in reverberant conditions.

- Combining spectral subtraction preprocessing with all-pass processing improves the performance in both noisy and reverberant conditions.

The main advantages of the proposed TDE methods over the other approaches are summarized below.

1. Signals are required from only **two microphones**.

2. Both methods are more **robust to reverberation** and can be used in environments with high reverberation times and low direct to reverberation ratios.

3. Both methods are more **robust to additive noise**.

4. The GCC-based method has **low computational complexity** and can be used in **real-time applications**.

5. The GCC-based method performs well in **time-varying environments** even with high reverberation, and is **not sensitive to the type of input signal**.

3.1.1 Time Delay Estimation Based on Adaptive Inverse Filtering

In this section, we present a Time Delay Estimation (TDE) method based on Adaptive Inverse Filtering (AIF). In order to calculate the TDOA between the microphones, the inverse filter of each Room Impulse Response (RIR) h_i^{-1}, $i = 1, 2$, is estimated using the method presented in Section 2.1. TDE estimation based on AIF requires a *good* inverse filter with a specific characteristic, namely a dominant peak that exponentially decays in reverse time (although it may have two dominant peaks). To check if the inverse filter is good, it is sufficient to check the monotonicity of the envelope of the inverse filter in reverse time from the maximum. For example, a RIR with RT60 = 1000 ms obtained using the image method [3] is shown in Figs. 3.1 (a) and (d). The inverse filter of this RIR was estimated using inverse filter lengths of 3000 and 6000 samples, and the results are shown in Fig. 3.1 (b) and (e), respectively. Fig. 3.1 (b) illustrates a *poor* inverse filter as it does not have an envelope which is exponentially decaying in reverse time. Fig. 3.1 (e) shows a *good* inverse filter, which indicates that increasing the filter length results in an inverse filter which is suitable for TDE applications. This is because the dominant peak can be clearly identified. The equalized impulse responses obtained by convolving the RIR with the estimated inverse filters are shown in Figs. 3.1 (c) and (f). Both are impulse-like functions and thus both inverse filters may be suitable for speech enhancement applications, but the shorter inverse filter is not suitable for TDE applications.

In order to calculate the TDOA between two microphones, the inverse filter of each RIR must be estimated, but estimating these separately results in high computational complexity. Here, this complexity is reduced by estimating the inverse filter for the second microphone using the estimated filter for the first microphone as the initial inverse filter. This improves the convergence rate and decreases the amount of input speech data required. As an example, RIRs with RT60 = 400 ms were obtained using the image method [3] for two microphones 1.118 m apart. The inverse filters for the two microphones using an all-pass filter as the initial filter are shown in Figs. 3.2 (a) and (b), and the filter for the second microphone using (a) as the initial filter is shown in Fig. 3.2 (c). The average LP residual skewness for each estimation iteration corresponding to these inverse filters is shown in Fig. 3.2 (d). This shows that the algorithm converges much faster when (a) is used as the initial filter. In addition, the required input speech data is approximately halved.

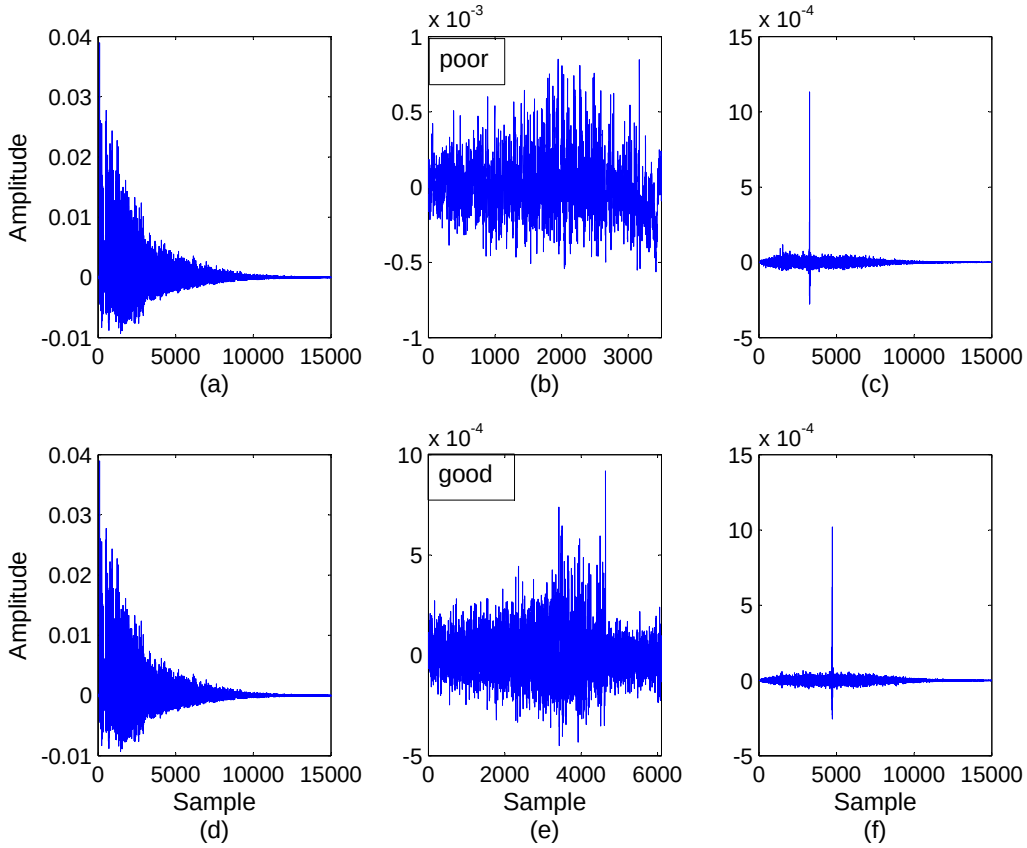

Figure 3.1: RIR with RT60 = 1000 ms (a) and (d), inverse filters estimated using different filter lengths (b) and (e), and the corresponding equalized impulse responses (c) and (f), respectively. By definition, (b) represents a *poor* inverse filter and (e) a *good* inverse filter.

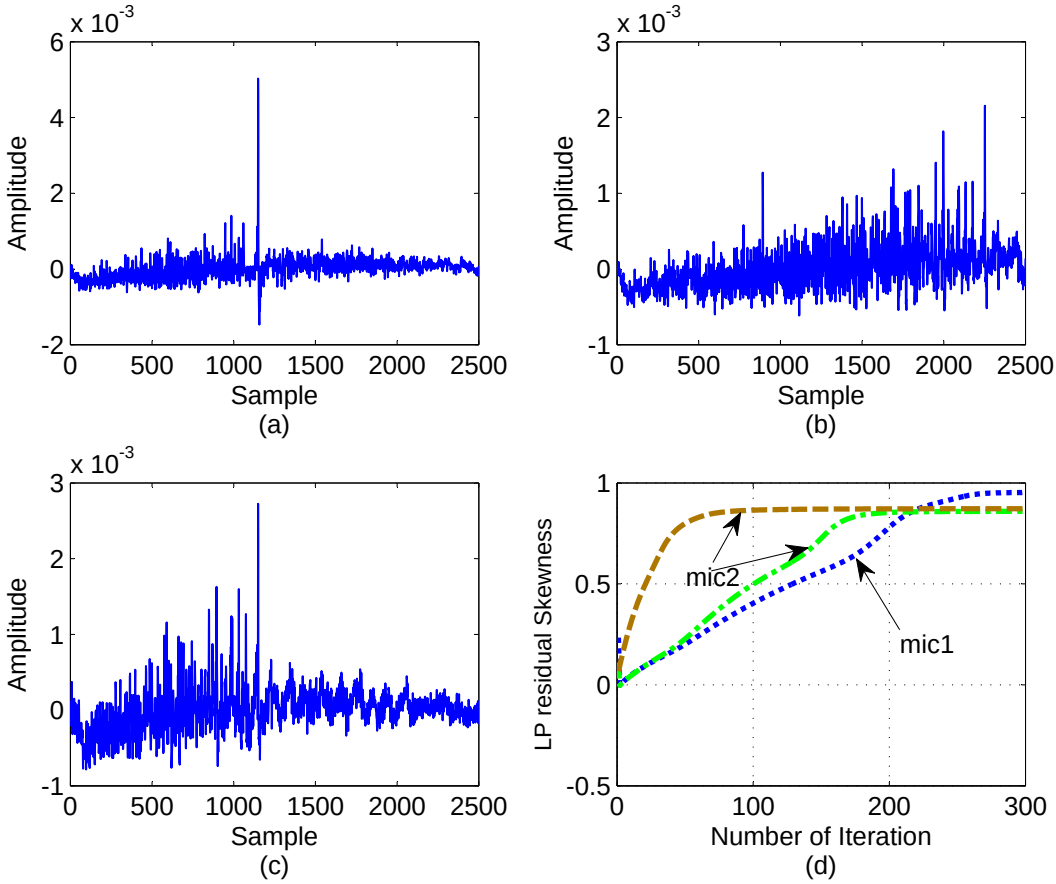

Figure 3.2: The inverse filters for two RIRs with RT60 = 400 ms (a) and (b) inverse filters of the first and second RIRs using an all-pass filter as the initial filter, respectively, (c) inverse filter of the second RIR using (a) as the initial filter, and (d) the average LP residual skewness for each iteration of the inverse filter estimation in (a)-(c).

Relationship between the Inverse Filter and Propagation Delay

If we have a *good* inverse filter estimate for the i-th microphone then

$$h_i[n] \star \hat{h}_i^{-1}[n] \approx \kappa_i \delta[n - D_i], \tag{3.1}$$

where $h_i[n]$ and $\hat{h}_i^{-1}[n]$ are the RIR and estimated inverse filter, D_i is the unknown delay, κ_i is a constant, and $\star$ denotes convolution. As an example, two RIRs both with reverberation time RT60 = 400 ms are given in Figs. 2.13 (a) and (b) for speaker-microphone distances $d = 1$ m and $d = 2$ m, respectively. The corresponding inverse filters are shown in Figs. 2.13 (c) and (d). These results show that the inverse filters are similar to delayed time-reversed versions of the RIRs. The equalized impulse responses obtained by convolving the inverse filters with the corresponding RIRs are shown in Figs. 3.4 (a) and (b).

Since it can be assumed that the RIR is a white random process [10], the convolution of $h_i[n]$ with $h_i[-n]$ is approximately proportional to an unit impulse

$$h_i[n] \star h_i[-n] \approx \kappa_h \delta[n]. \tag{3.2}$$

This is confirmed by Figs. 3.4 (c) and (d), which show the autocorrelation functions of the RIRs in Figs. 2.13 (a) and (b), respectively. From this figure, the autocorrelations of the RIRs are approximately unit impulses located at the origin. This also holds for the equalized impulse responses. Comparing (3.1) and (3.2), we obtain the following approximation

$$\hat{h}_i^{-1}[n] \approx \frac{\kappa_i}{\kappa_h} h_i[D_i - n]. \tag{3.3}$$

This is confirmed by the results in Fig. 2.13. Thus the index of the maximum value of the inverse filter and the corresponding RIR are related according to

$$D_i = \arg\max_n \hat{h}_i^{-1}[n] - \arg\max_n h_i[n]. \tag{3.4}$$

For two microphones, from (3.4) we have that

$$\text{TDOA} = D_{inv} - D, \tag{3.5}$$

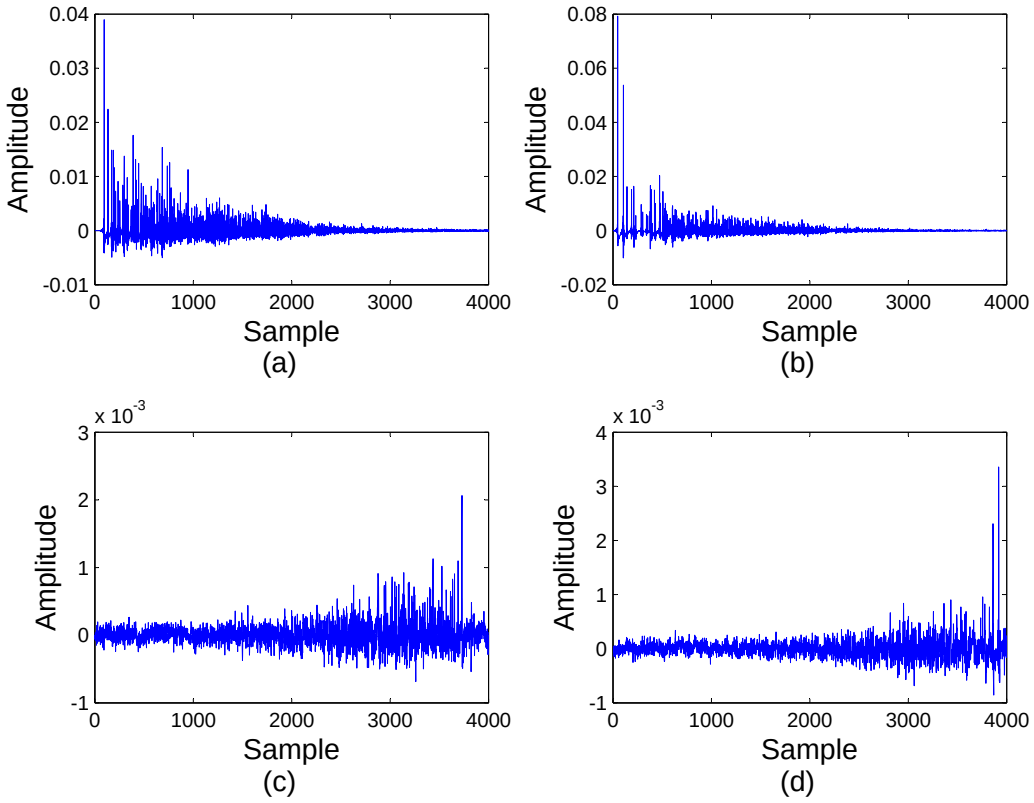

Figure 3.3: (a) and (b) RIR with RT60 = 400 ms and $d = 1$ and $d = 2$ m, respectively, and (c) and (d) the corresponding inverse filters.

where

$$\text{TDOA} = \arg\max_n h_2[n] - \arg\max_n h_1[n], \tag{3.6}$$

$$D_{inv} = \arg\max_n \hat{h}_2^{-1}[n] - \arg\max_n \hat{h}_1^{-1}[n], \tag{3.7}$$

$$D = D_2 - D_1. \tag{3.8}$$

The Proposed TDE method based on adaptive inverse filtering

A block diagram of the TDE method based on the two-channel AIF algorithm is shown in Fig. 3.5. The inverse filter estimate for the first microphone is $h_1^{-1}[n]$. This estimate is used as the initial filter for estimating the inverse filter of the second microphone $h_2^{-1}[n]$. In contrast to the method in [31], the required delay D is estimated using the

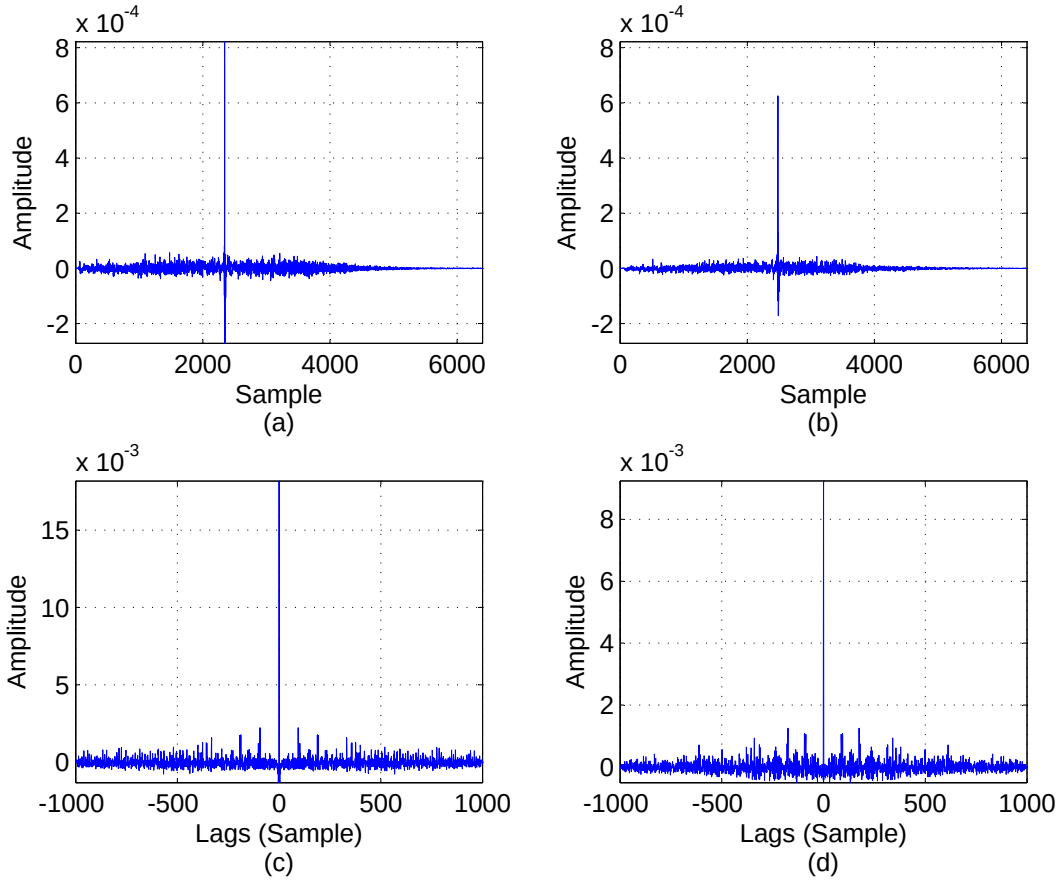

Figure 3.4: (a) and (b) the equalized impulse responses for the RIRs in Figs. 1 (a) and (b), respectively, and (c) and (d) the corresponding autocorrelation functions.

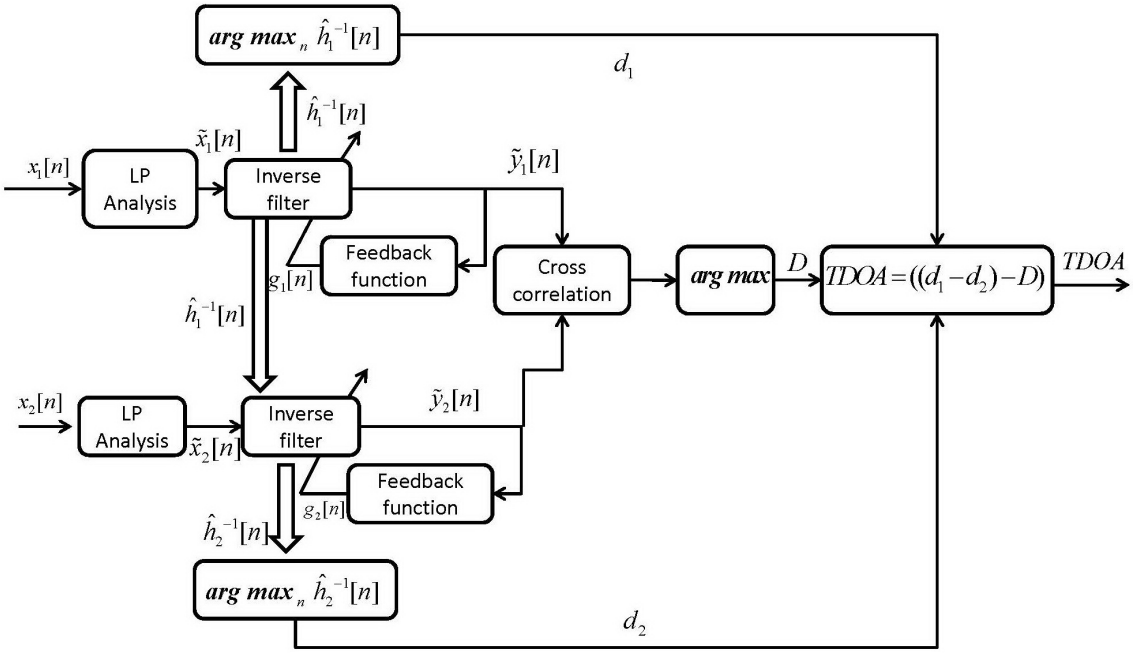

Figure 3.5: Block diagram of the TDE method based on the two-channel AIF algorithm.

inverse-filtered LP residual signals. The estimated inverse filters used to calculate the inverse-filtered LP residual signals are

$$\tilde{y}_1[n] = \tilde{x}_1[n] \star h_1^{-1}[n], \tag{3.9}$$

$$\tilde{y}_2[n] = \tilde{x}_2[n] \star h_2^{-1}[n], \tag{3.10}$$

where $\tilde{x}_1[n]$ and $\tilde{x}_2[n]$ are the LP residual signals for the first and second microphones, respectively. The cross-correlation of $\tilde{y}_1[n]$ and $\tilde{y}_2[n]$ is

$$R_{12}[n] = \tilde{y}_1[n] \star \tilde{y}_2[-n], \tag{3.11}$$

and the index of the maximum of $R_{12}[n]$ is

$$D = \arg\max_n \{R_{12}[n]\}. \tag{3.12}$$

Finally, the TDOA between the two microphones is estimated using (3.5).

This method requires that the inverse filter of each channel be estimated separately. To reduce the computational complexity, an estimate of the inverse filter of the second channel can be obtained using the estimated inverse filter of the first channel.

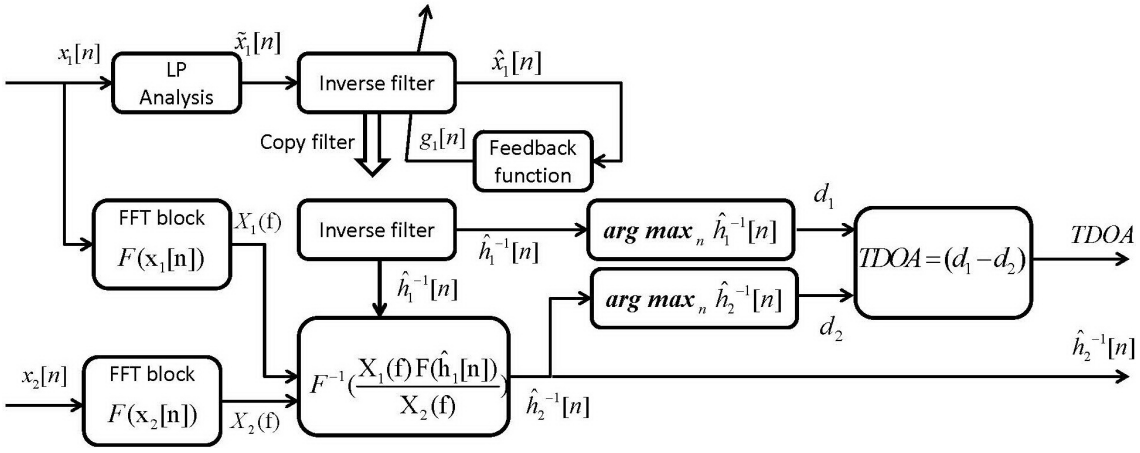

Figure 3.6: Block diagram of the proposed TDE method based on the one-channel AIF algorithm.

This is most effective when the two RIRs have similar characteristics, which is often the case.

Let the received speech signals obtained by the two microphones be $x_i[n]$, $i = 1, 2$. The corresponding received speech signals can be modeled as

$$x_1[n] = s[n] \star h_1[n], \tag{3.13}$$

$$x_2[n] = s[n] \star h_2[n], \tag{3.14}$$

where $s[n]$, $h_1[n]$, and $h_2[n]$ represent the clean speech signal, the first RIR and the second RIR, respectively. In the frequency domain, these can be written as

$$X_1(f) = S(f)H_1(f), \tag{3.15}$$

$$X_2(f) = S(f)H_2(f), \tag{3.16}$$

where $X_1(f)$, $X_2(f)$, $H_1(f)$, $H_2(f)$), and $S(f)$ are the fast Fourier transforms (FFTs) of $x_1[n]$, $x_2[n]$, $h_1[n]$, $h_2[n]$, and $s[n]$, respectively. The estimated inverse filters can be expressed as

$$\hat{h}_1^{-1}[n] = h_1^{-1}[n] \star \delta[n - D_1], \tag{3.17}$$

$$\hat{h}_2^{-1}[n] = h_2^{-1}[n] \star \delta[n - D_2]. \tag{3.18}$$

Thus $\hat{h}_2^{-1}[n]$ can be estimated using (3.15), (3.16) and (3.17), and the FFT of $\delta[n-D_1]$ which is $e^{-j2\pi f D_1}$, giving

$$\begin{aligned}
\hat{h}_2^{-1}[n] &= F^{-1}\left(\frac{X_1(f)}{X_2(f)} F(\hat{h}_1^{-1}[n])\right) && (3.19)\\
&= F^{-1}\left(\frac{S(f)H_1(f)}{S(f)H_2(f)} F(h_1^{-1}[n] \star \delta[n - D_1])\right)\\
&= F^{-1}\left(\frac{H_1(f)}{H_2(f)} \frac{e^{-j2\pi f D_1}}{H_1(f)}\right)\\
&= F^{-1}\left(\frac{e^{-j2\pi f D_1}}{H_2(f)}\right)\\
&= h_2^{-1}[n] \star \delta[n - D_1], && (3.20)
\end{aligned}$$

where F and F^{-1} denote the FFT and inverse FFT, respectively.

For implementation purposes, the reverberant speech signals $x_1[n]$ and $x_2[n]$ can be segmented into blocks of length L (the length of the inverse filter), giving $x_1[n, k]$

and $x_2[n,k]$, respectively, where k is the block number. Then the FFT of the averaged blocks is

$$\overline{X}_1(f) = F\left(\frac{1}{K}\sum_{k=1}^{K} x_1[n,k]\right), \tag{3.21}$$

$$\overline{X}_2(f) = F\left(\frac{1}{K}\sum_{k=1}^{K} x_2[n,k]\right), \tag{3.22}$$

where K is the number of blocks. $\hat{h}_2^{-1}[n]$ is then estimated using (3.19) as

$$\hat{h}_2^{-1}[n] = F^{-1}\left(\frac{\overline{X}_1(f)F(\hat{h}_1^{-1}[n])}{\overline{X}_2(f)}\right). \tag{3.23}$$

Equation (3.20) shows that the estimated inverse filter for the second RIR using the proposed algorithm has the same delay D_1 as the first channel and thus $D = 0$ in (3.5). A block diagram of the TDE method based on the one-channel AIF algorithm is given in Fig. 3.6.

Figure 3.7 illustrates the proposed algorithm with two RIRs with RT60 = 400 ms and speaker-microphone distances of $d = 2.06$ m (Fig. 3.7 (a)), and $d = 1.41$ m (Figs. 3.7 (d) and (g)). The inverse filter of the first RIR was estimated using the AIF algorithm in [28], and the resulting filter and corresponding equalized impulse response are shown in Figs. 3.7 (b) and (c), respectively. Equation (3.23) was then used to estimate the inverse filter of the second RIR and the resulting filter and corresponding equalized impulse response are shown in Figs. 3.7 (e) and (f), respectively. This clearly indicates that the estimated inverse filter obtained using (3.23) is accurate and has a delay similar to that of the equalized impulse response. However, if the AIF algorithm in [28] is used to estimate the inverse filter of the second RIR independently, the resulting filter has a delay that differs from that of the equalized impulse response as shown in Figs. 3.7 (h) and (j).

In the unlikely event that the second estimated inverse filter obtained using the one-channel AIF algorithm is not suitable for TDE applications, i.e., it is not a *good* inverse filter (which may occur as when the two microphones are far apart), the second inverse filter can be estimated using the two-channel AIF algorithm. Based on this, the AIF algorithm for TDE is summarized below.

1. The first inverse filter is estimated using the method presented in Section 2.1.

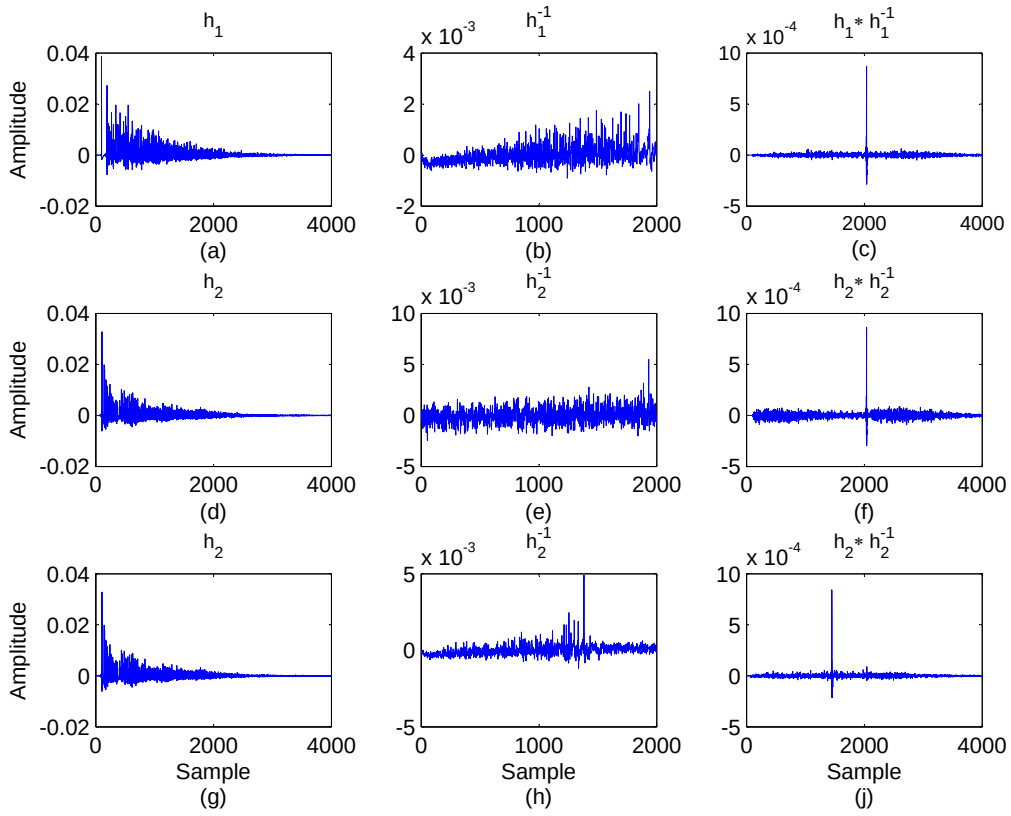

Figure 3.7: The RIRs on the left, their estimated inverse filters in the middle, and the convolution of each pair on the right. The estimated inverse filters (b) and (h) were obtained using AIF, while the estimated inverse filter (e) was obtained using (3.23). h_1 has RT60 = 400 ms and $d = 2.06$ m, and h_2 has RT60 = 400 ms and $d = 1.41$ m.

2. The second inverse filter is estimated using (3.23).

3. This second estimated filter is checked to see whether it is a *good* inverse filter.

4. If it is good, the TDOA is estimated by subtracting the indexes of the estimated inverse filters. Otherwise, the second inverse filter is estimated using the first inverse filter as the initial filter. Then the TDOA is estimated using (3.5).

In practice, the distance between microphones used for source localization is not large, and the one-channel AIF algorithm almost always produces acceptable results.

Improved maximum value selection

As mentioned previously, an appropriate estimated inverse filter has the characteristic of an exponentially decaying function in reverse time. According to (3.3), an inverse

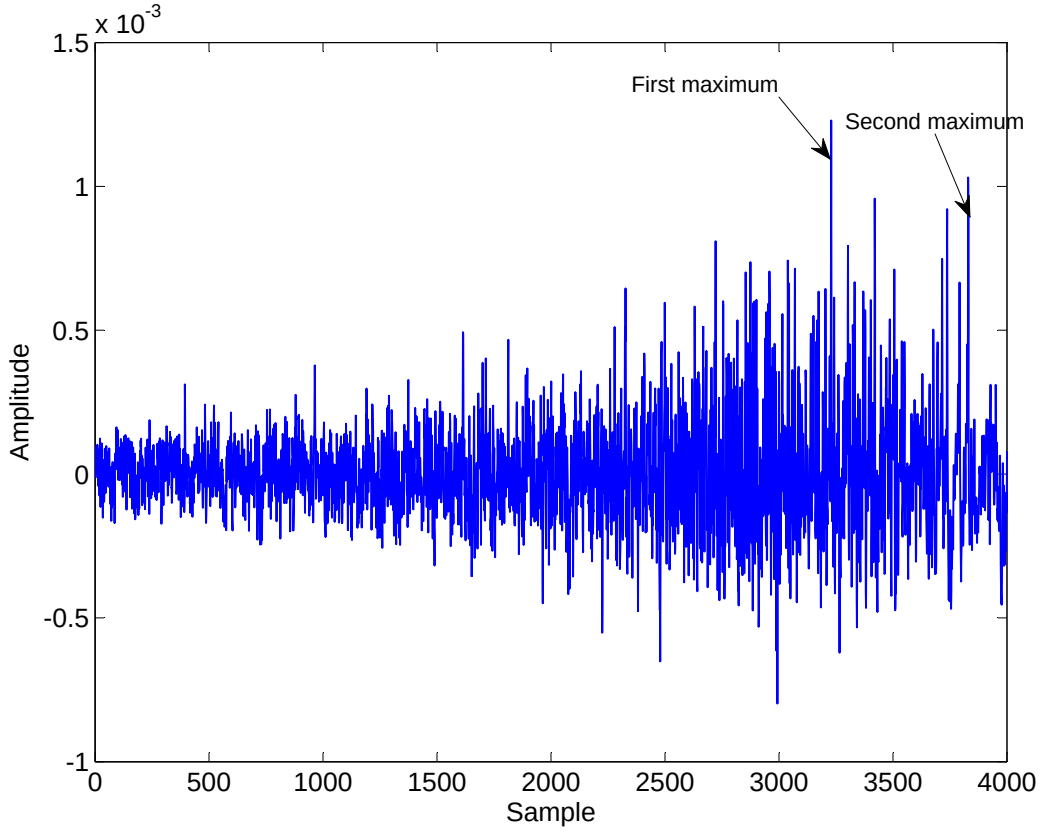

Figure 3.8: An example of an estimated inverse filter which does not begin in reverse time with the maximum value. This is an unusual case where the maximum corresponds to early reverberation while the second highest value corresponds to the direct component of the RIR.

filter in reverse time should begin with a maximum value (related to the direct component of the RIR), followed by smaller amplitude values corresponding to the early reverberation of the RIR. However, it may happen that an inverse filter does not begin with the maximum value in reverse time. Instead, it may begin with the second, third or subsequent highest value. For example, Fig. 3.8 shows an inverse filter which begins in reverse time with the second highest value, while the maximum corresponds to early reverberation. As a result, instead of using the index of the maximum of the inverse filter, the index of the first significant value of the inverse filter (which has an exponentially decaying characteristic), should be used in (3.7). With this slight modification, the number of TDOA estimation failures becomes negligible.

Figure 3.9 compares the percentage of failures for the proposed AIF method using the above modification and the GCC method for TDOA estimation with 105 different

pairs of RIRs and 8 different utterances (4 male speaker and 4 female speaker from the TIMIT database sampled at 16 kHz). An acceptable estimate is defined as one that satisfies

$$|\text{TDOA}| \leq f_s \frac{d_m}{c}, \tag{3.24}$$

where d_m is the distance between microphones, $f_s = 16000$ is the sampling rate, and $c = 340$ m/s is the velocity of sound. The percentage of failures with the two methods is shown in Fig. 3.9. The upper plot is for different reverberation times (15 RIRs for each RT60) and no noise, while the bottom plot is for different SNRs with RT60 = 1000 ms and a speaker-microphone distance of $d = 2$ m. With reverberation and no noise, the PHAT method has better performance than the other GCC methods, while in noisy conditions the CC method outperforms the others. Thus the PHAT method was used to obtain the results in Fig. 3.9 (a), and the CC method to obtain the results in Fig. 3.9 (b). This figure clearly indicates that the proposed AIF method is superior to the existing methods in terms of the percentage of failures.

Improved AIF-based TDE Method for Time-Varying Environments

In [28], it was shown that inverse filtering can be used in slow time-varying environments where the speaker pauses after moving. This pause should be sufficiently long to allow the inverse filter to be updated. The current inverse filter can be used to initialize the algorithm. There are two problems with using this technique in a time-varying environment. First, depending on the reverberation and changes in the Direct to Reverberation Ratio (DRR)[1], 5 to 10 s of input data are required in order to accurately estimate the new inverse filter. Second, this update requires a sufficient number of iterations to converge and so creates additional delay in the TDE. However, these problems can be overcome by using the one-channel AIF algorithm.

Using the one-channel AIF, the inverse filter can be estimated from the received reverberant speech signal using the method in [28]. This signal and the estimated inverse filter are considered to be $x_1[n]$ and $h_1^{-1}[n]$. Then the inverse filter $h_2^{-1}[n]$ can be updated after each movement of the speaker even if the pause is very short (less than 500 ms for high reverberation), using (3.23) where $x_2[n]$ is the received speech signal after the speaker moves. The reason that the pause can be very short is that the minimum length for $x_1[n]$ and $x_2[n]$ in (3.23) is the length of the inverse filters $h_1^{-1}[n]$ and $h_2^{-1}[n]$. According to [25], the maximum inverse filter length in high

[1]The DRR is the ratio of the direct path energy to the total reflective energy [28].

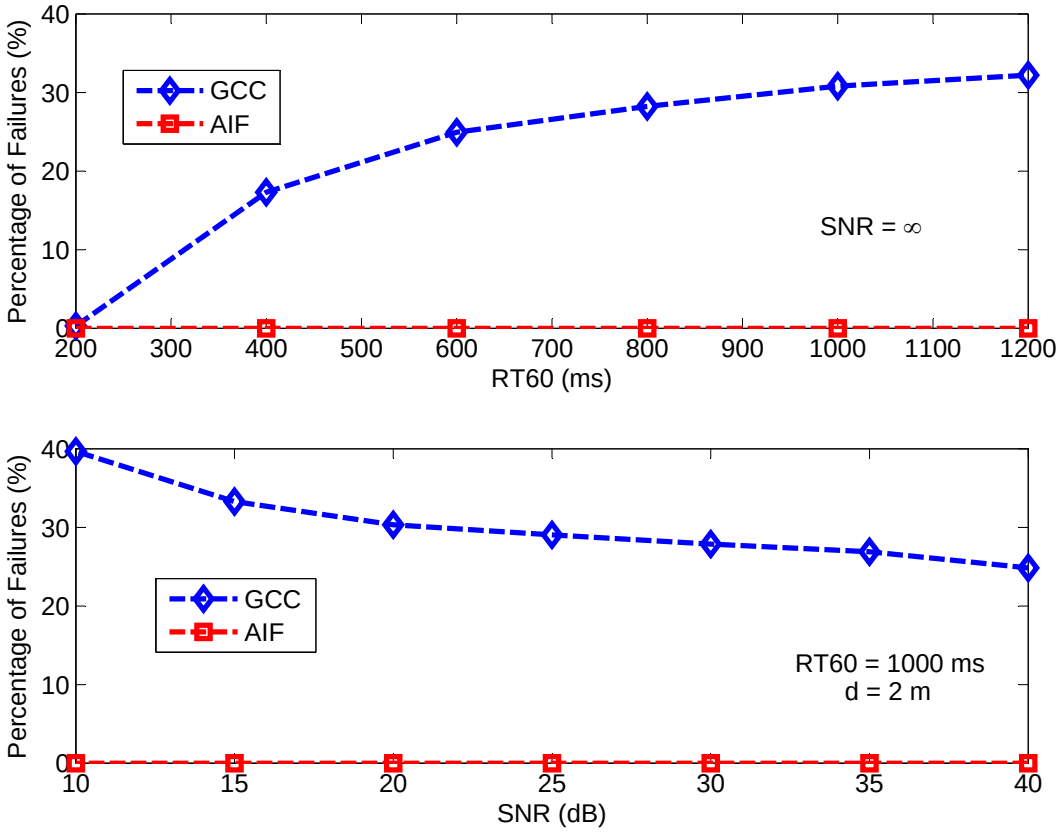

Figure 3.9: The percentage of failures using the GCC [51] and proposed AIF methods. The upper plot shows the results without noise and the bottom plot shows the results for different SNRs when RT60 = 1000 ms and $d = 2$ m.

reverberation environments is only 6000 (375 ms), so the updates can be performed quickly using a non-iterative technique.

Input Signal Types for the AIF-based TDE Method

The AIF method is an adaptive gradient-ascent algorithm with an objective function which uses skewness (normalized third order moment), as a measure of the asymmetry of the data around the sample mean [28]. In fact, the input signal should have an asymmetric distribution with sufficient skewness so that the reverberation moves it towards a symmetric Gaussian distribution with zero skewness. Then the inverse filter of the room reverberation can be estimated by maximizing the skewness and forcing the distribution to be asymmetric. However, speech signals typically have a symmetric distribution (e.g. Laplacian or more generally super Gaussian) [28]. The reason these signals can still be used with AIF methods is that a sufficiently long

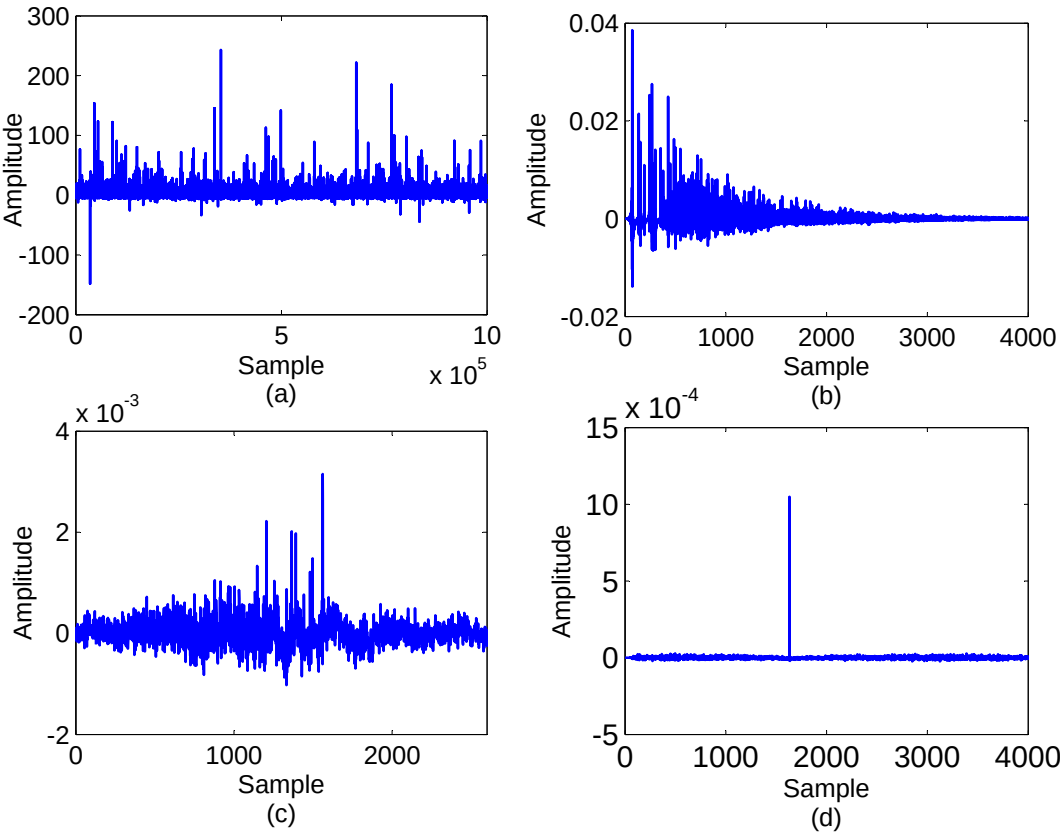

Figure 3.10: (a) a sequence of random variables from an asymmetric pdf with an alpha-stable distribution [63], (b) a RIR with RT60 = 400 ms and d = 1.5 m, (c) the estimated inverse filter using the proposed AIF method, and (d) the equalized impulse response.

LP residual of the speech signal has higher skewness compared to the original speech signal [28]. This is the reason that Figs. 3.5 and 3.6 have LP analysis blocks to generate the LP residual signal.

As an example of a non-speech signal, a sequence of random variables with an asymmetric pdf using the alpha-stable random number generator was generated [63]. This signal is shown in Fig. 3.10 (a), and was convolved with the RIR with RT60 = 400 ms given in Fig. 3.10 (b). The inverse filter of the RIR for this signal was estimated using the technique in [28], and the result is shown in Fig. 3.10 (c). The convolution of the RIR and the inverse filter (equalized impulse response) given in Fig. 3.10 (d) verifies that the inverse filter has been accurately estimated. Note that the LP analysis block is not employed when the input signal is not speech.

3.1.2 Time Delay Estimation Using All-Pass Component Processing and Spectral Subtraction

Although the proposed TDE method based on AIF is very accurate, it has some drawbacks. First, the inverse filter estimation has high computational complexity. This means that it requires a long input signal (e.g. 20 s for high reverberation times) to converge to an acceptable inverse filter. Also, it requires approximately 300 iterations to estimate the inverse filter. Thus it may not be suitable for real-time applications and time-varying environments. Second, it does not perform well when the noise level is high (e.g. SNR < 10 dB). Finally, it is limited to input signals which have an asymmetric pdf with sufficient skewness such as the LP residual of a speech signal. To resolve these problems, a robust TDE method is presented in this section based on all-pass component and spectral subtraction processing.

A typical signal $x[n]$ has non-minimum phase and thus consists of a minimum phase component $x_{min}[n]$ convolved with an all-pass component $x_{all}[n]$, i.e.

$$x[n] = x_{min}[n] \star x_{all}[n]. \tag{3.25}$$

In the frequency domain, (3.25) can be written as

$$X(f) = X_{min}(f)X_{all}(f), \tag{3.26}$$

where $X(f)$, $X_{min}(f)$ and $X_{all}(f)$ are the FFTs of $x[n]$, $x_{min}[n]$ and $x_{all}[n]$, respectively. $X_{min}(f)$ has no poles or zeros outside the unit circle. In addition, $|X(f)| = |X_{min}(f)|$ and $|X_{all}(f)| = 1$.

Effect of All-Pass Component Calculation on the Cross-Correlation for TDE

Let the received signal from the first and second microphones be

$$x^{(i)}[n] = s[n] \star h^{(i)}[n] \quad i = 1, 2, \tag{3.27}$$

respectively, where $s[n]$ is the clean speech signal and $h^{(1)}[n]$, and $h^{(2)}[n]$ are the corresponding RIRs. The noise is first assumed to be negligible and thus is ignored. The effect of noise will be examined later. In the frequency domain, (3.27) can be

written as

$$X^{(i)}(f) = S(f)H^{(i)}(f) \qquad i = 1, 2, \tag{3.28}$$

where $S(f)$, $H^{(1)}(f)$ and $H^{(2)}(f)$ are the FFTs of $x^{(i)}[n]$, $h^{(1)}[n]$, and $h^{(2)}[n]$, respectively. Equation (3.27) can be written in terms of the corresponding minimum-phase and all-pass components as

$$\begin{aligned} x^{(i)}_{min}[n] \star x^{(i)}_{all}[n] &= (s_{min}[n] \star s_{all}[n]) \star (h^{(i)}_{min}[n] \star h^{(i)}_{all}[n]), & (3.29) \\ &= (s_{min}[n] \star h^{(i)}_{min}[n]) \star (s_{all}[n] \star h^{(i)}_{all}[n]) \qquad i = 1, 2. & (3.30) \end{aligned}$$

The all-pass components for the microphones are

$$x^{(i)}_{all}[n] = s_{all}[n] \star h^{(i)}_{all}[n] \qquad i = 1, 2. \tag{3.31}$$

In the frequency domain, these can be written as

$$X^{(i)}_{all}(f) = S_{all}(f)H^{(i)}_{all}(f) \qquad i = 1, 2. \tag{3.32}$$

The cross-power spectrum of $x^{(1)}[n]$ and $x^{(2)}[n]$ is

$$G_{x^{(1)}x^{(2)}}(f) = G_{ss}(f)H^{(1)}(f)conj(H^{(2)}), \tag{3.33}$$

where $G_{ss}(f)$ is the power spectrum of $s[n]$ and $conj(\cdot)$ denotes complex conjugate. As a simple example, let $h^{(1)}[n] = \delta[n]$ and $h^{(1)}[n] = \alpha\delta[n - n_d]$, where α is the amplitude and n_d is the delay. Then (3.33) can be written as

$$G_{x^{(1)}x^{(2)}}(f) = \alpha G_{ss}(f)e^{-j2\pi f n_d}. \tag{3.34}$$

In order to estimate the delay n_d using the Cross-Correlation (CC) method [50], the index of the maximum peak of the CC must be estimated. The CC of $x^{(1)}[n]$ and $x^{(2)}[n]$ is given by the IFFT of the cross-power spectrum given by

$$R_{x^{(1)}x^{(2)}}[n] = \alpha R_{ss}[n] \star \delta[n - n_d]. \tag{3.35}$$

One interpretation of (3.35) is that the unit impulse has been spread or smeared by the signal spectrum $G_{ss}(f)$. If the source $s[n]$ is white noise, its Fourier transform is a

unit impulse and no spreading occurs. However, the signal spectrum such as that for speech is typically far from a unit impulse thus it smears the cross-power spectrum. Thus the CC peak at n_d is not sharp. This makes it difficult if not impossible in realistic situations with multiple delays to distinguish the peaks, which is a serious problem for TDE.

In order to better understand the smearing problem, a TDE test was conducted using the CC method with the white noise and voiced speech segments shown in Fig. 3.11. The two RIRs illustrated in Fig. 3.12 were considered. One has a single impulse component at a delay of 25 samples (Fig. 3.12 (a)), while the second has a broader response with decreasing amplitudes from samples 25 to 41 (Fig. 3.12 (b)). First, the white noise segment as $x^{(1)}[n]$ was convolved with the single-delay RIR resulting in $x^{(2)}[n]$. The CC of these two signals is given in Fig. 3.13 (a) and shows a prominent peak at sample 25. Then, the white noise segment as $x^{(1)}[n]$ was convolved with the multiple component RIR resulting in $x^{(2)}[n]$, and the corresponding CC is shown in Fig. 3.13 (b). This also has a prominent peak at sample 25. Thus when $s[n]$ is a white noise source, no spreading takes place and the CC always has an identifiable peak at the correct delay (here sample 25). Second, the speech segment as $x^{(1)}[n]$ was convolved with the two RIRs. For the single component RIR, the CC is shown in Fig. 3.13 (c). This again has a peak at sample 25. However, with the multiple component RIR, the corresponding CC shown in Fig. 3.13 (d) has a peak at sample 31, which is incorrect. Thus, the speech segment smears the CC which results in a TDE error when the RIR has multiple components. This is a serious problem as such an RIR is more typical than a single component RIR.

To solve the above problem, we propose using the all-pass component of the received signals to calculate the CC. The cross-power spectrum of $x_{all}^{(1)}[n]$ and $x_{all}^{(2)}[n]$ is

$$G_{x_{all}^{(1)}x_{all}^{(2)}}(f) = G_{s_{all}s_{all}}(f)H_{all}^{(1)}(f)conj(H_{all}^{(2)}), \tag{3.36}$$

$$= H_{all}^{(1)}(f)conj(H_{all}^{(2)}), \tag{3.37}$$

since $G_{s_{all}s_{all}}(f) = |S_{all}|^2 = 1$. For the simple case considered above

$$G_{x_{all}^{(1)}x_{all}^{(2)}}(f) = e^{-j2\pi f n_d}. \tag{3.38}$$

This indicates that all-pass processing eliminates the signal spectrum and thus no

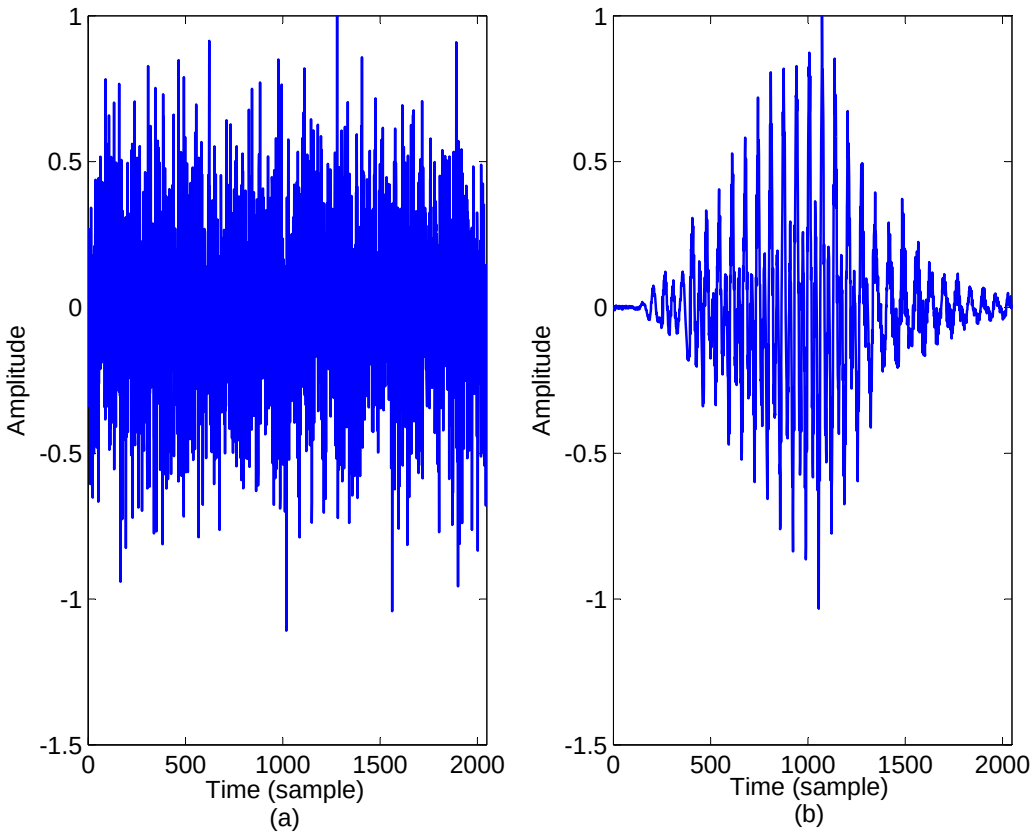

Figure 3.11: (a) a sequence of random variables from an asymmetric pdf with an alpha-stable distribution [63], (b) a RIR with RT60 = 400 ms and $d = 1.5$ m, (c) the estimated inverse filter using the proposed AIF method, and (d) the equalized impulse response.

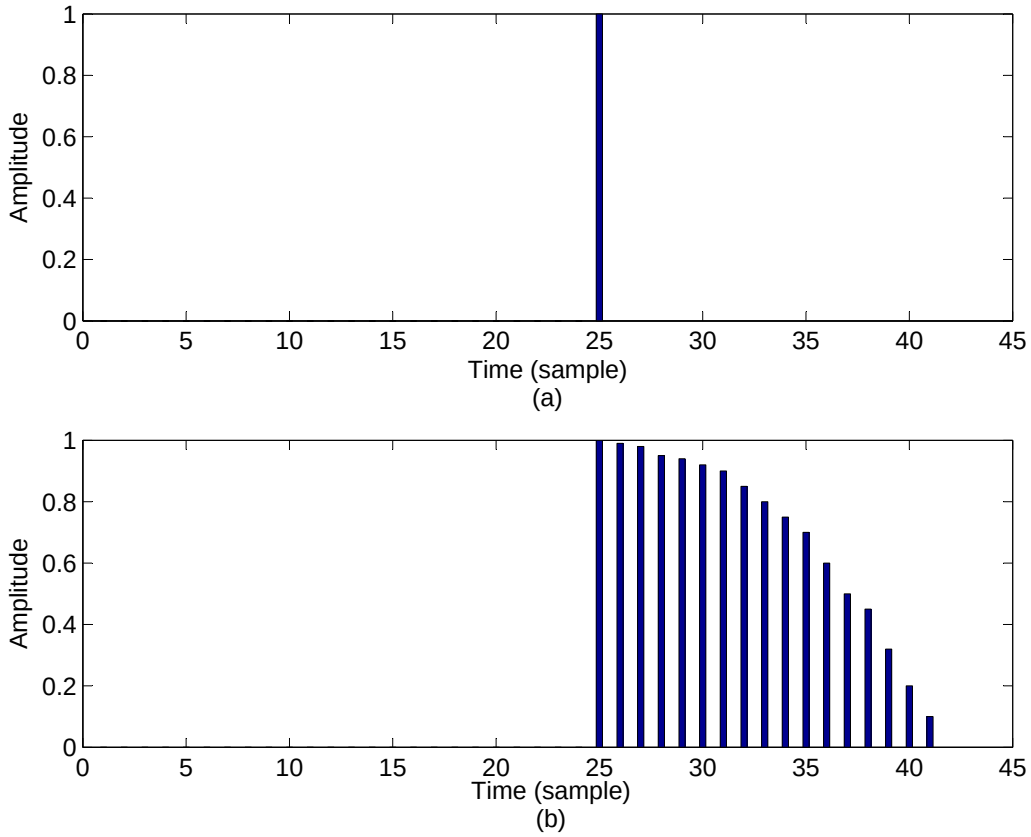

Figure 3.12: The (a) single component RIR, and (b) multiple component RIR, used for TDE using the CC method in [50].

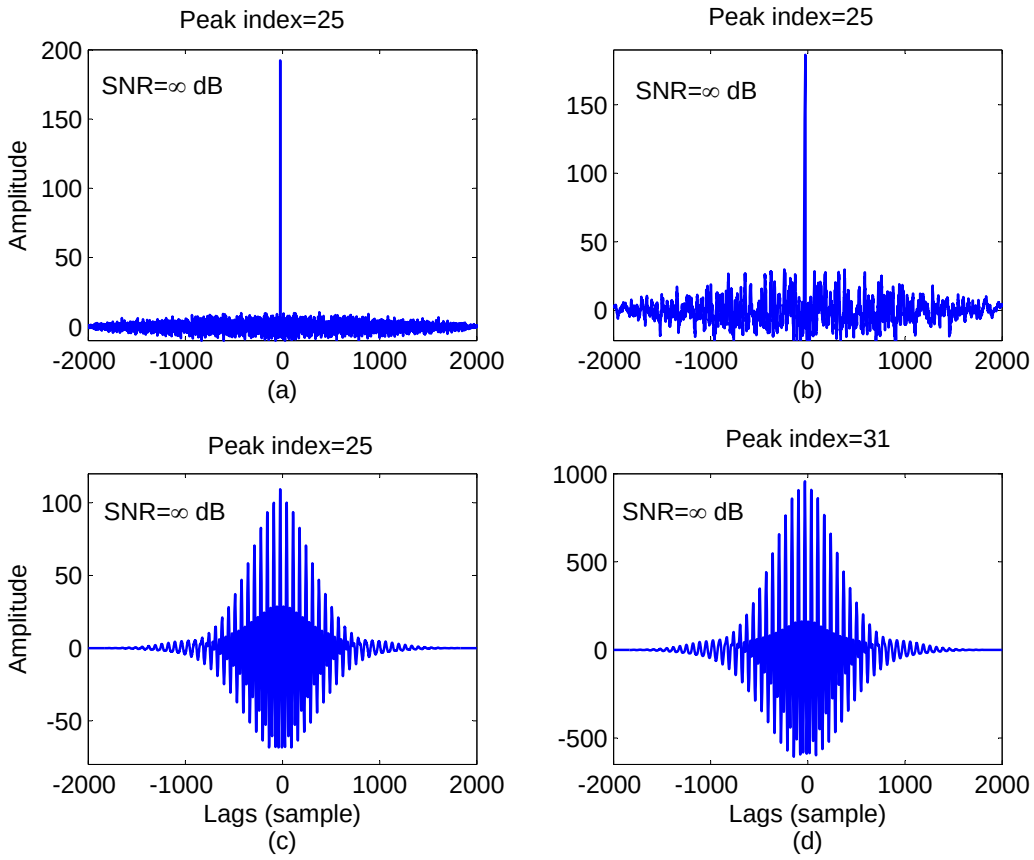

Figure 3.13: The CC for the white noise source with (a) single component RIR, and (b) multiple component RIR; and the CC for the speech segment source with (c) single component RIR, and (d) multiple component RIR.

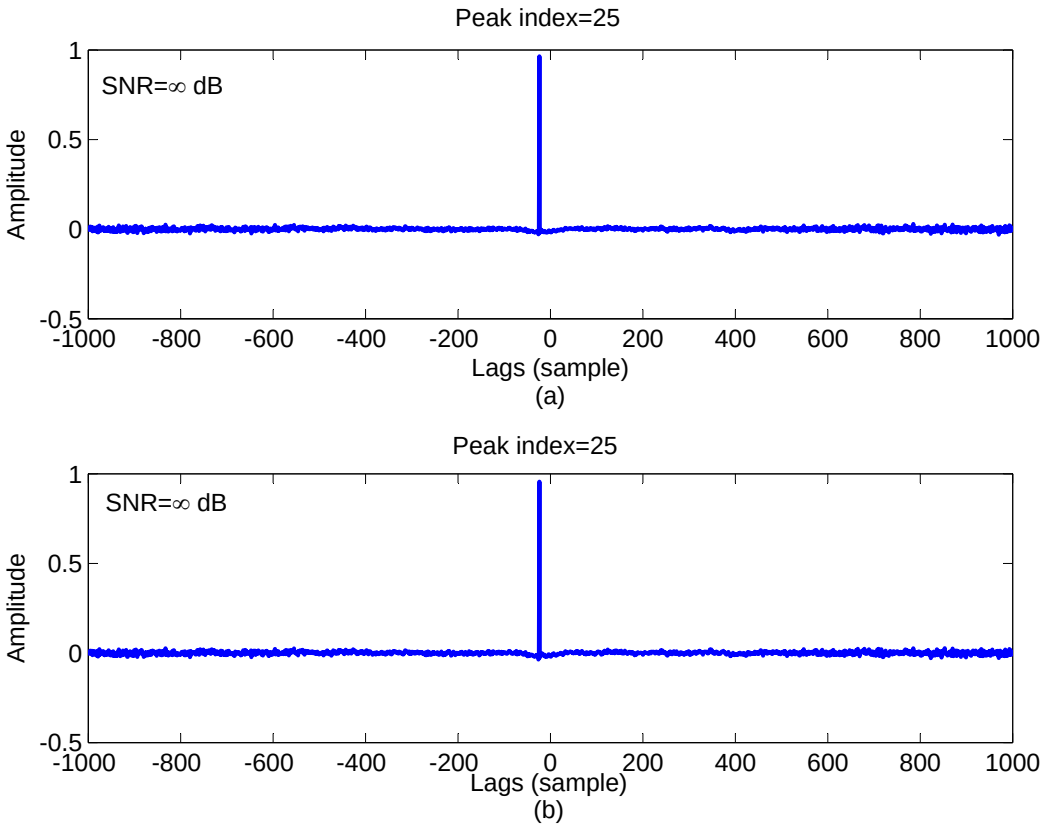

Figure 3.14: The CC for the speech segment source using all-pass processing with (a) single component RIR, and (b) multiple component RIR.

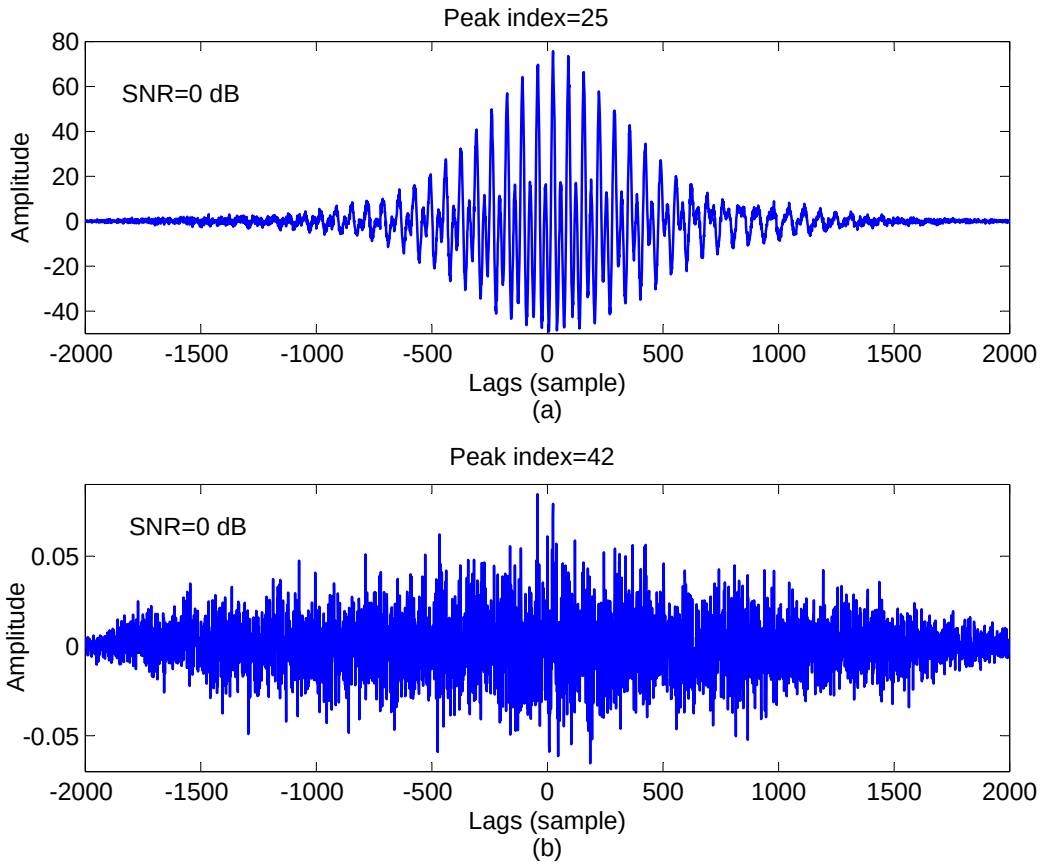

Figure 3.15: (a) CC for the speech source with a single component RIR, and (b) CC for the speech source using all-pass processing with a single component RIR. In both cases the SNR = 0 dB.

spreading of the unit impulse occurs in the CC. To examine the effectiveness of all-pass processing for TDE using the CC method, the previous test with the speech segment as the source $s[n]$ was repeated. The CC for the single component RIR is given in Fig. 3.14 (a), and correctly shows the peak at sample 25. The CC for the multiple component RIR is given in Fig. 3.14 (b), and this also shows the peak at sample 25. These results are similar to those obtained previously with the white noise segment as the source. Therefore, the problem of cross-power spectrum smearing can be solved by using all-pass processing.

The performance in noisy conditions is now examined. Let $n^{(1)}[n]$ and $n^{(2)}[n]$ be the additive noise for the first and second received signals, respectively. The cross-

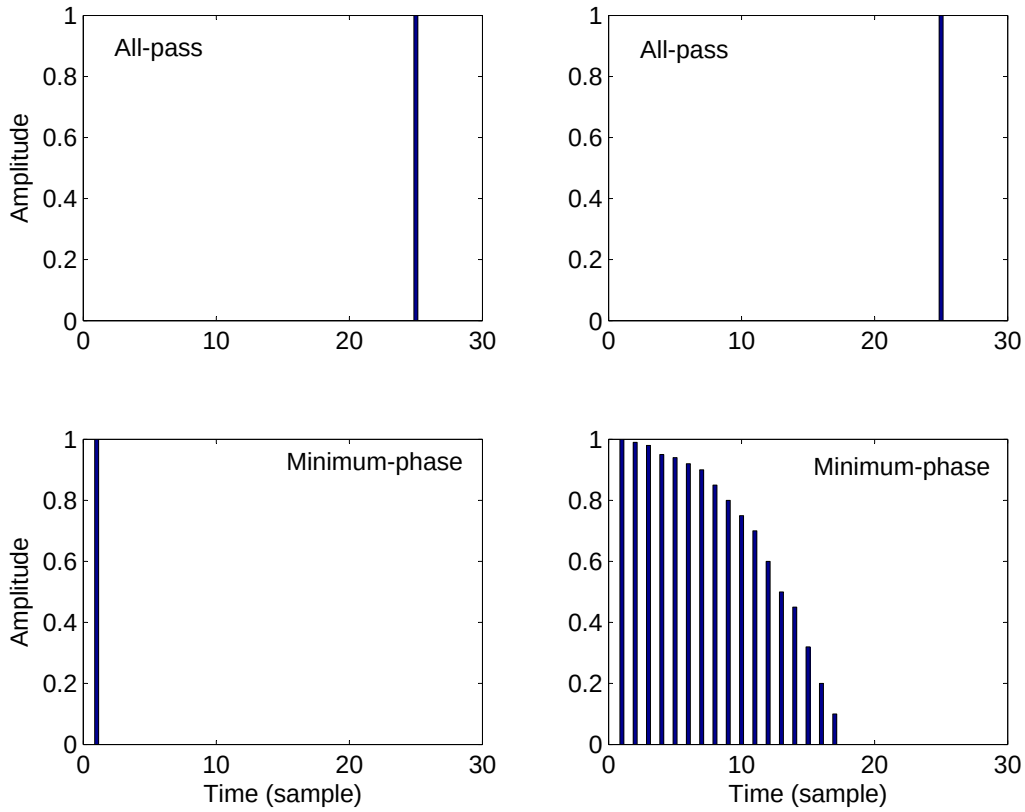

Figure 3.16: Minimum-phase and all-pass components for the RIRs of Fig. 3.12: left plots for the single component RIR and right plots for the multiple component RIR.

power spectrum for a single component RIR can be approximated as[2]

$$G_{x^{(1)}x^{(2)}}(f) \approx \alpha G_{ss}(f)e^{-j2\pi f n_d} + G_{n^{(1)}n^{(2)}}(f), \tag{3.39}$$

where $G_{n^{(1)}n^{(2)}}(f)$ is the cross-power spectrum of the noise. When the SNR is not very low (e.g. SNR ≥ 0), $G_{x^{(1)}x^{(2)}}(f)$ can be characterized by the source spectrum $G_{ss}(f)$ which has a non-flat spectrum and is largely sparse. Thus the cross-power spectrum has characteristics similar to those of the source spectrum, so the CC method should provide adequate performance under noisy conditions as long as the first term in (3.39) is much larger than the second term. However, when all-pass processing is used, (3.38) indicates that all-pass component of the source signal has a flat spectrum and thus a cross-power spectrum which is similar to that of the noise. Therefore the cross-power spectrum is more sensitive to noise when all-pass processing is used. To illustrate this, the CC for the speech segment given in Fig. 3.11 (b) using the single component RIR in Fig. 3.12 (a) when SNR = 0 dB is shown in Fig. 3.15 (a). This CC is broader than of the noise-free result given in Fig. 3.13 (c), but the peak still occurs at sample 25. However, when all-pass processing is employed the CC is smeared significantly and appears random as shown in Fig. 3.15 (b). This CC has a peak at sample 42, which is far from correct. Thus, while all-pass processing improves the CC performance when the noise is not significant, this will be degraded by higher noise levels, resulting in performance worse than without this processing.

Figure 3.16 gives the all-pass and minimum-phase components of the single component RIR (left plots) and the multiple component RIR (right plots). This shows that the all-pass component not only preserves the delay information (at sample 25), but reduces the effects of secondary peaks after the direct path (called early reverberation). In the next section, this feature is exploited to improve the TDE.

TDE using All-Pass Processing in a Reverberant Environment

Reverberation typically affects the minimum phase and all-pass components of the RIR differently, so these components will have distinct features. Two RIRs generated using the image method [3] with RT60 = 200 ms are shown in Fig. 3.17 (a)-(b). Their minimum-phase and all-pass components are shown in Fig. 3.17 (c)-(d) and (e)-(f), respectively. The minimum phase components contain a dominant peak at the origin

[2]Here the noise and signal are assumed to be uncorrelated so the noise and signal cross terms are ignored.

followed by several secondary peaks of smaller amplitude with an envelope that decays rapidly, so the energy is concentrated near the origin [61]. On the other hand, the all-pass component preserves the direct path delay information as it has a dominant peak at the same position as in the original RIR. Comparing the secondary peaks of the all-pass component with those of the original RIR, it is clear that the early reverberation energy is greatly attenuated in the all-pass component. Thus all-pass processing can be used to reduce the early reverberation while preserving the direct path delay information, which is ideal for TDE applications. To better illustrate this characteristic, 15 different RIRs with different speaker-microphone positions for RT60 ranging from 200 to 1200 ms were generated using the image method [3]. The Early to Late Reverberation energy Ratio (ELRR) is defined as

$$\text{ELRR} = 10\log_{10}\left(\frac{\sum_{n=n_d}^{N} h^2[n]}{\sum_{n=n_d}^{\infty} h^2[n] - \sum_{n=n_d}^{N} h^2[n]}\right), \tag{3.40}$$

where n_d is the direct path position in samples and N is the boundary which determines the early reverberation. Here, this boundary is chosen such that subsequent reflection amplitudes are no more than 10% of the direct path amplitude. The average ELRR for the 15 different RIRs for each value of RT60 and for the corresponding all-pass components is shown in Fig. 3.18. It is clear that the ELRR is lower for the all-pass components regardless of the reverberation time. Thus, using the all-pass component will decrease the reverberation effects, in particular those due to early reverberation.

The decomposition of the speech signals $x[n]$ into their minimum phase $x_{min}[n]$ and all-pass $x_{all}[n]$ components using homomorphic filtering [61] is shown in Fig. 3.19. The input speech sequence is first zero-padded and then the cepstrum sequence $c_x[n]$ is determined. This is achieved by taking the FFT of $x[n]$ to get $X(f)$, then calculating the complex logarithm of $X(f)$, and finally taking the IFFT. The complex cepstrum of the minimum phase sequence $c_x^{min}[n]$ is obtained by multiplying $c_x[n]$ with $2u[n] - \delta[n]$ where $u[n]$ and $\delta[n]$ are the unit step and unit impulse functions, respectively. Taking the FFT and then the exponential of $c_x^{min}[n]$ gives the minimum phase component in the frequency domain, $X_{min}(f)$. The IFFT of $X_{min}(f)$ is the minimum phase component $x_{min}[n]$. Finally, the all-pass component in the frequency domain $X_{all}(f)$ is obtained by dividing $X(f)$ by $X_{min}(f)$. The IFFT of $X_{all}(f)$ is the all-pass component $x_{all}[n]$. Note that phase wrapping should not be done in the

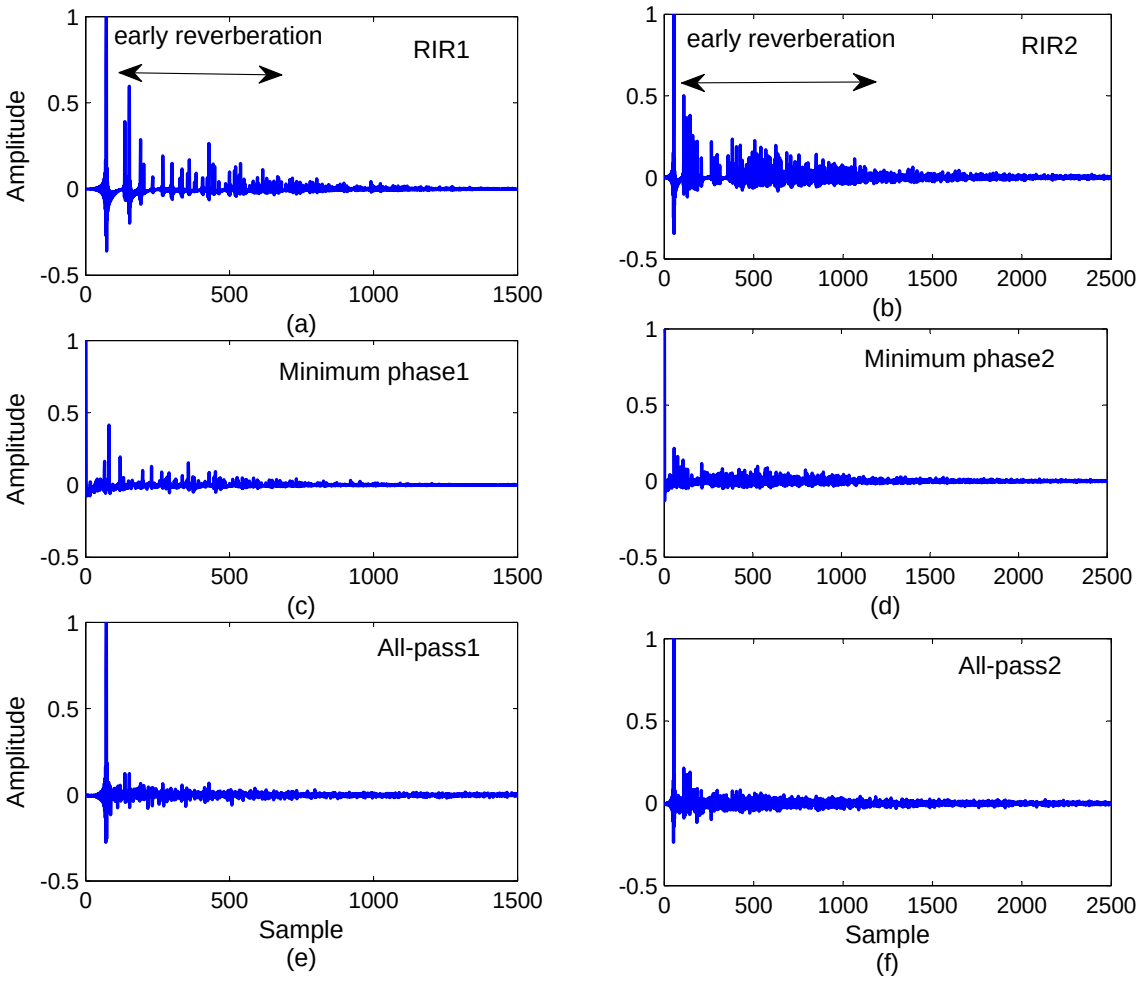

Figure 3.17: (a)-(b) two RIRs with RT60= 200 ms generated using the image method, (c)-(d) the corresponding minimum-phase components, and (e)-(f) the corresponding all-pass components.

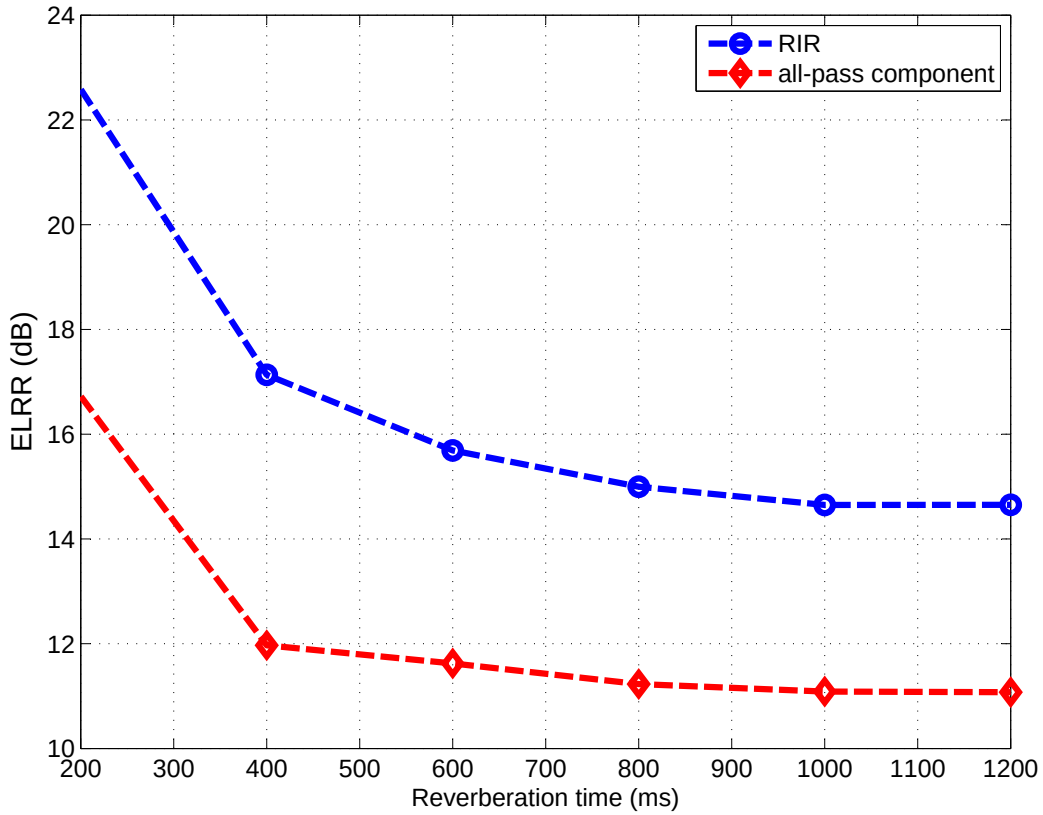

Figure 3.18: The Early to Late Reverberation energy Ratio (ELRR) for the RIRs and the corresponding all-pass components.

FFTs and IFFTs.

The Proposed TDE Method in a Noisy Reverberant Environment

Figure 3.17 (e)-(f) shows that reverberation effects remain in the all-pass component in the form of small amplitude late impulses after the direct path. This will affect the TDE especially when the reverberation time is high. In addition, noise will degrade TDE performance. Spectral subtraction is used to deal with the problems of late reverberation and additive noise. Thus the noise reduction method proposed in [25] is first employed as it provides reasonably good performance in the presence of reverberation. Then, the spectral subtractions method proposed in [19] is used to deal with the remaining reverberation effects. Finally, all-pass processing is used.

The noise reduction technique in [25] proceeds as follows. First the Short-Time Fast Fourier (STFT) transform of $x[n]$, denoted $X(l, f), l = 0, \ldots, L-1, f = 0, \ldots, F-1$, is calculated using a Hamming window of 16 ms duration with an 8 ms overlap

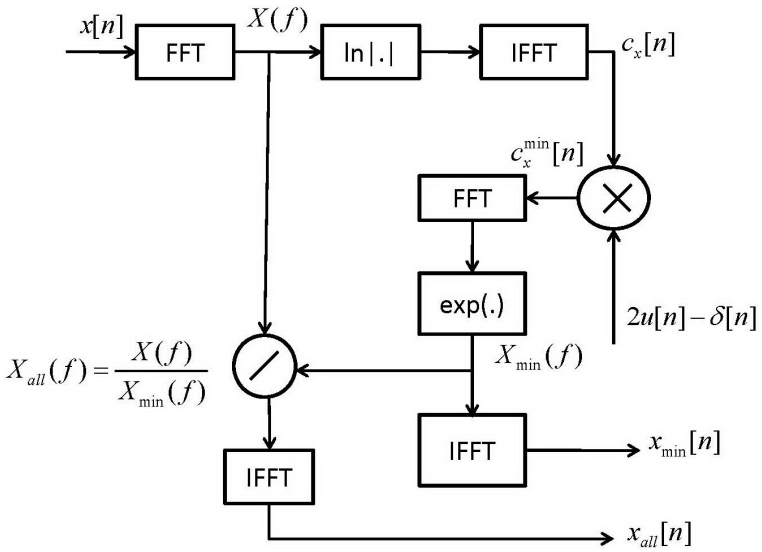

Figure 3.19: Block diagram of the homomorphic filtering for minimum phase and all-pass component decomposition.

between frames. Then the amplitude of the enhanced speech spectra $\overline{X}(l,f)$ is obtained by multiplying the amplitude of $X(l,f)$ with the spectral denoising weight filter $G_n(l,f)$ giving

$$|\overline{X}(l,f)| = |X(l,f)|G_n(l,f). \tag{3.41}$$

The optimal denoising spectral weighting can be calculated as follows [25]

$$G_n(l,f) = \begin{cases} \min\left(0.1\sqrt{\frac{P_{noise}(l,f)}{|X_{all}(l,f)|^2}}, 1\right) & \text{for } \sqrt{\frac{P_{noise}(l,f)}{|X_{all}(l,f)|^2}} \geq \frac{1}{o(l,f)+0.1} \\ 1 - o(l,f)\sqrt{\frac{P_{noise}(l,f)}{|X_{all}(l,f)|^2}} & \text{otherwise} \end{cases}, \tag{3.42}$$

where $P_{noise}(l,f)$ is the noise Power Spectral Density (PSD) which can be estimated using the minimum statistics approach as long as the noise is reasonably stationary [33]. $o(l,f)$ is the subtraction factor which depends on the SNR and is given by [25]

$$o(l,f) = \begin{cases} \sqrt{1 - \frac{2}{25} \times \min\left(\max\left(10\log_{10}\frac{\sum_{f=0}^{F-1}|X_{all}(l,f)|^2}{\sum_{f=0}^{F-1}P_{noise}(l,f)},\ -5\right),\ 20\right) - 20} \\ \quad \text{for } \sum_{f=0}^{F-1} P_{noise}(l,f) > 0;\ f = 0,\ldots,F-1 \\ 1 \quad \text{for } \sum_{f=0}^{F-1} P_{noise}(l,f) = 0;\ f = 0,\ldots,F-1 \end{cases}$$

Next spectral subtraction is used to reduce the late reverberation as follows. The PSD of the late impulse components is calculated as [19]

$$P^{late}(l,f) = 0.32w(l - D_l) \star |\overline{X}(l,f)|^2, \tag{3.43}$$

where $w[n]$ is assumed to be the Rayleigh distribution [19]. Assuming a 50 ms delay between the early and late impulses [19] and considering a frame shift of 8 ms for FFT analysis, the respective delay for the weight function D_l is set to 7. The final enhanced speech spectrum is obtained by multiplying the amplitude of $\overline{X}(l,f)$ with the spectral dereverberation weight filter $G_d(l,f)$ giving

$$|\widehat{X}(l,f)| = |\overline{X}(l,f)|G_d(l,f). \tag{3.44}$$

The optimal dereverberation spectral weighting is then calculated as follows

$$G_d(l,f) = \max\left\{\sqrt{\frac{|\overline{X}(l,f)|^2 - P^{late}(l,f)}{|\overline{X}(l,f)|^2}}, 0.02\right\}. \tag{3.45}$$

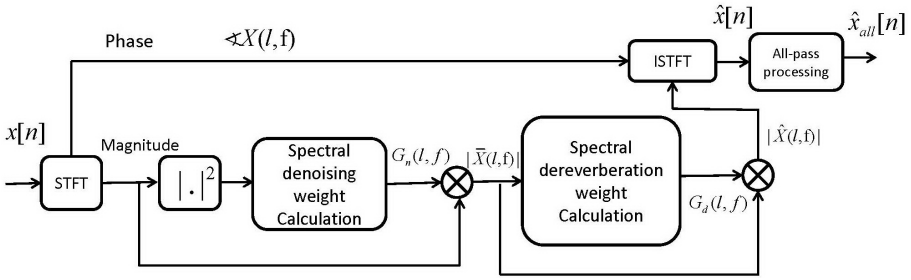

Figure 3.20: Block diagram of the proposed preprocessing stages for TDE methods.

Finally, the modified magnitude spectrum and the original phase of $X(l, f)$ are combined, and an Inverse Short-Time Fourier Transform (ISTFT) using the overlap-add technique is employed to obtain the enhanced signal.

Finally, all-pass processing is used to decrease the early reverberation. While the spectral subtraction and all-pass processing operations can be done in any order, our extensive experiments show that spectral subtraction followed by all-pass processing for the GCC-based TDE method provides better performance.

A block diagram of the preprocessing stages for the proposed TDE method is shown in Fig. 3.20. The input signal is first transformed to the frequency domain using the STFT with a Hamming window of size 128 ms with a 50% overlap. The speech signal is then processed using the spectral denoising weight given in (3.42). The resulting signal is used to obtain the spectral dereverberation weight from (3.45), the corresponding enhanced spectrum is combined with the phase of the original signal to get the final enhanced signal using an ISTFT. Finally, all-pass processing is performed using the procedure in Fig. 3.19 to obtain the all-pass component. After performing these preprocessing stages for both received speech signals, the TDOA is estimated using the GCC method [51] or CC method [50].

3.1.3 The Cramer-Rao Lower Bound of the Reverberant Speech Signal in Noisy Speech

In this section, we formulate again the Cramer-Rao Lower Bound (CRLB) for a more general noisy reverberant speech signal which previously was calculated for simple case of a delayed source signal plus noise in [64]. Let us assume that the received

speech signal are denoted by $x_1[n]$ and $x_2[n]$ as

$$x_i[n] = y_i[n] \star n_i[n], \tag{3.46}$$
$$y_i[n] = s[n] \star h_i[n] \quad i = (1, 2) \tag{3.47}$$

where $y_i[n]$ and $h_i[n]$ are the reverberant speech and RIR for i-th microphone, respectively. In frequency-domain, the above equations can be written as

$$X_i(f) = Y_i(f)N_i(f), \tag{3.48}$$
$$Y_i(f) = S(f)H_i(f) \quad i = (1, 2) \tag{3.49}$$

where $X_i(f)$, $Y_i(f)$, $N_i(f)$, $H_i(f)$ and $S(f)$ are the FFTs transform of $x_i[n]$, $y[n]$, $n_i[n]$, $h_i[n]$, and $s[n]$, respectively. The signal coherence can be defined as follows.

$$\gamma(f) = \frac{X_1(f)conj(X_2(f))}{\sqrt{|X_1(f))|^2|X_2(f))|^2}} \tag{3.50}$$

where $conj(.)$ denotes complex conjugate. The signal coherence has to be confined as $\gamma(f) \leq 1$. The CRLB can be written as [64]

$$CRLB = 2\pi \left(\sum_{f=0}^{F-1} (\frac{2\pi}{F} f)^2 \times SNR_{TD}(f) \right)^{-1} \tag{3.51}$$

where F is the number of frequency bins. $SNR_{TD}(f)$ is a SNR-like expression that is defined as follows.

$$SNR_{TD}(f) = \left\{ \frac{1}{|\gamma(f)|^2} \left[1 + \left(\frac{|Y_1(f)|^2}{|N_1(f)|^2} \right)^{-1} \right] \left[1 + \left(\frac{|Y_2(f)|^2}{|N_2(f)|^2} \right)^{-1} \right] - 1 \right\}^{-1} \tag{3.52}$$

In Section 3.1.4, we will use the $\sqrt{CRLB}$ as the standard deviation of the CRLB for our comparison.

3.1.4 Performance Results

In this section, the performance of the TDE methods is evaluated and compared in reverberant and noisy conditions. Towards this end, 8 different utterances (4 male speaker and 4 female speaker) from the TIMIT database sampled at 16 kHz are used.

The image method [3] was used to generate the RIRs for a $5 \times 4 \times 6$ m^3 rectangular room assuming omnidirectional microphones. The proposed TDE methods based on AIF, and the CC method [50] and GCC method (PHAT is chosen here) [51] with the proposed preprocessing stages, are compared with the eigenvalue decomposition method (EVD) [56] and the original CC and GCC methods. A Hamming window of length 128 ms was used with the GCC-based methods. The two preprocessing stages (spectral subtraction and all-pass processing), are considered separately and combined with the CC and GCC methods to clearly illustrate their effects. They are denoted "s" for spectral subtraction and "ap" for all-pass processing. Thus, "CC-ap" represents the CC method with only all-pass processing and "GCC-s-ap" denotes the GCC method with both spectral subtraction and all-pass processing.

Ten RIRs with different microphone-speaker positions having RT60 in the range 200 ms to 1200 ms were generated. For these RIRs, the source location is [2 1 1] (m), and the microphone locations are

[1.5 2 2.5], [2 1 2], [1 2 3], [1 1 1], [2 3 1], [3 1 2], [1 1 2], [1 .5 2], [0.5 1 1], and [2 1 2.5] (m)

The Direct to Reverberation Ratio (DRR) values for the RIRs with ten different microphone positions are shown in Fig. 3.21. The DRR values for the RIRs are calculated according to [27]. The TDOA estimate for each RT60 value was obtained for all 45 possible pairs of RIRs with the 8 utterances using the TDE methods. An acceptable estimate is defined according to (3.24).

Performance in Reverberant Noise-Free Conditions

In this section, the performance of the TDE methods is examined under different reverberant conditions when SNR $= \infty$. The average estimation error and the standard deviation (STD) of this error for the acceptable estimates for each of the methods are shown in Figs. 3.22 and 3.23, respectively.

These figures show that both spectral subtraction and all-pass processing improve the performance of the methods. However, all-pass processing provides better performance than spectral-subtraction. It can be seen that using both preprocessing stages for the GCC method results in better performance than the GCC-based TDE methods. These results also confirm that the proposed AIF method is very accurate even in highly reverberant conditions. To the best of our knowledge, no other TDE method based on two microphones can provide similar performance. The proposed

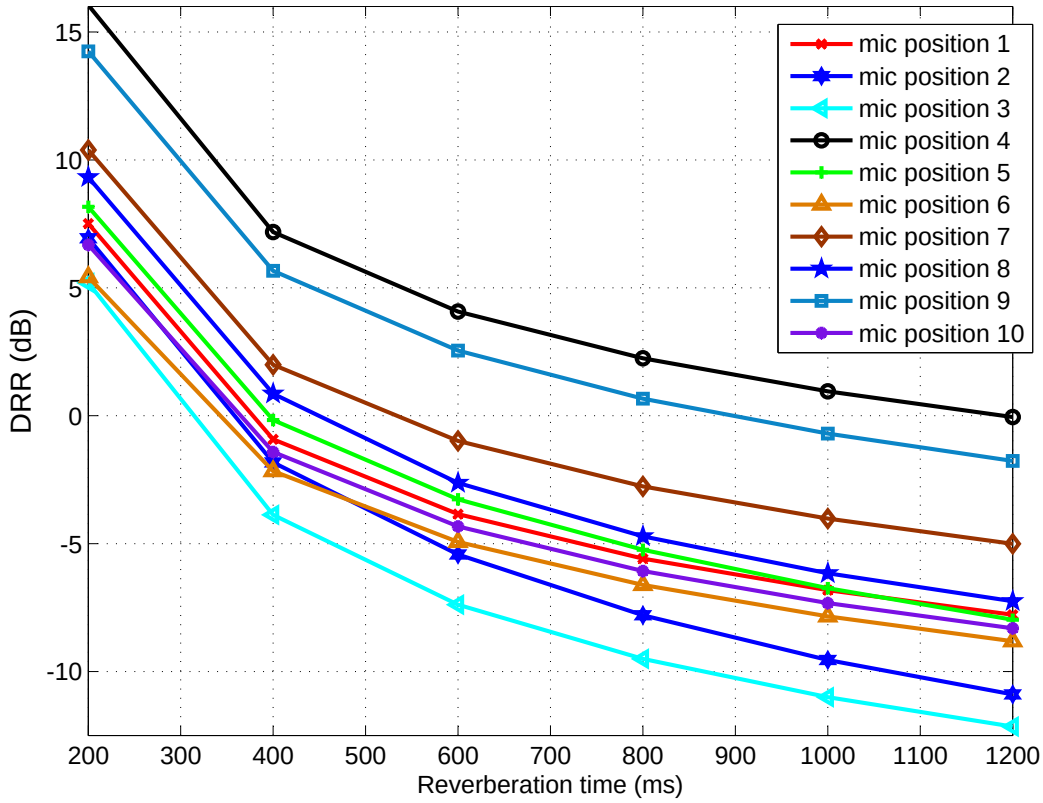

Figure 3.21: The DRR values for the RIRs with ten different microphone positions, having RT60 in the range 200 ms to 1200 ms.

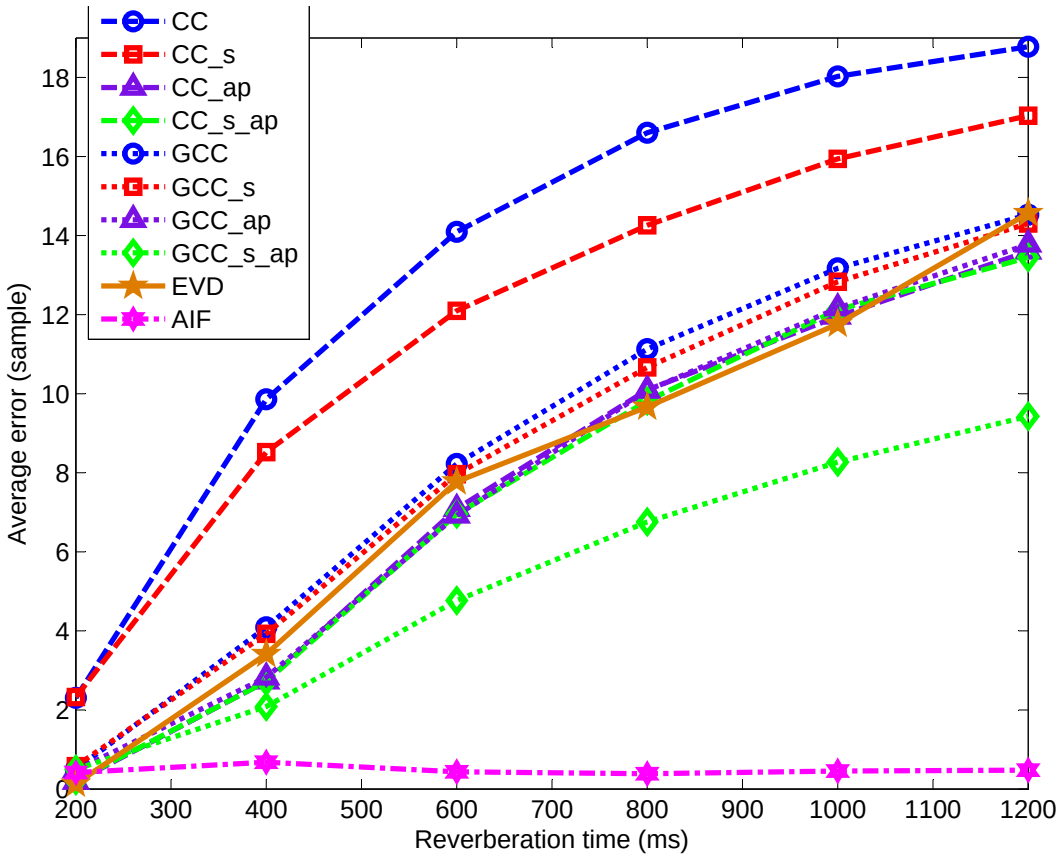

Figure 3.22: TDOA average estimation error for the TDE methods in different reverberant environments using 8 speech utterances.

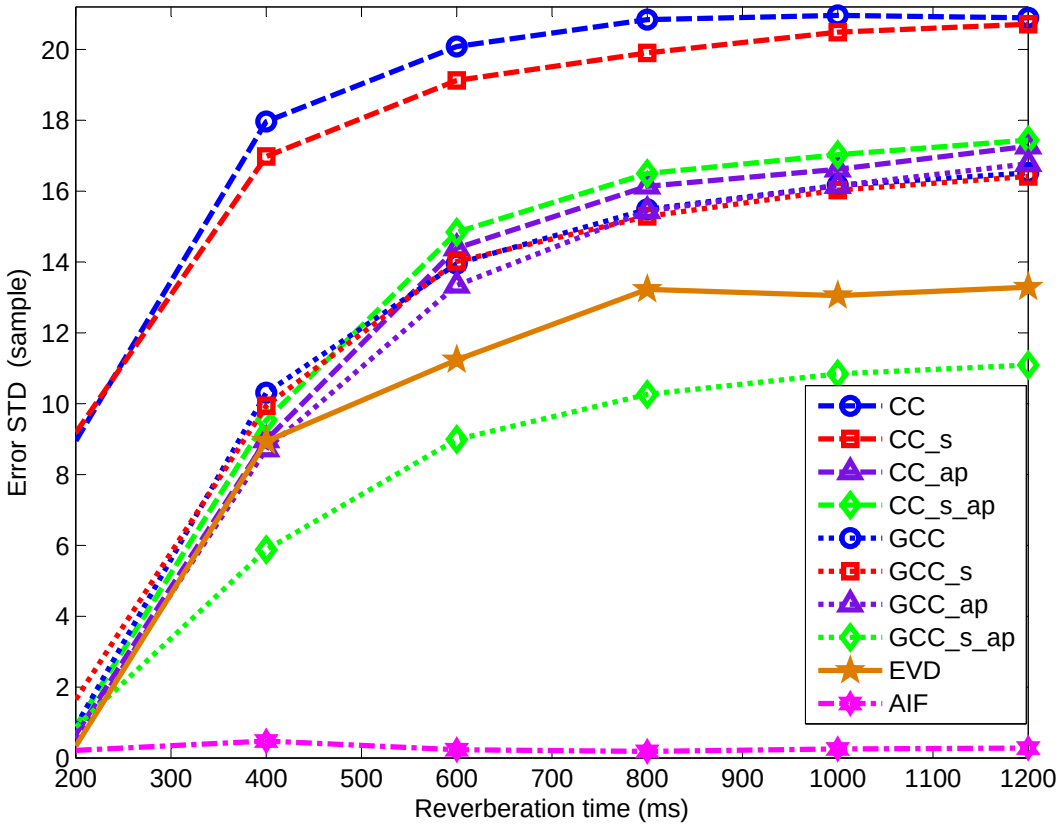

Figure 3.23: TDOA estimation error standard deviation (STD) for the TDE methods in different reverberant environments using 8 speech utterances.

TDE method using both spectral subtraction and all-pass processing ("GCC-s-ap") has the next best performance with a low estimation error even with high reverberation.

Performance in Noisy Reverberant Conditions

In this section, the performance of the TDE methods is evaluated in different noisy reverberant conditions. Ten RIRs with RT60 = 400 ms and different speaker-microphone positions were employed. The reverberant speech signal was added to ten realizations of computer-generated white Gaussian noise with SNRs ranging from 40 dB to -10 dB. The average estimation error and STD of this error for the acceptable estimates are shown in Figs. 3.24 and 3.25, respectively. It can be seen from Fig. 3.24 that when SNR $>$ 25 dB, all-pass processing improves the performance of both the GCC and CC methods (compare "GCC-ap" with "GCC" and "CC-ap" with "CC"). As expected, spectral subtraction also provides an improvement so that

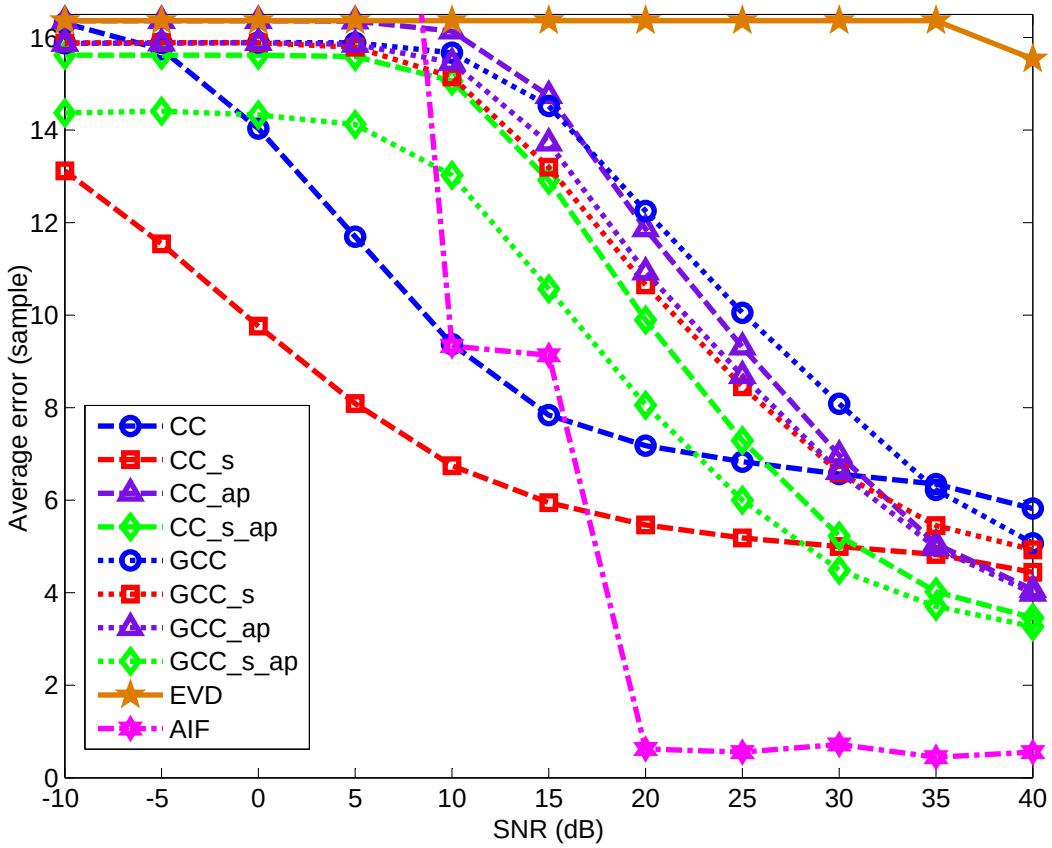

Figure 3.24: TDOA average estimation error for the TDE methods in different noisy reverberant environments using 8 speech utterances with RT60 = 400 ms.

"GCC-s-ap" and "CC-s-ap" are better when SNR > 25 dB. However, when the SNR is low the CC method is better than the GCC method, and all-pass processing does not improve the performance of the CC method as discussed in Section 3.1.2. Thus the CC method with spectral subtraction ("CC-s") provides the best performance compared to the other TDE methods at low SNR levels. The EVD method based on two microphones does not work well in noisy conditions. Further, these results confirm that the proposed AIF method has much better performance when SNR > 10 dB. The Cramer-Rao lower bound (CRLB) for the STD [64] is also given in Fig. 3.25. Only the proposed AIF method is comparable to the CRLB when SNR $\geq$ 20 dB. In addition, this figure shows that spectral subtraction preprocessing improves the STD of the CC method in low SNR conditions, which confirms that the CC method with this preprocessing ("CC-s") performs better than the other TDE methods.

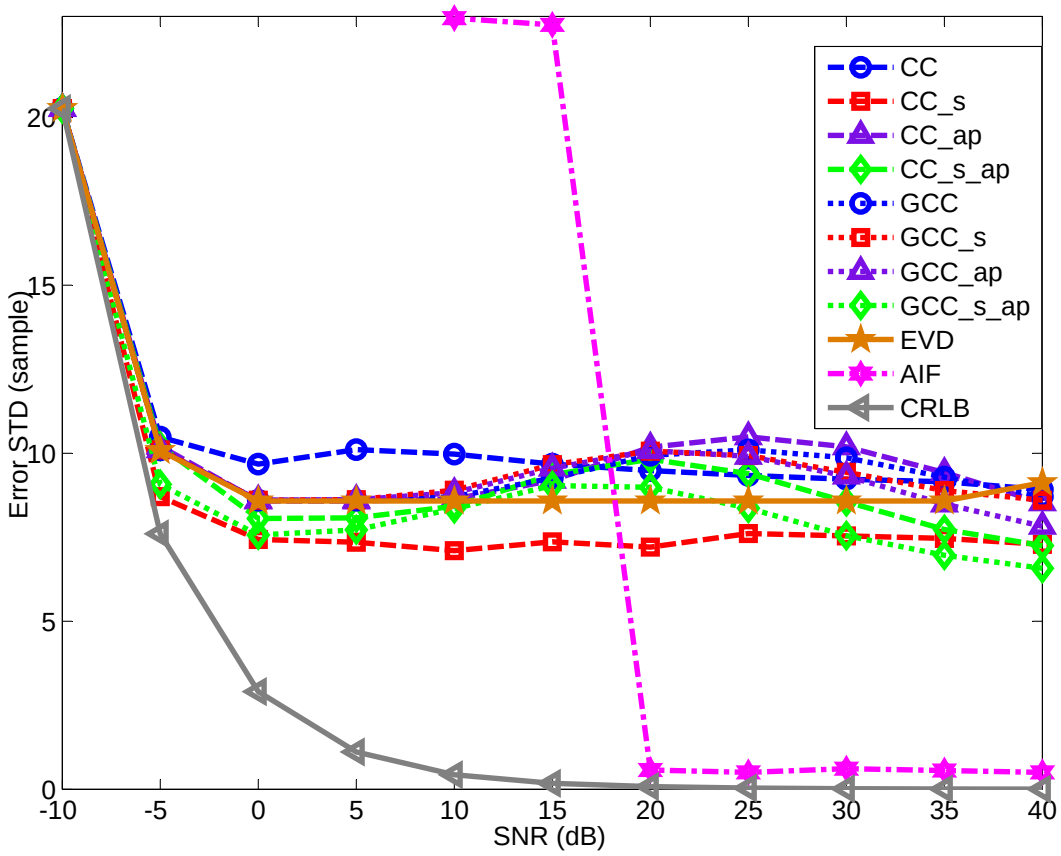

Figure 3.25: TDOA estimation error standard deviation (STD) for the TDE methods in different noisy reverberant environments using 8 speech utterances with RT60 = 400 ms.

Performance with a White Gaussian Noise Input Signal

Thus far, a speech signal has been used as the input. While this is typical in audio applications, there are many situations where the TDOA must be estimated for input signals other than speech. White Gaussian noise has a symmetric pdf and thus methods such as AIF cannot be used with this type of input signal. Conversely, GCC methods are more general and thus suitable for any input signal. Again ten RIRs with different microphone-speaker positions having reverberation times in the range 200 ms to 1200 ms are used. This is the same as in Section 3.1.4 except that the input is a white Gaussian noise signal. The average estimation error and the STD of this error for the acceptable estimates obtained using the TDE methods are shown in Figs. 3.26 and 3.27, respectively. It can be seen from these figures that spectral subtraction does not provide any improvement for the CC and GCC methods when the input is white Gaussian noise. However, all-pass processing results in a significant improvement with both methods. In fact, GCC with all-pass processing ("GCC-ap") has the best performance followed by CC with this processing.

Performance in a Real Room Environment

In this section, the performance of the TDE methods is evaluated in a real reverberant environment. Results were obtained for the meeting room binaural RIR from the Aachen Impulse Response (AIR) database [41] which has RT60 = 0.67 s. The RIR was measured without a dummy head using only the left channel. Five microphones in different locations were used for the meeting room [41]. As before, 8 clean utterances (4 female and 4 male speakers), were convolved with the measured RIR to obtain the reverberant speech signals. The average TDOA estimation error and error standard deviation (STD) for the TDE methods are presented in Table 3.1. These results show that spectral subtraction only slightly improves the performance of the CC method, but all-pass processing provides a significant improvement in performance. As expected, the proposed AIF method is best among the TDE methods while the GCC-based methods with all-pass processing are the next best.

Next the real room performance is examined in noisy conditions. Computer-generated white Gaussian noise with different SNRs was added to the reverberant speech signals described above. The average and STD results for the TDE methods are shown in Figs. 3.28 and 3.29, respectively. It can be seen that all-pass processing improves the performance of the CC method when SNR > 20 dB, but fails to work for

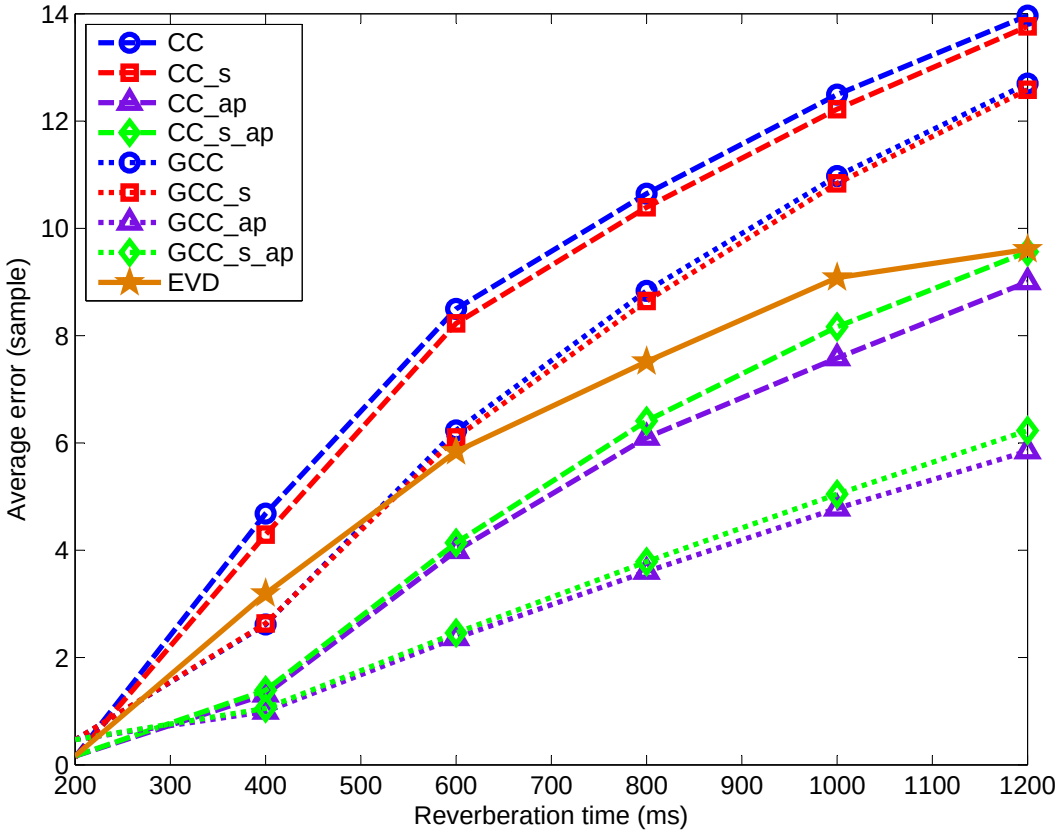

Figure 3.26: Average TDOA estimation error for the TDE methods in different reverberant environments for a white Gaussian input signal.

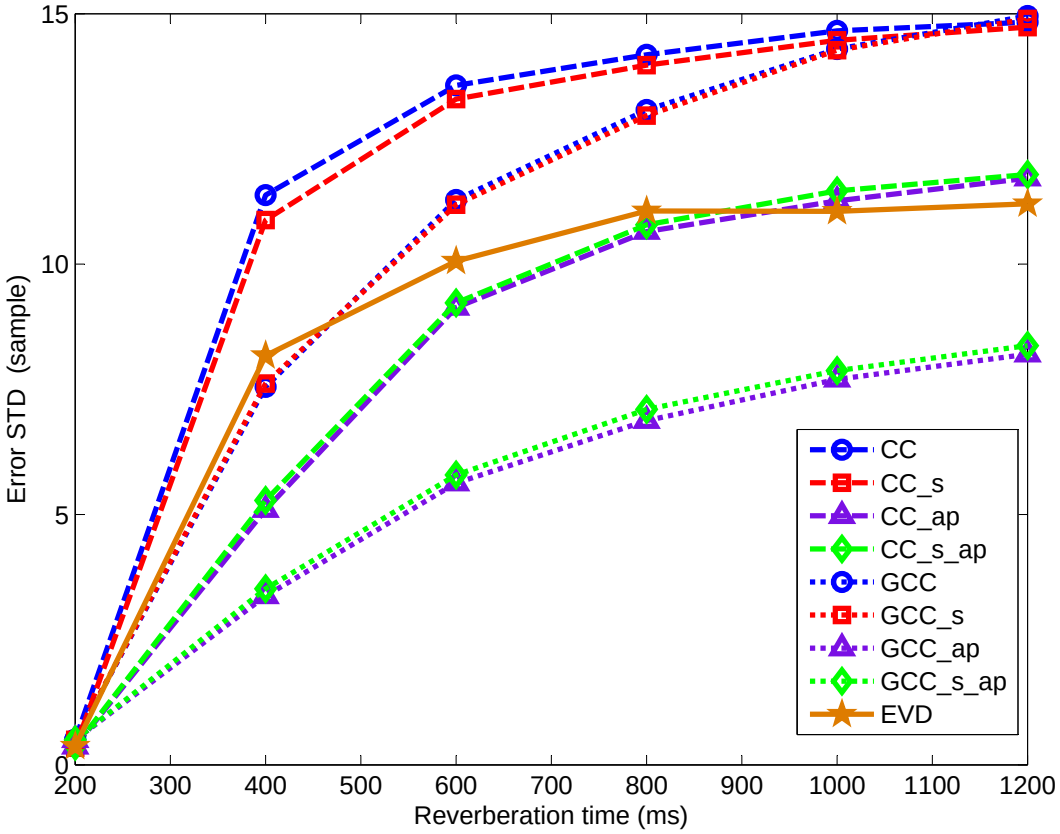

Figure 3.27: TDOA estimation error standard deviation (STD) for the TDE methods in different reverberant environments for a white Gaussian input signal.

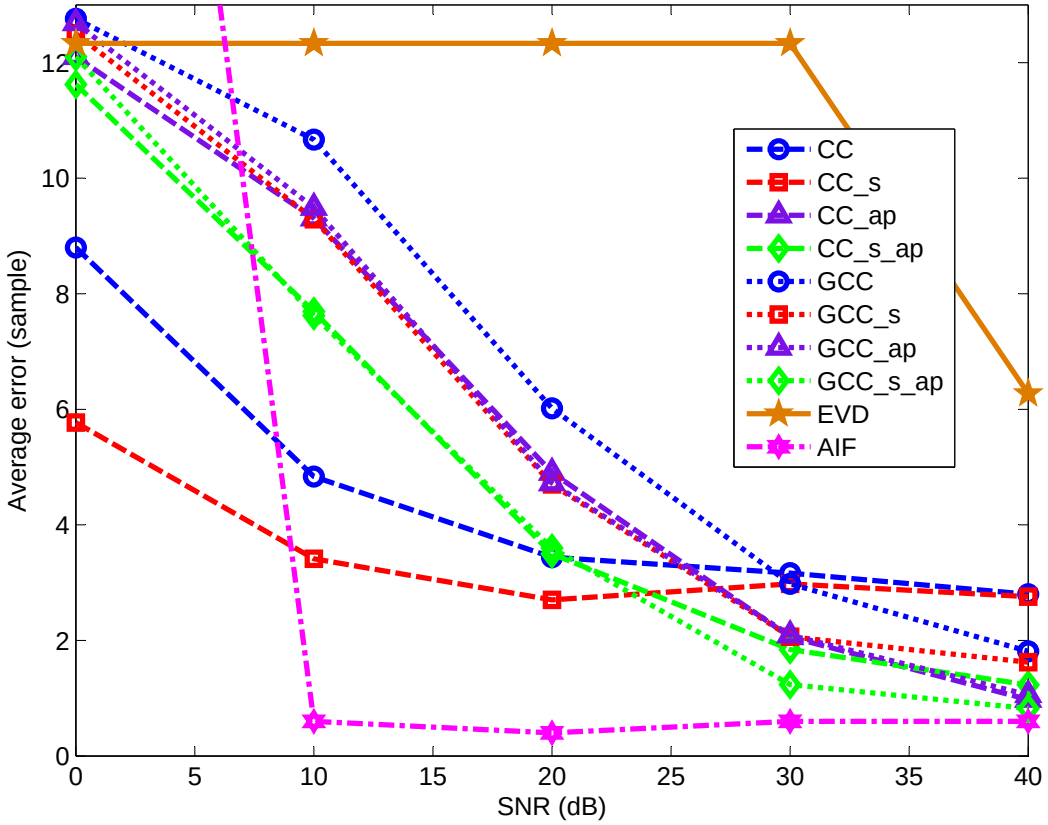

Figure 3.28: Average TDOA estimation error for the TDE methods in a real meeting room with RT60 = 0.67 s and additive white Gaussian noise.

Table 3.1: Average TDOA Estimation Error and Error Standard Deviation (STD) for the TDE Methods in a Real Meeting Room with RT60 = 0.67 s.

	CC	CC-s	CC-ap	GCC	GCC-s	GCC-ap	EVD	AIF
average	2.47	2.13	**0.79**	1.52	1.52	**0.98**	1.20	**0.50**
STD	3.60	2.91	**2.07**	2.44	2.50	**2.18**	2.29	**0.51**

this method at higher noise levels. However, this preprocessing improve the performance of the GCC method for all noise levels. Overall, the CC method with spectral subtraction preprocessing ("CC-s") has the best performance in low SNRs, while the GCC-based methods with both preprocessing stages are best at higher SNRs. As expected, the proposed AIF method provides the most accurate results when SNR > 10 dB, which confirms the effectiveness of this method in real conditions.

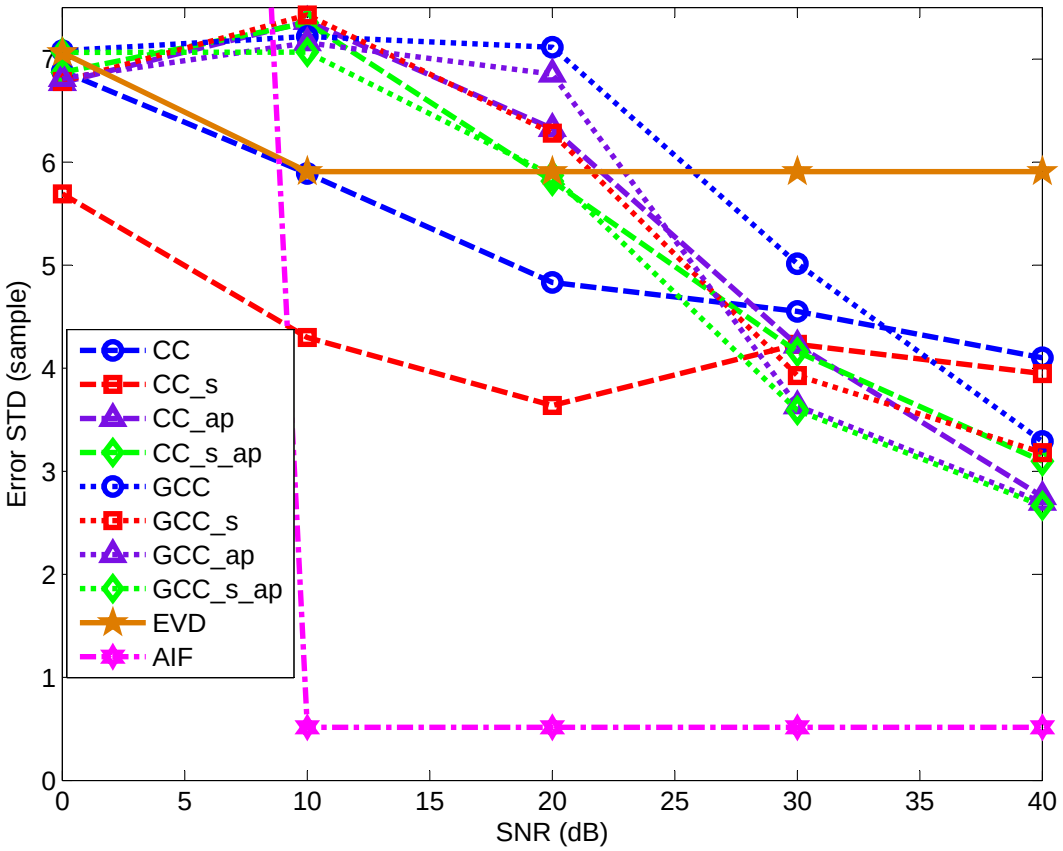

Figure 3.29: TDOA estimation error standard deviation (STD) for the TDE methods in a real meeting room with RT60 = 0.67 s and additive white Gaussian noise.

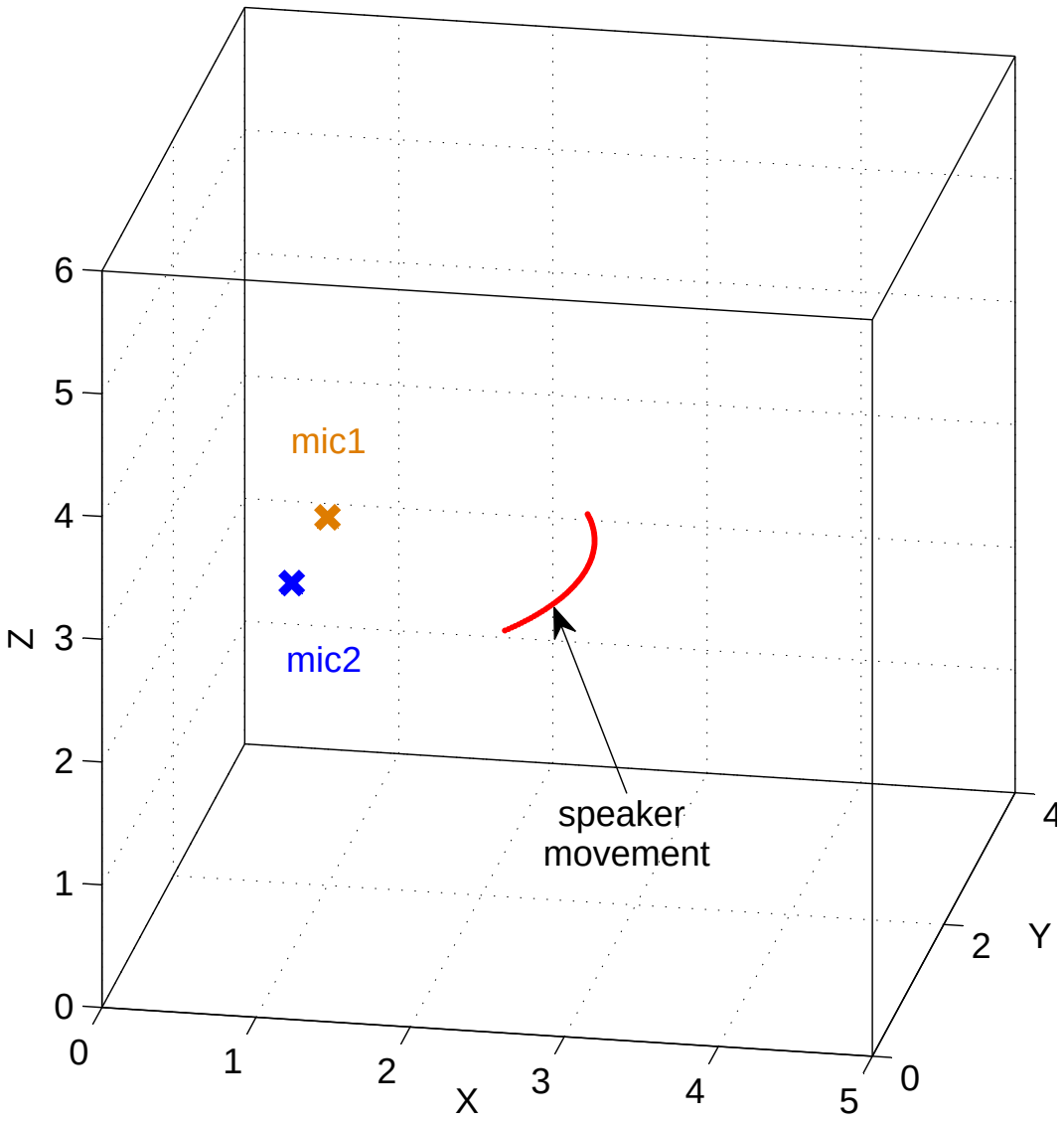

Figure 3.30: The position of the two microphones and the speaker movement in a room.

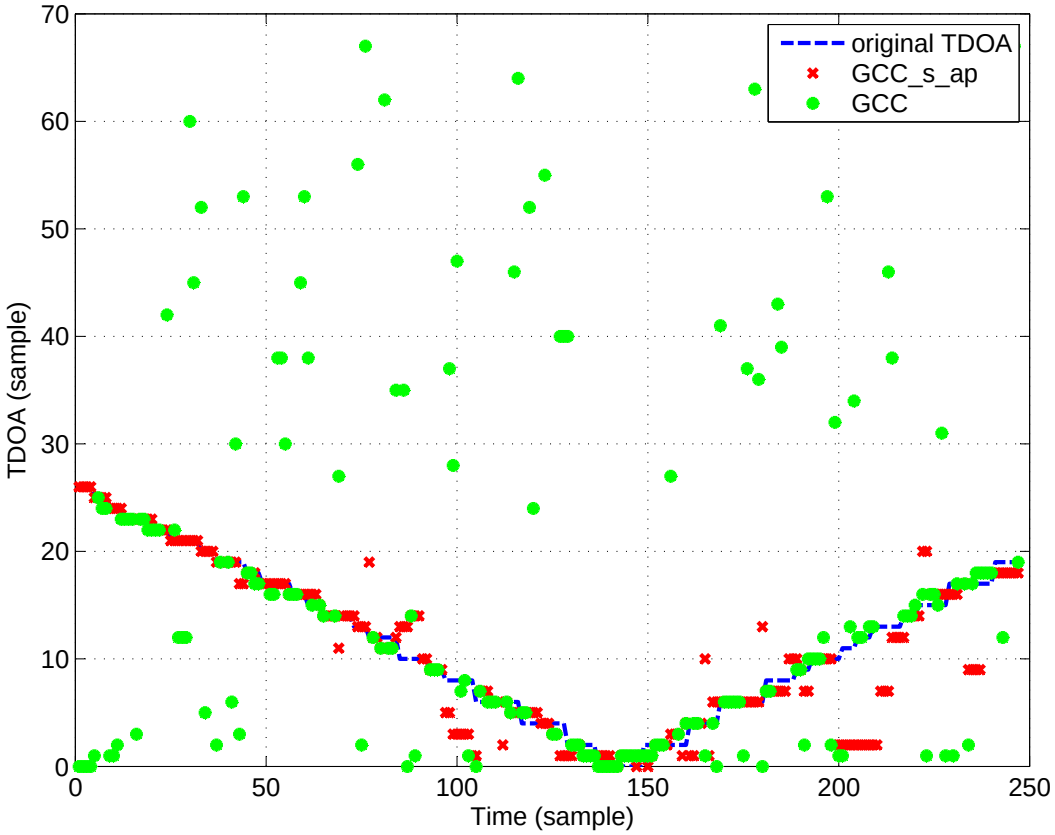

Figure 3.31: TDOA estimation in a time-varying reverberant environment using the proposed method (GCC with spectral subtraction and all-pass processing), and the GCC method alone [51].

Performance in Time-Varying Reverberant Conditions

In this section, the performance of the proposed TDE method using all-pass component with spectral subtraction is evaluated in a time-varying reverberant environment having a high reverberation time of 1000 ms. The performance is compared with the GCC method as it has been shown to be effective in high reverberation, time-varying conditions. Two microphones are fixed at locations [1 1 3] (m) and [1 2 3] (m) in a rectangular room with dimensions [5 4 6] (m). The initial speaker position is [2.5 0.5 3] (m) and he moves every 128 ms as shown in Fig. 3.30 [3]. The TDOA was estimated using the GCC method with and without the two preprocessing stages every 128 ms. The estimation results in the time progression compared with the actual TDOA are given in Fig. 3.31. This shows that the proposed method ("GCC-s-ap") follows the changes in the TDOA more accurately than the GCC method. As a result, the proposed method is more suitable for use in real-time applications and should provide good performance in many time-varying reverberant environments.

3.1.5 Conclusions

Two Time Delay Estimation (TDE) methods have been proposed which require signals from only two channels. The first estimates the TDOA between the microphones by adaptively estimating the inverse filters of the RIRs. This technique is very robust to reverberation, and so is accurate even in highly reverberant conditions. Moreover, it was shown that it is effective in low to moderate noise conditions. However, the computational complexity is high due to the inverse filter estimations, so it may not be suitable for real-time applications and time-varying environments. This is a common problem with most TDE methods. In addition, since this solution is based on maximizing the skewness of the LP residual of the speech for inverse filter estimation, it is not suitable for input signals with a symmetric pdf.

The second method resolves the above problems by employing two preprocessing stages to reduce the effects of reverberation and additive background noise. It was proven that all-pass processing as a preprocessing stage in the CC method [50] can be used to improve the performance in reverberant, low noise environments. This improvement is limited to high SNR conditions since otherwise the performance will be significantly degraded for the CC method. In general, all-pass processing will

[3] This was simulated using the free software available at `http://home.tiscali.nl/ehabets/signal_generator.html`.

decrease the effects of reverberation, particularly early reverberation which can be very detrimental to TDE performance. Thus all-pass processing can be used improve the performance of TDE techniques such as the GCC method [51]. Spectral subtraction was used to reduce both the noise and reverberation, making the GCC methods more robust. Using both spectral subtraction and all-pass processing as preprocessing stages results in excellent performance even in very noisy reverberant conditions. Further, the performance is good in time-varying environments, and this solution is not sensitive to the type of input signal. Since the computational complexity with this approach is low, it is suitable for real-time applications.

Performance results in noisy reverberant conditions were presented which demonstrated that the two proposed solutions are superior to other TDE methods. The performance was also evaluated in real recorded environments to show that they are more robust in real conditions. Finally, experimental result in time-varying environments were presented to illustrate the superiority of the proposed GCC-based method even in high reverberation conditions (RT60 = 1 s).

Chapter 4

Future Work

Generally, the proposed research can be categorized into two main subjects. The first concerns single-microphone speech enhancement and the second speaker localization in a noisy reverberant environment.

The first subject was presented in Chapter 2. In that chapter, a single-microphone speech enhancement method was proposed which uses inverse filtering and spectral subtraction. The inverse filtering can successfully deal with the problem of early reverberation and the spectral subtraction reduces the effects of additive background noise and late reverberation. To the best of our knowledge, this is the best single microphone technique for suppression of the effects of **both early and late reverberation** in **noisy speech** even when the **reverberation is high**. However, the proposed two-stage method can not be used in time-varying environment. This is because, in the first stage, we utilize an adaptive technique which requires convergence after at least 300 iterations. Moreover, this method needs enough input signal (20 sec for high reverberation time) which decreases the computational efficiency and its usage for real-time applications. On the other hand, the second stage of our method is based on spectral processing which are done in frequency domain using the FFT transform. This method can be used in real-time applications and can be applicable for time-varying environments.

The second subject was presented in Chapter 3. In that chapter, two new methods were presented to estimate the TDOA in a reverberant room. TDOA-based methods usually achieve the most accurate source localization in passive systems, and accurate TDOA estimation is the key to the effectiveness of these methods. In this dissertation, we proposed two different techniques to deal with the problem of TDE in noisy reverberant conditions. At first, we proposed a method based on adaptive inverse

filtering in Section 3.1.1. This method has some advantages and some disadvantages. For the advantages, it is very accurate and the most reliable method in reverberant conditions. It also can be used when the noise level is not very high (SNR>10 dB). For the disadvantages, it has still high computational complexity and it only works for the asymmetric input signals. It can not work in fast time-varying environment and in real-time processing. Secondly, we proposed two efficient preprocessing for the CC-based TDE methods in Section 3.1.2. This method is not as accurate as the previous TDE method but it is not sensitive to the type of input signals. It also has very low computational complexity, making it applicable for time-varying environment and real-time processing.

Although the proposed methods for both single-microphone speech enhancement and speaker localization have some advantages over the other conventional methods, there are still some issues that need to be further resolved. These issues include the computational complexity, to be more robustness to additive background noise and reverberation, tracking a moving source, and the number of sensors. Toward this end, some main research directions are planned for further work, and these are presented below.

4.1 Future Research on Speech Enhancement

As we discussed above, our proposed method has some benefits over the other methods but it still has some limitations. To overcome some of these limitations, we bring some new ideas for our future work.

1. One of the limitation of our adaptive inverse filter method is its high computational complexity due to its adaptively convergence with high enough input data. One idea is use the Compressed Sensing (CS) and sparse signal processing. CS can be used to reconstruct sparse vector from less number of measurements, provided the signal can be represented in sparse domain. Sparse domain is a domain in which only a few measurements have non-zero values. CS can also be applied to speech signals. The application of structured sparsity for joint speech localization-separation in reverberant acoustics has been investigated for multiparty speech recognition. Now we can use CS to reduce the length of inverse filter and input speech data, making the adaptive inverse filtering algorithm converge faster and more efficient. Then the estimated inverse filter

can be used to enhance the reverberant speech signal. This would reduce the computational complexity and makes the method more efficient.

2. In reverberant environments, a moving speaker yields a dynamically changing source-microphone geometry giving rise to a spatially-varying RIR. It is therefore desirable to reduce the effect of reverberation in such a time-varying environment. Our idea is the use of a model-based approach such as the one proposed in [65] in which the sound source is modeled by a block-based time-varying all-pole filter, and the channel by a linear time-varying all-pole filter. In this case, the parameters of all-pole filter can be estimated for each block of speech segment (e.g. every 128 ms) so that the method can be applicable in time-varying environment and real-time applications. This method can be combined with our efficient second-stage method to reduce the effects of additive noise and the remaining reverberation.

3. In the second-stage of our method, we use the spectral subtraction method in which the amplitude of the speech spectrum is only used to drive the spectral weight filter. However, it is shown in [66] that knowledge of the clean speech spectral phase can be employed for a more robust amplitude estimation. In addition, instead of using the original phase of the corrupted signal, the phase can be modified along with the amplitude to further improve the performance in noisy reverberant conditions. To do this, one should estimate the phase of the clean speech signal as this is partially done in [67] for speech separation application. The idea is to improve our second stage method by incorporating the phase information for estimating the spectral weight filter to enhance the speech spectra, and utilize the modified phase instead of the original, and corrupted phase.

4. In this dissertation, we propose two different methods that are combined to deal with the problem of noise and reverberation. One may use each of these algorithms combined with other methods. For example, our second stage can be combined with the method based on modifying the LP residual signal proposed in [16]. This can be very effective as our method can deal with problem of noise and late reverberation and the LP residual based method can deal with the remaining reverberation effects, in particular due to early reverberation.

4.2 Future Research on Speaker Localization

In this section, we bring four new ideas for speaker localization in reverberant environment. The first one is based on the TDE that can work under noisy reverberant conditions. The other three methods are based on speaker-microphone distance detection. Knowledge of the actual distance between the source and receiver can be advantageous not only for source localization but also in other audio and speech applications such as denoising, dereverberation, and separation. To date, most source localization methods use the received signal from a microphone array to obtain the location of the speaker. However, these methods have high computational complexity because of the number of microphones. Further, in some applications few microphones are available to localize the position of the source. Thus it is advantageous to determine the position or at least the speaker microphone distance using as few microphones as possible. This is possible if channel identification or estimation of the time of arrival of the direct path is employed. In the literature, the RIR has been estimated blindly using multi-channel processing, but estimation errors are a problem.

1. Our idea is to eliminate the source spectrum in the logarithm frequency domain which is generally called the cepstral domain. The key idea is subtracting the logarithm of the signal spectrum for the received signals and thus eliminate the effect of the source spectrum and then extract the TDOA from the phase information. This leads to a very fast and effective technique in reverberant conditions. We also suggest adding a noise reduction preprocessing stage before this in order to reduce the effect of additive noise in noisy reverberant conditions.

2. The problem of absolute distance estimation is becoming more difficult inside enclosed spaces where reverberation can be a significant component of the received signals. Hence, this problem is closely related to the estimation of the direct-to-reverberant Ratio (DRR) and this has been highlighted in several studies [68]-[70]. The DRR is typically extracted from a measured room impulse response (RIR), but in practice RIRs are not always available and its estimation is very difficult due to its non-minimum phase characteristic. We proposed an inverse filtering method in Section 2.1. In Section 3.1.1, we showed that the estimated inverse filter using our algorithm has the direct delay information as there is a clearly peak at a position related to the direct path. In addition, the

energy of the impulses next to the peak in the inverse filter is related to the early reverberation energy in the corresponding RIR. In most of the literature (e.g. [71]-[72]) some features which are related to the DRR are extracted from the received signal and then these features are used to train a classifier based on Gaussian Mixture Models (GMM) in order to detect the distance. Our idea is to use inverse filter estimation to classify the different choices of known distance and detect the correct distance based on this. This method would be more robust to reverberation and would also work under certain noisy conditions.

3. In Section 3.1.1, we showed how the inverse filter can be used to estimate the TDOA between two microphones. Thus the direct path information which is incorporated in the estimated inverse filter. Therefore we are able to estimate the direct path and also the microphone-speaker distance using only the inverse filter instead of the corresponding RIR. Our idea is to estimate the inverse filter of the RIR and calculate the inverse-filtered signal. Then the direct path information of the RIR can be estimated using the peak index of the inverse filter and the TDOA between the inverse-filtered signal and the reverberant signal. Therefore, we can directly estimate the direct path of the RIR using only single microphone information.

4. In Section 3.2.2, it was shown that the all-pass component of the RIR preserves the direct path information. In addition, the minimum-phase component of the RIR starts at time zero and does not depend on the direct path information. Therefore, these two components can be used to estimate the direct path of the RIR directly from the received speech data. Then the location of the speaker can be obtained using the speaker-microphone distances.

Chapter 5

Conclusions

This dissertation considered single-microphone speech enhancement and speaker localization in a noisy reverberant room. These difficult yet important problems. Some of the main obstacles are summarized below.

- Speech signals have colored and non-stationary characteristics with complex temporal dynamics. Existing techniques are known to be optimal, in particular for the source localization problem, when the input signal is stationary with a known spectrum. However, these techniques are of limited use with speech signals whose spectra can change dynamically and are non-stationary.

- The received speech signal in a room is inevitably contaminated by additive background noise. This noise makes the speech enhancement and source localization problems much more difficult, and requires modifications to existing techniques or the addition of preprocessing.

- In addition to additive noise, problems can occur due to reflections in an enclosed space. This reverberation can severely degrade the performance of speech enhancement and source localization techniques. The main difference between noise and reverberation is that the latter is dependent on the desired signal, whereas noise can be assumed to be independent of the signal.

- Another problem is related to the nature of the RIR. The RIR has non-minimum phase and thus its inverse filter is unstable. Moreover, the RIR is usually long and can change rapidly due to motion (it can be quite sensitive to the source position, temperature, locations of room furnishings, and movements in the

room). This can have a significant effect on speech enhancement and source localization techniques, and in general makes these problems much harder.

The solution to the problems of speech enhancement and source localization in noisy reverberant environments is thus very difficult. In this dissertation, solutions for these problems were presented and compared with existing techniques. For the speech enhancement problem, a two-stage method was proposed which uses inverse filtering to reduce early reverberation, and spectral subtraction to reduce late reverberation and additive noise. It is known that early and late reverberation affect speech signals differently [10]. The effects of reverberation can be considered in a two-dimensional perceptual space. The two components are colouration caused by early reverberation and echoes caused by late reverberation. The echoes smear the speech spectra and reduce the intelligibility and quality of the speech signals. Coloration results from the non-flat frequency response of the early reflections. It was shown that the proposed method can deal with both early and late reverberation in different noisy reverberant environments, and is superior to existing methods. To the best of our knowledge, this is the best technique for suppressing both early and late reverberation in noisy speech using only one microphone.

For the speaker localization problem, two techniques were proposed to estimate the TDOA. Generally, TDOA-based approaches are the most effective. The key issue with these techniques is the accuracy of the TDE method. A TDE method was presented which can accurately estimate the TDOA between two spatially separated microphones even in highly reverberant rooms. The main drawback of this method is the computational complexity. Thus a GCC-based TDE method was proposed which is robust to reverberation and has lower computational complexity. It can also be used in real-time applications and can be employed to track a speaker moving in a room.

Bibliography

[1] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, *Statistical recovery of wideband speech from narrowband speech.* IEEE Trans. Audio, Speech, Lang. Process., vol. 2, no. 4, pp. 544–548, Oct. 1994.

[2] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition, Signal Processing*, A. Oppenheim, Series Ed., Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, FL: CRC Press, 2007.

[5] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, pp. 977–980, Apr. 1991.

[6] K. Kumar and R. M. Stern, "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation," in *Proc IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 4282–4285, March 2010.

[7] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," in *Proc IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 840–843, Apr. 2003.

[8] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[9] K. Lebart and J. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust.*, vol. 87, pp. 359-366, 2001.

[10] E. A. P. Habets, Single- and multi-microphone speech dereverberation using spectral enhancement, Ph.D. Thesis, Eindhoven Univ. of Tech., The Netherlands, 2007.

[11] J. S. Erkelens and R. Heusdens, "Single-microphone late-reverberation suppression in noisy speech by exploiting long-term correlation in the DFT domain," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 3997–4000, Apr. 2009.

[12] E. A. P. Habets, S. Gannot, and I. Cohen, "Speech dereverberation using backward estimation of the late reverberant spectral variance," in *Proc. IEEE Conf. Electrical and Electronics Engineers in Israel*, pp. 384–388, Dec. 2008.

[13] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 817-820, May 2006.

[14] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Advances in Signal Process.*, vol. 2009, Article ID 437807, 9 p., 2009.

[15] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.,* vol. 16, no. 9, pp. 770–773, Sept. 2009.

[16] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech and Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[17] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 17, no. 2, pp. 137–148, Feb. 2009.

[18] H. Padaki, K. Nathwani and R. M. Hegde, "Single channel speech dereverberation using the LP residual cepstrum," in *Proc. National Conf. on Commun.*, pp. 1–5, Feb. 2013.

[19] M. Wu and D. L. Wang, "A two-stage algorithm for one microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, no. 3, pp. 774–784, May 2006.

[20] E. A. P. Habets, N. Gaubitch, and P. A. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, pp. 4577–4580, Apr. 2008.

[21] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 758-764, 2000.

[22] J. S. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Speech and Audio Process.*, vol. 18, no. 7, pp. 1746–1765, Sept. 2010.

[23] H. R. Abutalebi, and B. Dashtbozorg, "Speech dereverberation in noisy environments using an adaptive minimum mean square error estimator," *IET Signal Process.*, vol. 5, no. 2, pp. 130–137, 2011.

[24] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[25] S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 322–335 , July 2013.

[26] S. Mosayyebpour, T. A. Gulliver, and M. Esmaeili, "Single-microphone speech enhancement by skewness maximization and spectral subtraction," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–4, Sep. 2012.

[27] S. Mosayyebpour, A. Sayyadiyan, M. Zareian, and A. Shahbazi, "Single channel inverse filtering of room impulse response by maximizing skewness of LP residual," *Proc. IEEE Int. Conf. on Signal Acquisition and Process.*, pp. 130–134, Feb. 2010.

[28] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based approach for inverse filtering of room

impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, July 2012.

[29] S. Mosayyebpour, A. Keshavarz, M. Biguesh, A. Gulliver, and M. Esmaeili "Speech-model based accurate blind reverberation time estimation using an LPC filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1884–1893, Aug. 2012.

[30] S. Mosayyebpour, H. Lohrasbipeydeh, M. Esmaeili, and T. A. Gulliver, "Time delay estimation via minimum-phase and all-pass component processing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, Vancouver, BC, pp. 4285–4289, May 2013.

[31] S. Mosayyebpour, A. Sayyadiyan, E. Soltan Mohammadi, A. Shahbazi, and A. Keshavarz, "Time delay estimation using one microphone inverse filtering in highly reverberant room," in *Proc. IEEE Int. Conf. on Signal Acquisition and Process.*, Bangalore, India, pp. 140–144, Feb. 2010.

[32] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.,* vol. 10, no. 7, pp. 204–207, July 2003.

[33] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 5, pp. 504–512, July 2001.

[34] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Process.*, vol. 86, no. 6, pp. 1215–1229, June 2006.

[35] D. Mauler and R. Martin, "Noise power spectral density estimation on highly correlated data," in *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pp. 1–4, Sep. 2006.

[36] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, pp. 208-211, Apr. 1979.

[37] I. Cohen, "Speech Enhancement Using a Noncausal A Priori SNR Estimator," *IEEE Signal Process. Lett.,* vol. 11, no. 9, pp. 725–728, Sept. 2004.

[38] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[39] S. Kamath, and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, p. IV-4164, May 2002.

[40] K. Furuya, and A. Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1579–1591, July 2007.

[41] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *Proc. Int. Conf. Digital Signal Process.*, pp. 1–5, July 2009.

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

[43] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[44] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *Proc. IEEE Int. Conf. on Robotics and Automation,* pp. 1033–1038, Apr.-May 2004.

[45] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[46] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.* vol. 34, no. 3, pp. 276–280, May 1986.

[47] M. McCloud and L. Scharf, "A new subspace identification algorithm for high resolution DOA estimation," *IEEE Trans. Antennas Propag.*, vol. 50, no. 10, pp. 1382–1390, Oct, 2002.

[48] T. G. Dvorkind and S. Gannot "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.,* vol. 85, no. 1, pp. 177–204, 2005.

[49] M. Brandstein, J. E. Adcock, and H. Silverman, "A practical time delay estimator for localizing speech sources with a microphone array," *Computer Speech and Language,* vol. 9, no. 2, pp. 153–169, 1995.

[50] G. C. Carter, *Coherence and Time Delay Estimation: An Applied Tutorial for Research, Development, Test, and Evaluation Engineers*, IEEE Press, Piscataway, NJ, 1993.

[51] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. 24, no. 4, pp. 320-332, Aug. 1976.

[52] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE,* vol. 61, no. 10, pp. 1497–1498, Oct. 1973.

[53] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Process.,* vol. 5, no. 3, pp. 288–292, May 1997.

[54] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for time delay estimation under reverberant conditions," *Signal Process.*, vol. 59, pp. 253–266, 1997.

[55] M. S. Brandstein, "A pitch-based approach to time-delay estimation of reverberant speech," in *Proc. IEEE ASSP Workshop on Appls. of Signal Process. to Audio Acoustics*, 4 p., Oct. 1997.

[56] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, pp. 384–391, Jan. 2000.

[57] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[58] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband

independent component analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1490–1503, Aug. 2011.

[59] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 246–260, Jan. 2012.

[60] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 157–160, Mar. 2007.

[61] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1975.

[62] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, July 2003.

[63] Mark Veillette, Alpha-Stable Distributions in MATLAB, Jul 2012 http://www.mathworks.com/matlabcentral/fileexchange/37514-stbl-alpha-stable-distributions-for-matlab.

[64] R. J. Kozick and B. M. Sadler, "Source Localization with Distributed Sensor Arrays and Partial Spatial Coherence," *IEEE Trans. Sig. Proc.,* vol. 52, no. 3, pp. 601–616, March 2004.

[65] J. R. Hopgood and C. Evers, "Block-based TVAR models for single-channel blind dereverberation of speech from a moving speaker," *in IEEE Workshop on Statistical Signal Processing,* pp. 274–278, 2007.

[66] T. Gerkmann and M. Krawczyk, "Mmse-optimal spectral amplitude estimation given the stft-phase," *Signal Processing Letters, IEEE,* vol. 20, no. 2, pp. 129–132, Feb. 2013.

[67] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," *Proceedings of the International Conference on Spoken Language Processing,* 2012.

[68] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Speech Audio Process.,* vol. 18, no. 7, pp. 1793–1805, Sep. 2010.

[69] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Trans. Speech Audio Process.,* vol. 19, no. 8, pp. 2374–2384, Nov. 2011.

[70] M. Kuster, "Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity," *J. Acoust. Soc. Amer.,* vol. 130, no. 6, pp. 3781–3787, 2011.

[71] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos , "Speaker distance detection using a single microphone, " *IEEE Trans. Speech Audio Process.,* vol. 19, no. 7, pp. 1949 –1961 , 2011.

[72] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos , "Sound Source Distance Estimation in Rooms based on Statistical Properties of Binaural Signals " *IEEE Trans. Speech Audio Process.,* vol. 21, no. 8, pp. 1727 –1741 , 2013.