TWO STAGE BLIND DEREVERBERATION BASED ON STOCHASTIC MODELS OF SPEECH AND REVERBERATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MEHMET KAVRUK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JANUARY 2019

Approval of the thesis:

TWO STAGE BLIND DEREVERBERATION BASED ON STOCHASTIC MODELS OF SPEECH AND REVERBERATION

submitted by **MEHMET KAVRUK** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Tolga Çiloğlu
Head of Department, **Electrical and Electronics Eng.**

Prof. Dr. Tolga Çiloğlu
Supervisor, **Electrical and Electronics Eng., METU**

Examining Committee Members:

Prof. Dr. Çağatay Candan
Electrical and Electronics Eng., METU

Prof. Dr. Tolga Çiloğlu
Electrical and Electronics Eng., METU

Assoc. Prof. Dr. Fatih Kamışlı
Electrical and Electronics Eng., METU

Dr. Sevinç Figen Öktem
Electrical and Electronics Eng., METU

Prof. Dr. Özgül Salor Durna
Electrical and Electronics Eng., Gazi University

Date: 15.01.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Mehmet Kavruk

Signature:

ABSTRACT

TWO STAGE BLIND DEREVERBERATION BASED ON STOCHASTIC MODELS OF SPEECH AND REVERBERATION

Kavruk, Mehmet
Master of Science, Electrical and Electronics Engineering
Supervisor: Prof. Dr. Tolga Çiloğlu

January 2019, 128 pages

Distant speech processing is popular nowadays due to wide use of the hands-free communication with smart devices. The quality of microphone signals in an enclosed area is degraded by environmental noise and reverberation in distant speech communication. Although there are powerful denoising algorithms in the literature, there is no robust dereverberation method which works independent of recording conditions. This work proposes a statistical model based blind dereverberation algorithm which suppresses reverberation part without causing serious degradation in the source signal in different speaker to microphone configurations. The proposed algorithm successively uses minimum variance distortionless response (MVDR) and linear prediction methods. The parameters of the MVDR algorithm are estimated using the statistical nature of reverberation. The linear prediction algorithm is applied to the output of MVDR in order to handle residual reverberation. The dereverberation filter in this stage is generated using the statistical models of speech and reverberation. None of the algorithms require any deterministic prior knowledge about the system due to the used statistical models. The experimental results demonstrate that the proposed algorithm suppresses reverberation in the distant recordings without degradation on the source signal with respect to the objective quality measures under different conditions.

Keywords: Dereverberation, MVDR, Linear Prediction

ÖZ

SES VE REVERBERASYONUN STOKASTIK MODELLERINE DAYALI ÇİFT KANALLI KÖR DEREVERBERASYON

Kavruk, Mehmet
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği
Tez Danışmanı: Prof. Dr. Tolga Çiloğlu

Ocak 2019, 128 sayfa

Akıllı cihazlarda uzaktan komut teknolojisinin gelişimi sayesinde uzak ses işlemesi günümüzde popüler bir hale gelmiştir. Uzaktan komut sisteminde, bir odada kaydedilen sesin kalitesi çevresel gürültü ve reverberasyon yüzünden düşer. Literaturde çevresel gürültüleri bastırmak için belirli algoritmalar olmasına rağmen reverberasyonu bastırmak için kullanılan genel bir algoritma yoktur. Bu çalışma istatistiksel modellere dayanarak değişik konfigürasyonlarda asıl konuşmada bir bozulma oluşturmayan ve aynı zamanda sesin reverberasyonunu bastırabilen bir yöntem sunmaktadır. Sunulan yöntem MVDR ve doğrusal tahmin algoritmalarını art arda kullanmaktadır. MVDR algoritmasının parametreleri reverberasyonun istatistiksel modeli göz önüne alınarak bulunmuştur. MVDR algorithmasının çıkışı doğrusal tahmin algorithmasını besler ve MVDR algoritmasında bastırılamayan reverberasyonlar burada bastırılmaya çalışılır. Doğrusal tahmin algoritmasında kullanılan filtre sesin ve reverberasyonun istatistiksel modellerine göre bulunur. Bu istatiksel modeller sayesinde sistemde hiçbir ön bilgiye gerek duyulmaz. Testlerin sonucunda objektif sonuçlara göre, sunulan sistem değişik konfigurasyonlarda asıl ses sinyaline önemli bir zarar vermeden reverberasyonu bastırabilmektedir

Anahtar Kelimeler: MVDR, Doğrusal Tahmin, Reverberasyon

To my dearest family

ACKNOWLEDGMENTS

First of all, I would like to thank my thesis advisor Prof. Dr. Tolga Çiloğlu for his help, comments, remarks and guidance. None of this would have been possible without him.

I am also grateful to my family for their endless support and encouragement throughout my life and during the writing process of this thesis.

I would like to thank all my friends for their understanding and support during this thesis.

Finally, I would like to acknowledge the support of ASELSAN Inc. during this thesis.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1. Motivation

The substantial rise in the use of smart portable devices results in an improvement in hands-free speech communication interfaces. Hands-free communication with the device is used in various applications such as personal assistants, speech to text conversions and speaker identifications. In addition, spread of the internet develops multimedia applications. It is very common to hold conferences by sending videos and voice via internet. This thesis is motivated by a growing demand for hands-free speech interfaces with smart devices and other applications where quality of speech signals is important.

Speech is acquired by microphones in the hands-free applications. Advent of bluetooth technology provides a low cost, high quality, wireless headset. In this way, the microphone can be held close to mouth during communication. Since these headsets bring some restrictions for the speaker and are not feasible for multiple speaker situations such as teleconferences, wireless headsets are not commonly preferred. The main purpose of hands-free speech interface is to provide a natural way of communication for users and comfort to user movements, so the microphones cannot always be located near a speaker. In fact, there is generally a distance between speaker and microphone in the applications. Therefore, a speech processing algorithm which works independent of speaker-microphone configuration is required.

In hands-free applications, the distance between a speaker and microphone is generally between 0.5 m - 2 m. In this scenario, the recordings not only contain the desired speech but also other interferences. The main detrimental interference is the background noise. If the sound is produced in an enclosed area, reverberation effect is also inevitable in distant recordings. These two distinct components jointly contribute to an overall degradation in speech signals, which reduces the quality of the perceived speech. The resulting lower quality of recordings reduces the performance of hands-free speech interface.

Reverberation originates when a sound is produced in an enclosed area. It is caused by reflection of sound waves from walls and surrounding objects, as it can be seen in Figure 1.1. In fact, carefully generated controlled reverberation strengthens the intelligibility of observed speech signal [1]. However, speaker localization and intelligibility of sound are seriously degraded in uncontrolled severe reverberant conditions because speech phonemes are blurred, and their characteristics change with reverberation [2]. The effect of reverberation is clearly seen in spectrograms. Although, the formants and phonemes of anechoic speech are well separated in time, reverberation causes a smearing effect, and the subsequent phonemes are overlapped. These effects are shown in Figure 1.2. In Figure 1.2.a, all characteristics of anechoic speech phonemes are easily seen; however, they are blurred in reverberant environments as in Figure 1.2.b.

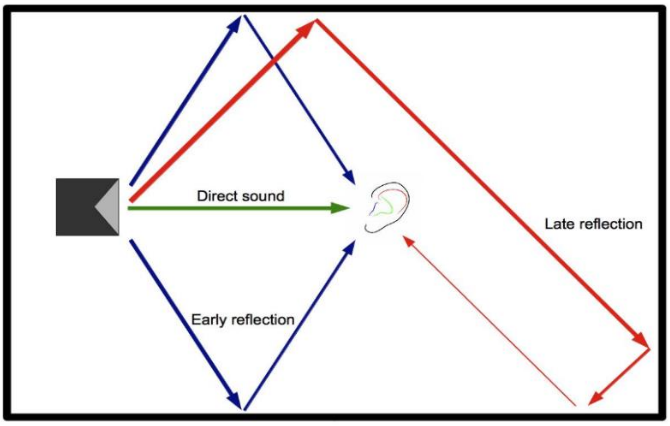


Figure 1.1. Reverberation arises due to reflections of sound waves in an enclosed region.

Background noise arises due to the environmental conditions such as traffic and other audible talkers. When the level of noise is comparable with the desired sound, the perceptual quality of the observed speech is reduced significantly. Noise in the observed speech directly affects the performance of automatic speech recognition, speaker localization and identification [3].

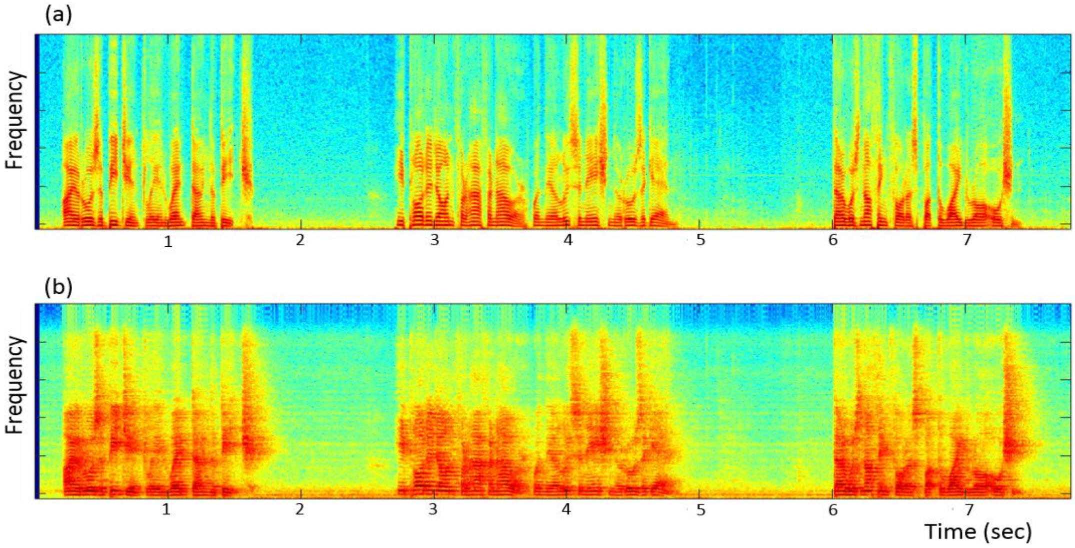


Figure 1.2. The spectrogram of (a) anechoic speech and (b) reverberant speech.

Reverberation and background noise both impair speech intelligibility, so they are the main problems of hands-free speech interfaces. Many signal processing techniques exist in order to fix the problems caused by background noise. Noise reduction methods have been extensively investigated, and lots of significant contributions and robust solutions have been already offered [4]. However, reverberation problem has drawn much less attention, and a robust solution to this problem has not been developed yet. In order to use hands-free speech interface effectively in daily life, the reverberation problem must be solved. This work offers to use an integrated algorithm which consists of minimum variance distortionless response (MVDR) and linear prediction to deal with reverberation problem. If the direction of the speaker is estimated, the directional filter of MVDR can strengthen the desired signal while it suppresses reverberation. However, reverberation reaches the microphone from all directions, and the reverberant parts which are at the direction of the speaker can not be handled in this way. Therefore, this work proposes to use a single channel algorithm at the output of the MVDR in order to handle residual reverberation. In addition, the parameters of the single channel algorithm are estimated more accurately than a single microphone case by the microphone array thanks to this integrated algorithm. As a result, the main objective of this thesis is to benefit from two-stage algorithm which consists of microphone array and linear prediction. Feasibility and advantages of the proposed solution are evaluated in this work. In addition, the weaknesses of the algorithm due to non-ideal models of speech and reverberation are explained, and possible solutions to these problems are explained.

In reverberant environments, it is difficult to estimate the direction of the speaker, so the performance of MVDR reduces drastically. In this work, a method is developed to reduce reverberation effects and the speaker direction is estimated more accurately.

The essential difference between background noise and reverberation is that reverberation depends on the desired signal; thus, its characteristics change depending on the speaker and environment, whereas background noise is independent of the

desired speech. In fact, reverberation is roughly nothing but a delayed copy of the desired speech signal due to reflections. Therefore, the contributions proposed to deal with background noise cannot produce a solution to reverberation problem [2].
The algorithms in this work are adjusted to the nature of reverberation.

1.2. Overview of Literature

Signal processing methods which deal with the reverberation problem are called dereverberation. Although there is still no robust method for dereverberation, some contributions and solutions are proposed in the literature. A detailed overview of dereverberation approaches is given in Chapter 2. However, basics of the proposed dereverberation methods are given shortly in this section to identify the problem. Although there are different techniques for dereverberation, they can be divided into two main classes: reverberation cancellation and reverberation suppression [5].

Reverberation cancellation is based on estimating the inverse of acoustic impulse response. In order to remove the effects of reverberation, the inverse filter must be estimated precisely. By using the inverse of acoustic impulse response, dereverberation is easily achieved by convolving the filter and observed speech signal. However, this approach always introduces some artifacts in the process because acoustic impulse response is estimated blindly, and it changes dramatically with environmental effects. If the inverse of impulse response cannot be found exactly, significant degradations in the desired signal and additional noise are inevitable.

Although reverberation cancellation can provide complete dereverberation potentially, reverberation suppression can obtain better results. There is no estimate of impulse response in this class. Instead of estimating acoustic impulse response of a room, reverberation part is isolated and treated as a noise when it is suppressed. Isolation of reverberant part is a difficult task due to the non-deterministic nature of reverberation. However, this is a more achievable option than estimating the exact impulse response.

Both of the classes require knowledge of lots of parameters; therefore, dereverberation is still an unsolved problem in the speech processing literature. Although the methods in the first class can provide dereverberation completely in theory, the work in this thesis belongs to the second class. The aim of this work is to propose a solution which can be used in practical applications independent of speaker to microphone configuration.

In this work, the non-deterministic nature of reverberation is considered, and the algorithms are adjusted with respect to this nature. However, the nature of reverberation is simplified due to the complexity of required computations. The proposed system combines two different approaches in order to achieve better dereverberation. In the first stage, a microphone array structure is used to isolate the desired speech from reverberation, and the reverberant part is filtered. Microphone array can not suppress the reverberation signals that reach the microphone from the direction of the desired signal. Therefore, in the second stage, a single channel dereverberation method based on linear prediction is applied to the output of the microphone array in order to suppress the residual reverberation. This linear prediction algorithm is based on the parameters used in the first stage. In this way, a compact dereverberation solution is obtained.

1.3. Thesis Contributions

In this thesis, the contributions can be gathered as follows,

- Beamformer and spectral subtraction method in denoising algorithms are used in dereverberation with the probabilistic models of speech and reverberation [6].
- Although there are multiple arrival directions to microphone in a reverberant environment, the arrival direction of the desired sound is estimated by a novel approach. First, the speech onsets without reverberation are detected by voice activity detection algorithm in the microphone array observations [7]. Since

there is no reverberation in these onsets, correlation between the onsets gives the phase differences of the direct parts in the recordings. Direction of arrival of the direct part is estimated by using these phase differences.

- Weighted prediction error (WPE) in the second stage is originally based on iterations to find power spectral density of early speech components [8]. However, since the spectral density of reverberation is estimated by microphone array in the first stage the method reaches the solution in a single step. The advantages of two stage dereverberation algorithm are also explained in this work.
- Reverberation suppression is achieved blindly without causing serious degradation in the desired speech with respect to the objective measures (‘*PESQ*’, ‘*LLR*’, ‘*CD*’) [9].

1.4. Thesis Organization

This thesis focuses on dereverberation by the probabilistic models of speech and reverberation, and they are explained in the following chapters in detail. Also, a detailed literature survey and implementation details are included. The thesis is structured as follows,

Chapter 2 explains physics of sound waves and causes of reverberation in enclosed areas. Non-deterministic nature of reverberation is illustrated. The probabilistic models of reverberation which are generated according to these explanations are shown. The theory of microphone array and a wide literature survey of dereverberation are also included in this chapter.

Chapter 3 presents all the theory of the proposed methods successively. The chapter consists mainly of two different parts. Each stage of the algorithm is explained in these parts. The mathematical derivations of each stage are given.

Chapter 4 presents implementation details. The speech quality measures are explained. The results under different conditions are given in terms of these measures. Discussions of the results are also included in this chapter.

Chapter 5 summarizes implementation and results of the algorithm, and presents the future work.

CHAPTER 2

BACKGROUND

2.1. Introduction

In this chapter, mathematical modeling of dereverberation and existing solutions of dereverberation problem are reviewed. In addition, the fundamentals of statistical room modeling and statistical nature of sound waves are presented as a background of the proposed solution in this thesis. ‘Beamforming’ method is also included at the end of the chapter since it takes a significant role in the proposed dereverberation algorithm.

2.2. Mathematical Modeling of Reverberation

In order to deal with reverberation, reverberant sound must be analyzed mathematically according to its characteristics. For this purpose, it is necessary to study room impulse response (RIR) since it represents room acoustics. RIR is the filter between source and listener in a room. Finite impulse response or infinite impulse response structures are used to describe RIR. Figure 2.1 shows the components of a typical RIR.

The characteristics of a RIR can explain reverberation in an enclosed area. There are three different components of RIR that generate different parts of reverberant speech [10].

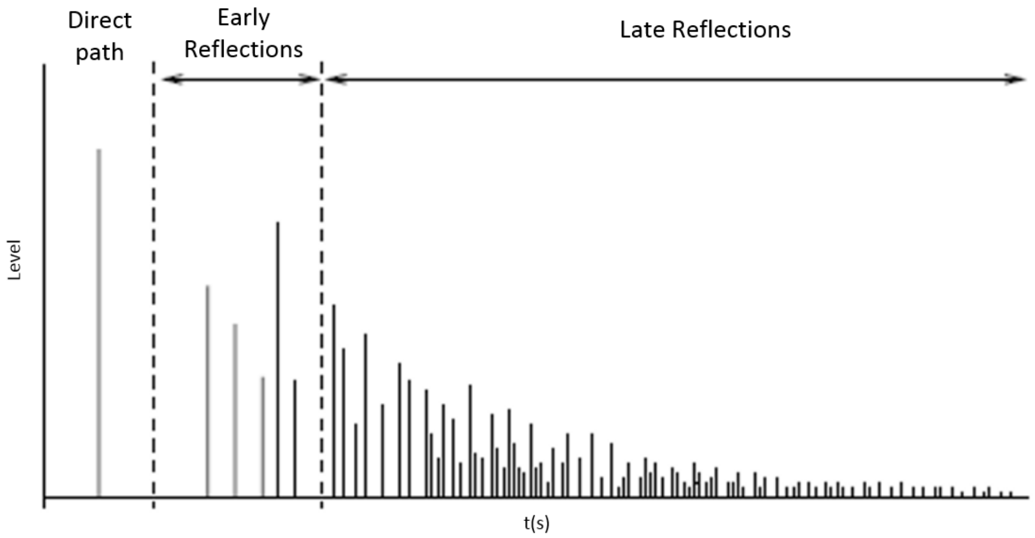


Figure 2.1. The structure of a typical RIR in an enclosed area.

Direct Path: The single non-zero amplitude shows only direct sound without reflections in this interval. The initial dead time in this component refers to propagation delay of direct sound between source and microphone. The magnitude of the direct part relative to other impulses is related to source-microphone distance and reflectivity of room.

Early Reflections: The reflections arriving approximately within the first 30 msec are called early reflections. Number of the impulses in this interval is low, and their magnitudes are large relative to the subsequent impulses. In general, these closely spaced reflections cannot be distinguished from direct sounds by human ears. Therefore, they reinforce direct response and they are considered useful for speech intelligibility.

Late Reflections: Late reflections follow early reflections. They generally reach the microphone after multiple reflections. They are randomly spaced, decaying impulses. This part is the major contribution to the notorious reverberation effects and destructive for speech intelligibility. Therefore, dereverberation methods aim at suppressing late reflections in the observed sound since they are the main cause of degradation of speech intelligibility.

During acoustic propagation in a room, each reflection absorbs some energy of sound waves. Therefore, the intensity of the reaching sound to microphones decreases with time. This situation can be observed from RIR pattern. Typical RIR has a tail structure which tends to zero by time; therefore, it is sufficient to take the first L values of RIR into consideration. The value of L depends on the acoustic features of room, and it is a significant parameter of a particular reverberation pattern.

Reverberant signal can be considered as result of convolution of a source signal and a causal RIR. Observed speech signal at discrete time n at m^{th} microphone can be written in time domain as:

$$x_m[n] = \boldsymbol{h}_m^H[n]\boldsymbol{s}[n] + v_m[n] \qquad (2.\ 1)$$

where $\boldsymbol{h}_m[n] = \left[h_{\mathrm{m},1}\ h_{\mathrm{m},2}\ h_{\mathrm{m},3} \ldots..\ h_{m,L}\right]^T$ is the vector of RIR of length L,

$\boldsymbol{s}[n] = [s[n]\ s[n-1]\ s[n-2] \ldots..\ s[n-L+1]]^T$ is the vector of anechoic source signal, $v_m[n]$ is noise signal at m^{th} microphone. Mathematically, the aim of speech dereverberation is to find the best estimate of $s[n]$. This is a very complicated problem since acoustics impulse response is not known.

RIR structure heavily depends on room acoustics. Therefore, studying room acoustics gives a better understanding with regards to artificial RIR modeling and dereverberation approaches.

2.3. Overview of Room Acoustics

Main contributors of an RIR are the room acoustic properties and wave motion characteristics. Room acoustic properties affect propagation and interaction of sound waves. The geometry of sound propagation, the rules of reflection and absorption make it possible to predict each sound wave location in a room. In this way, RIR can be found theoretically. However, it is evident that the procedure is very tedious and uncertain. Therefore, modeling room acoustics with the simplifying assumptions eases the situation. Before studying room acoustics, mathematical modeling of sound waves is helpful.

2.3.1. Sound Wave Propagation

Sound wave propagation through the air is described by Helmholtz equation [8]:

$$\nabla^2 p(\boldsymbol{q},t) - \frac{1}{c^2}\frac{\partial^2 p(\boldsymbol{q},t)}{\partial t^2} = 0 \tag{2. 2}$$

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2. 3}$$

where p is sound pressure level at related position in dB, $\boldsymbol{q}$ is position vector, t and c are time in sec and sound wave velocity in m/sec, respectively. The wave equation involves sound pressure level in acoustic field. By Eq. (2.2) the sound pressure level can be described as position and time dependent function.

For simplicity, assume that sound waves propagate only along x direction, and then general form of the solution is:

$$p = p_0 e^{(jwt-kx)}. \tag{2. 4}$$

Eq. (2.4) shows that sound waves propagate along a direction with the same amplitude but changing phase with respect to position and time.

Another simple wave model is the spherical wave whose curved wavefronts represent constant sound pressure levels. A spherical wave can be considered as emerging from a point source. The polar coordinate ‘r’ is used to describe the source distance. Helmholtz equation in spherical coordinates [11]:

$$\frac{\partial^2 p(r,t)}{\partial r^2} + \frac{2}{r}\frac{\partial p(r,t)}{\partial r} - \frac{1}{c^2}\frac{\partial^2 p(r,t)}{\partial t^2} = 0. \tag{2. 5}$$

Now, sound pressure value is a function of the source distance and time. General form of the solution is:

$$p(r,t) = \frac{P_0}{r} e^{(jwt-kr)} \ . \tag{2. 6}$$

As a result, there are two basic models for sound wave propagation which are spherical and plane wave models as shown in Figure 2.2. In the plane wave model, sound pressure does not depend on source distance and the waves propagate in one direction with a flat shape wavefront. However, plane wave is an idealized wave model that does not exist in real life. Spherical waves are different than plane waves as they are propagating in all directions simultaneously. Pressure magnitude is proportional to inverse of the source distance.

Spherical waves can represent sound waves more ideally. However, when a listener is far enough, the curved wavefront structure and amplitude decay of the wave between two points can be neglected, and sounds waves converge to a plane wave [12]. This approximation is valid for the source distance greater than:

$$r = \frac{d^2}{\lambda} \tag{2. 7}$$

where d is the distance between microphones in an array, λ is the wavelength of related sound. The units of both variables are in meters.

Speech signals are dominated by 600 Hz to 1000 Hz frequency content. This means that wavelengths of speech signals are mostly from 20 to 50 cm. Therefore, far field approximation generally becomes valid in distant speech processing applications (low frequency parts of speech may not support the model).

Near field and far field terms in speech processing are related to wave model approximation. Acoustic far field is defined as from the distance in Eq. (2.7) to infinity. In this region, sound waves can be modelled by plane waves. Near field is the interval where curved structure of sound waves cannot be neglected. In dereverberation, generally far field assumption is used since reflected sounds take a distance much more than Eq. (2.7). Although it is a very simple approximation, it reduces computational complexity of calculations.

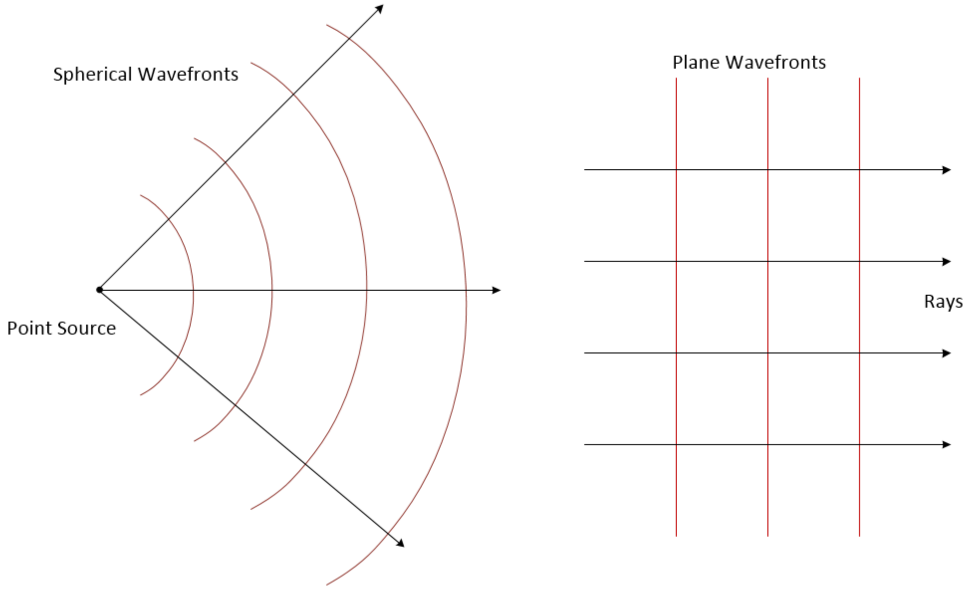


Figure 2.2. Representations of a spherical wave and a plane wave respectively.

2.3.2. Room Acoustics for Reverberation

In order to model entire room acoustics, one would study sound mapping at all different locations in a room. Instead of this exhausting work, finding a characteristic property of room that describes whole room acoustics is very tempting. For the sake of this purpose American physicist, Clement Sabine, discovered that reverberation was well suited for describing important aspects of room acoustics [13]. Sabine made lots of experiments with a sound source which was suddenly stopped. He found that there was persistence of the sound that could be heard after stopped and its intensity decayed with time.

Sabine measured auditability of the sound after the source was stopped. He discovered that the duration of auditability was same for all locations in a room. This is one of the most fundamental properties of reverberation. However, he found that duration of auditability was not a characteristic property of reverberant room, alone. At the next experiment, strength of sound was studied, and he found that a pistol shot reverberation was longer than a snap of fingers. Therefore, auditability of sound depends on both initial energy and reverberation time of room. Also, it was discovered that sound energy was half in equal time, independent of the initial energy. Sound always lost the same percentage of its energy at the same time interval. All of these experiments constituted the theoretical fundamentals of reverberation equations.

Mathematical representations of Sabine's work can be described as exponential decay function of sound energy density:

$$-\frac{dE}{E} = \frac{dt}{\tau} \tag{2. 8}$$

where τ is the characteristic time constant of room, E is energy density. Then integration of the exponential function gives:

$$lnE = lnE_0 - t/\tau \tag{2. 9}$$

where E_0 is initial energy. Function of energy with respect to time and initial value can be written explicitly:

$$E = E_0 e^{-\frac{t}{\tau}}. \tag{2. 10}$$

If the minimum energy value for auditability is determined, the time for auditability in a room can be found by Eq. (2.10). This equation is one of the characteristic acoustic features of room. Characteristic time constant τ makes this process distinctive for each enclosed region.

According to Sabine’s works, instead of working with place to place variation of sound signals, it is possible to form average aural characteristics of a room by reverberation. However, each reverberant sound wave at different locations in a room cannot be represented by just the same exponential energy equation. It would be too simplified representation of sound signals. Due to uncertainty of each sound wave in enclosed region, single sound wave has to be represented statistically. Therefore, theoretical basis of reverberant signals should be in statistical nature complemented with the reverberation equations.

2.3.2.1. Frequency Domain Reverberation Model

Room mode is collection of the resonance frequencies occurring when sound source is excited. Modal density of a room is related to the resonance numbers which are contained per frequency. Polack describes the modal density mathematically [14]

$$\frac{dN}{df} \approx \frac{4\pi V}{c^3} f^2 \tag{2. 11}$$

where N is the resonance number, V is the volume of room in m^3, f is the frequency of sound in Hz, c is sound velocity in m.

As it can be seen in Eq. (2.11), modal density increases with square of frequency. It is expected due to shorter wavelength of sound because resonance occurs when sound propagation distance is multiple of half of sound wavelength. In Figure 2.3, resonance frequencies can be seen as the peaks in the frequency response plot. They are well separated at low frequencies; however, when frequency of the sound increases, they get closer and modal density increases.

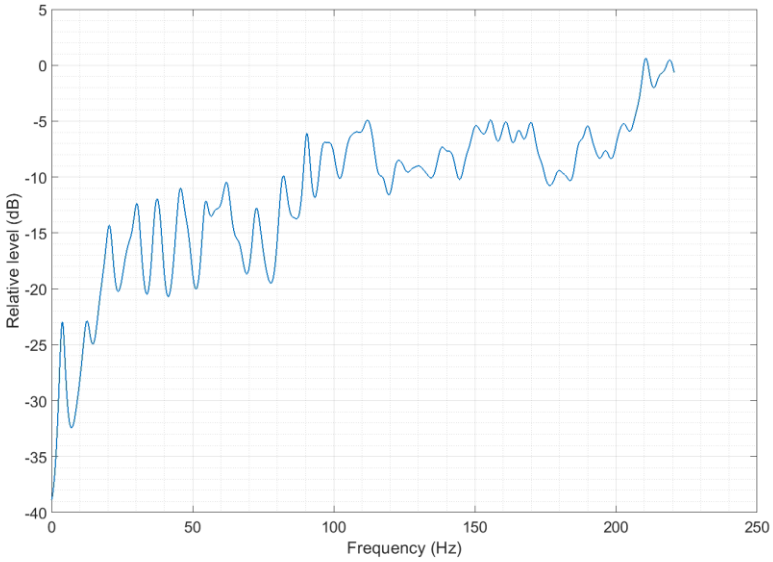


Figure 2.3. Resonance frequencies are the distinct peaks in frequency response plot.

Eq. (2.11) is the basis of room acoustical modeling of Schroeder [15]. Schroeder states that at high frequencies, resonance frequencies are too close to be distinct. Therefore, any resonance frequency affects several of neighbor frequencies. This means that, at high frequencies, single source signal excites several of room modes simultaneously. In a reverberant room, when a sine wave signal is excited, microphone captures sum of different phase and amplitude contributions of the room modes. Same phase and amplitude contributions never occur at different position and time in the room (assume that the room modes are independent). Contributions of the room modes can be explained statistically because of the uncertain nature of the process. In fact, since all room modes have equal probability to occur independently; according to central limit theory, summation of contributions converges to Gaussian shape. As a result, transfer function is described as position and frequency dependent Gaussian stochastic process. This statistical model is based on the assumption of Schroeder frequency that is defined as [15] :

$$f_{schroeder} \approx 2000\sqrt{\frac{T_r}{V}}. \qquad (2.\ 12)$$

If excited frequency is higher than Schroeder, room modes start to overlap. Therefore, frequency response of a room should be described statistically.

2.3.2.2. Time Domain Reverberation Model

Polack developed time domain model of the Schroeder's contributions [6]. In this model, RIR can be considered as a non-stationary stochastic process all over room. There are fast and slow variations in RIR. Fast variations are in a few msec time scales. These variations exist due to uncertain excitation nature of room modes, and they can be represented by Gaussian process if sufficient number of reflections occurs. Slow variations are in hundreds of msec scales. They exist due to sound energy absorption in room. The situation can be described as exponential decays. Polack defined RIR of reverberant part as a result of these assumptions:

$$h(t) = b(t)e^{-\beta t} \qquad (2.\ 13)$$

where $b(t)$ is Gaussian distribution stochastic process, while β is the decay parameter related to reverberation characteristics of room.

In time domain, there is a time interval for Polack model to become valid [16]. In order to define response as a stochastic process, a number of room mode overlapping must be existed across space and overlapped reflections must spread over room uniformly. The time interval for this process is named 'mixing time'. It is defined as:

$$t_{mixing} = \sqrt{V}\ (msec) \qquad (2.\ 14)$$

where V is the volume of room in m^3. Mixing time is the transition time from early reflections to late reflections in RIR. After several reflections, sound waves interact each other and spread all over room. RIR no longer corresponds to the arrivals of specific sound waves (It is named 'diffuse sound field'). General response starts to be described by a statistical process. There are two main conditions for Polack time domain reverberation model: mixing time and Schroeder frequency.

The main assumption of Polack reverberation model is diffuse sound field [17]. In the literature, diffuse field can be defined as uniform energy density across room and equal probability of energy flow in all directions with random phases. The proposed model accuracy depends on diffuseness of room. Structure of room and frequency content of sounds are main contributors of diffuseness.

Perfect diffuse sound field does not exist in real life; however, it is an acceptable assumption. All irregularities of a room help sound waves to distribute its energy in all possible direction. Roughness of room walls with respect to wavelengths scatters incident waves in wide range of directions. The ‘diffusely reflecting’ term expresses this situation, and it can be shown as in Figure 2.4. Practically, incident wave does not reflect only one direction as an ideal case. Furthermore, furniture in a room and irregular decorations all help sound waves to reflect diffusely. Even in partially diffuse room, diffuse sound field is also acceptable since when single wave comes to the diffuse part they scatter to all directions. After several reflections, scattered waves again spread across room. Therefore, even small diffusely reflective area exists; diffuse field in a room can be possible.

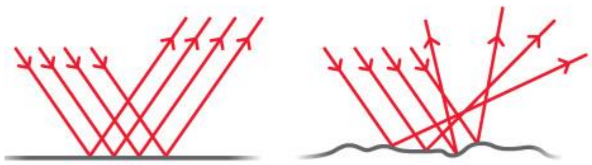


Figure 2.4. The roughness of the wall results in diffusely reflection.

As a result, Polack models reverberation as a random process that is stationary with respect to position, while non-stationary with respect to time. Different realizations of the same stochastic process are obtained at different positions at the same time. Time and initial energy change variance of the distributions. Temporal decay rate of the reverberation model depends on room acoustic characteristics. This decay rate is the significant contributor of RIR and it is used at dereverberation methods.

2.4. Reverberation Fundamentals

In the following chapters, a number of terms are used in reverberant signal processing. In order to grasp mathematical interpretations of the concepts, they are described in this section.

2.4.1. Reverberation Time

It is found that there is a persistence of sound after sound source is stopped suddenly. Auditability time of sound can represent room acoustic characteristics. However, initial level of sound source also affects time of auditability. This means auditability time cannot be characteristic property of a room, alone. It is necessary to form an objective rule related to the time in the literature in order to use it as an acoustic feature.

Reverberation time (RT_{60}) meets this claim. Reverberation time can be defined as the time that sound level decays to $1/1000000$ (60 dB in logarithmic scale) of initial value. This time quantity can characterize room acoustics, and it can be shown as in Figure 2.5:

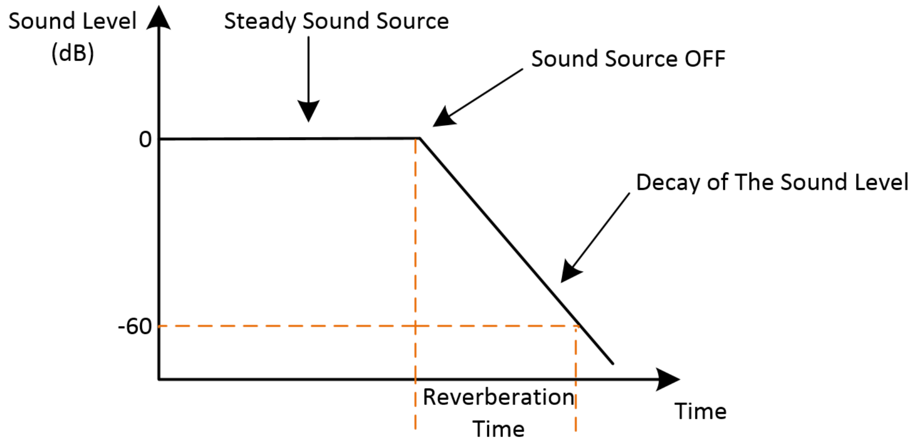


Figure 2.5. Reverberation time is the time that sound level decays 60 dB after source is stopped.

Schroeder proposed energy decay curve (EDC) to measure reverberation time [18]. If source signal is assumed as a white noise, and the noise is switched off at time t, ensemble average of the observed sound level at time $t_1 > t$ can be written as:

$$EDC_{t_1} = \int_{t_1}^{\infty} h^2(\tau)d\tau \tag{2. 15}$$

where $h(t)$ is assumed IIR type RIR, so reverberation time RT_{60} can be defined as:

$$10\,log\left(\frac{EDC_t}{EDC_{RT60}}\right) = 60. \tag{2. 16}$$

In the previous sections, it is stated that Polack models impulse response of reverberant parts as:

$$h(t) = b(t)e^{-\beta t} \tag{2. 17}$$

where β is the time constant related to reverberation time of a room. If reverberation time is known, corresponding RIR can be derived. Therefore, reverberation time is distinctive acoustic property of a room.

2.4.2. Sound Intensity

An important quantity related to sound waves is sound intensity. It is the measure of energy transportation in one second. Sound intensity is written as:

$$I = c.w \qquad (2.\ 18)$$

where c is sound wave velocity in $\mathrm{m/sec}$, w is energy density of sound in J / m^3.

2.4.3. Critical Distance

When source distance increases, strength of direct sound reduces while reverberant part is same due to uniform distribution of diffuse field. Critical distance is caused by this situation [17]. It is defined as the distance where energy density of direct part is equal to reverberant part of sound.

In a diffuse sound field, each propagation direction of sound waves has same probability, so intensity of sound waves in each direction is same. In addition, it is assumed that sound waves spread over room homogeneously, so energy density of sound is equal everywhere in room. This can be described as:

$$dw \ = \ I/c \ d\Omega \qquad (2.\ 19)$$

where Ω shows direction of the propagation in radian, w shows energy density for infinitesimal area in that direction. In diffuse field, energy density at a position can be obtained by integrating of Eq. (2.19) on all spherical angles:

$$w_r = \frac{4\pi I}{cA} \qquad (2.\ 20)$$

where A shows absorption coefficient related to exponential decay of sound. It can be written as:

$$A = 0.161\ \frac{V}{T} \qquad (2.\ 21)$$

where V is volume of enclosed space in m^3, T is reverberation time in sec.

If sound waves propagate spherically, direct sound energy density depends on the source distance and can be written as:

$$w_d = \frac{I}{c4\pi r^2}. \tag{2. 22}$$

Critical distance is where $w_r = w_d$, so it is found as:

$$D_c = \left(\frac{A}{16\pi^2}\right)^{1/2}. \tag{2. 23}$$

Critical distance is used in microphone placement in distant speech processing algorithms. If source distance is larger than critical distance, recorded speech quality will be very poor, and intelligibility of the sound will be heavily degraded. Furthermore, in order to measure reverberation time with EDC curve, source to speaker distance must be larger than critical distance because the effects of direct component reduced sufficiently with respect to reverberation in this configuration.

2.5. Literature Survey on Dereverberation Methods

Many dereverberation algorithms have been offered in speech enhancement literature since reverberation degrades quality of speech signal. Most of dereverberation algorithms combine different techniques to reduce reverberation effects. Therefore, it is not possible to classify clearly each dereverberation algorithm in the literature. However, each approach has been adapted mainly from a specific algorithm. In this section, dereverberation approaches can be classified according to the mainly used enhancement technique.

This section presents overview of the approaches which explicitly aim to speech dereverberation.

2.5.1. Beamforming Techniques

Microphone array is one of the most established approaches in speech processing literature. Therefore, beamforming technique is among the first dereverberation methods. The main concept of beamforming is simply aligning and adding the coherent parts of speech signals. In this way, the coherent part is amplified, whereas the incoherent part of speech is suppressed. In fact, beamformers have been used for capturing speech signals in noisy observations. However, they can be adapted to reverberant environments with some adjustments.

One of the simple beamforming techniques is delay and sum beamformer (DSB). Allen used DSB method for dereverberation [19] . Speech signals can be divided into two parts as early and late reverberant parts. Early reverberation is useful for intelligibility of speech, while late reverberation decreases speech intelligibility. Also, late reverberant parts at each microphone are mostly incoherent with each other. In this beamforming method, speech signals are divided into subbands since beamformers work with narrow band signals. DSB beamformer aligns and adds the subbands at each microphone with respect to phase of the coherent parts, i.e. early reverberation. As a result, high correlated parts at each microphone are amplified whereas incoherent parts (i.e. the late reverberant parts) are suppressed due to random phase structure. At the end, the dereverberated subbands are used to resynthesize the enhanced speech signals. Simple figure of a DSB is shown in Figure 2.6:

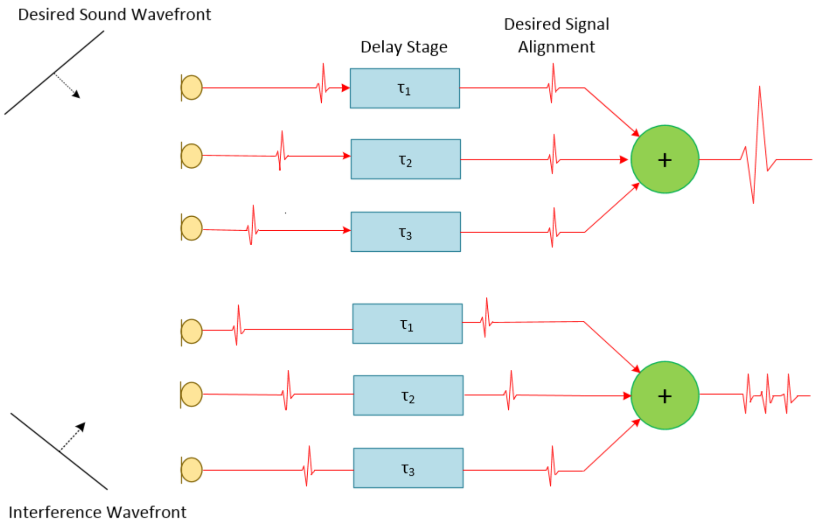


Figure 2.6. DSB method aligns the observed signals with respect to the desired signal.

DSB beamformers are proposed in several various ways in the literature. M.Kajala proposed to use constant FIR filters instead of simple delays at front end [20]. The filters are optimized in advance. This method is named ‘filter and sum’. Flanagan used two dimensional DSB to distinguish speech signals from reverberant signals [21]. Also, spherical array beamforming can be considered for dereverberation [22].

DSB is basic approach in beamforming methods. It very easy to implement; however, its ability to adapt changing environments and suppress undesired signals is limited because of the fixed beampattern. Therefore, adaptive beamformers can be used to achieve higher suppression. Frost developed data dependent, adaptive beamformer named ‘linearly constrained minimum variance algorithm’ (LCMV) [23]. In this approach, noise power is minimized while signal in the looking direction is preserved. Weights of the array are adjusted iteratively to improve directivity and suppress noise.

LCMV algorithm normally aims at denoising; however, it can also be used for dereverberation. Graffits and Jim developed two stages LCMV algorithm named ‘generalized sidelobe canceller’ (GSC) [24]. Gannot used GSC algorithm for dereverberation [25]. In GSC approach, the looking constraints and the minimization of reverberation power are separate stages. In the first stage, a fixed beamformer ensures distortionless response in the looking direction. The second stage blocks desired sound and provides reverberation reference signal. This reverberation reference is used to suppress reverberant part at the fixed beamformer output according to LMS. The basic GSC structure can be shown in Figure 2.7. Hoffman proposed GSC algorithm to cancel the desired speech in reverberant environments [26]. Dietzen combined whitening process with GSC algorithm due to estimate more reliable unbiased filter parameters [27]. Gannot estimate transfer function of input signals in standard GSC algorithm and developed TF-GSC algorithm [28]. Instead of using simple delays in the fixed beamformer stage, the transfer function ratios of the source signal at each microphone are estimated separately by single channel blind dereverberation. Then, fixed beamformer weights are replaced with respect to these transfer functions.

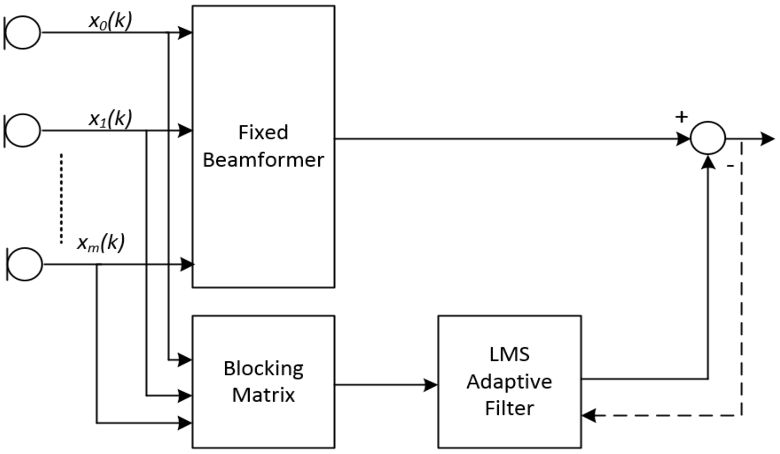


Figure 2.7. GSC algorithm use a fixed beamformer and blocking matrix.

Beamforming algorithms can be integrated with another methods in order to gain performance of systems. T. Dietzen combines GSC and spectral subtraction methods to obtain dereverberation in noisy environments [29]. He uses GSC for spatial filtering to perform denoising while spectral subtraction method provides deconvolution for the purpose of dereverberation in parallel. As it can be seen in Figure 2.8, there are two different filters in the integrated algorithm. They are estimated jointly by means of a single Kalman filter in a recursion. This integrated algorithm is similar to the proposed dereverberation in this thesis with respect to the methods. However, these methods are used one after the other to achieve better dereverberation in this work, while they are used for denoising and dereverberation separately in Figure 2.8.

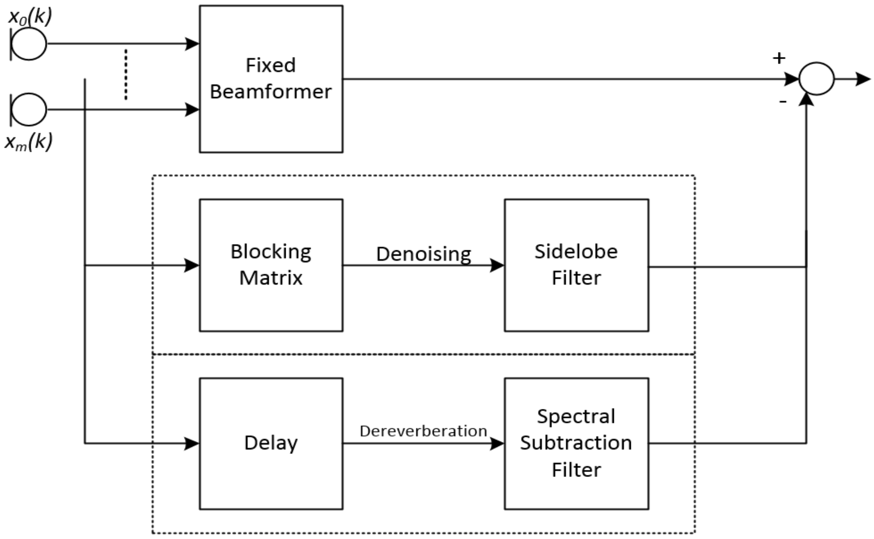


Figure 2.8. GSC algorithm use a fixed beamformer and blocking matrix.

2.5.2. Spectral Enhancement

Various spectral enhancement techniques have been used at dereverberation in the literature. An early dereverberation algorithm with cepstral processing was proposed by Duncan [30]. The study states that complex cepstral deconvolution can suppress reverberation. The complex cepstrum of sound can be described as:

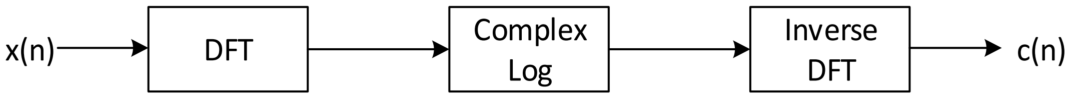


Figure 2.9. Complex cepstrum of a signal.

Deconvolution operation can be considered as subtraction in cepstrum analysis. Cepstrum of sound is a measure of frequency of variation in the log spectrum. Speech segment has slowly varying smooth log cepstrum. The fast variations in cepstrum of observed speech represent the reverberant part. These distinct peaks can be suppressed by averaging, and then the processed cepstrum coefficients are used to resynthesize the enhanced speech signal.

Flanagan proposed multi-microphone dereverberation approach [31]. In frequency bands, the microphone receiving the greatest average spectral power contributes output signal for related speech segment. In this way, output speech signal is generated by the most reliable received signals.

Spectral subtraction is very common approach for denoising and dereverberation problems. Erkelens proposed a spectral subtraction method [32]. The algorithm estimates late reverberation spectral variance (LRSV) blindly by analysis of long-term correlation in speech signals since the long term correlation is result of reverberation. After estimation of LRSV, dereverberation can be applied by spectral subtraction.

Another important method for spectral enhancement is linear prediction (LP) residual enhancement. Linear predictive coding (LPC) analysis is powerful tool for speech processing. The method considers speech as the output of an excitation signal and all pole filters. Excitation signal is quasi periodic pulses for voiced sound parts and random noise for unvoiced parts. If coefficients of the all pole filters are known, speech synthesis is possible. Main assumption for the LPC analysis is that the filter coefficients do not change with reverberation. Also, excitation signal of voiced speech contains extra peaks due to reverberation in addition to the original pulses. Therefore, dereverberation approach is to suppress the peaks due to reflections and resynthesize speech from original pulses. The simple diagram of LPC based dereverberation:

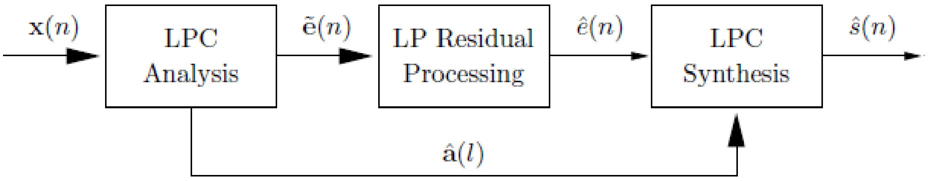


Figure 2.10. Dereverberation algorithm based on LPC analysis.

Yegnanarayana proposed LPC based single channel blind dereverberation approach [33]. The effects of reverberation on LPC residuals are studied. This approach divides speech into small segments which is approximately 2 msec and classifies them in three classes: high SRR, low SRR, only reverberant parts. Each segment is enhanced in terms of the LPC residuals with respect to the effects of reverberation. In the end, the processed residuals are used to resynthesize speech signal.

Bradford proposed a different approach with LPC analysis [34]. He combined the probabilistic distributions of LP residuals and adaptive filter for dereverberation. Amplitude distribution of LP residuals for clean speech is different from reverberant speech. Clean speech residuals consist of strong distinct pulses whereas reverberant speech residuals spread over time more randomly. Therefore, kurtosis of LP residual amplitude distribution is low for reverberant speech. Bradford states that LP residual kurtosis is reasonable metric to measure reverberation, as a result; an adaptive filter tries to increase kurtosis of LP residuals can be used to achieve dereverberation.

2.5.3. Inverse Filtering

Dereverberation process can be seen as inverse filtering of RIR. If RIR is known priori, this technique can achieve high performance in speech dereverberation.

However, this is not a realistic assumption in the practical applications. Structure of a simple inverse filtering can be shown as:

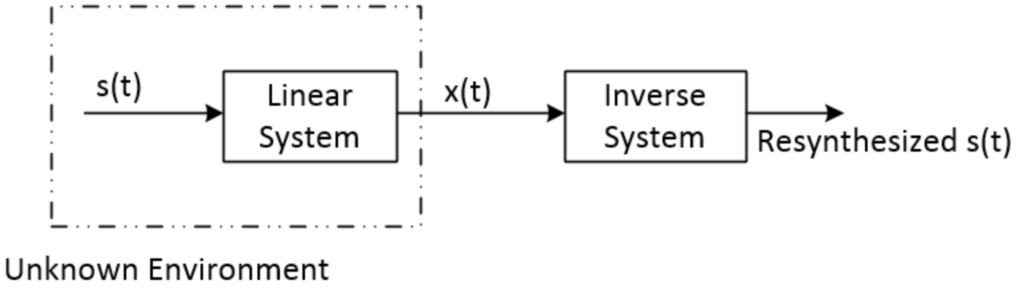


Figure 2.11. Inverse filtering of a system.

Inverse filtering method in Figure 2.11 is similar to deconvolution process. Transfer function is generally known in deconvolution problems. In fact, the system is a blind deconvolution in dereverberation problems since impulse response is not known. As a result, acoustic impulse response must be estimated somehow (at least approximately) in the inverse filtering algorithms.

Miyoshi considered the inverse of RIR as multiple of FIR filters instead of a single filter since inverse of RIR has sometimes unstable structure [35]. In this way, even if RIR has non-minimum phase, the inverse filter of the response can always be described exactly. According to this study, exact inverse response can be obtained by adding extra signal transmission channels (multi-microphone case) into the system.

Eric proposed that the channel coefficients could be estimated by higher order statistics of the observations from several microphones [36]. He uses covariance matrix of the observed signals and assumes that direct signal transfer function is orthogonal to noise subspace. The algorithm tries to find channel coefficients by eigenvectors of observed signal covariance matrix similar to MUSIC algorithm. This study is an example of blind channel identification in noisy environments. Gannot adapted this subspace methods to reverberant conditions in multi microphone case [37].

Furuya proposed a similar algorithm as Miyoshi's method [38]. However, the method estimates inverse filter blindly by observing correlation matrix in multi-microphone case. Early speech components are extracted by inverse filters and additional spectral subtraction algorithm for late reverberation suppression is used at the output of inverse filter.

Another blind inverse filtering approach is a single channel dereverberation based on harmonic structure of speech signals proposed by Nakatani [39]. This blind dereverberation method initially estimates fundamental frequency and harmonic structure of speech. Then, desired sound can be obtained by sum of the corresponding sinusoidal signals.

2.5.4. Statistical Model

Reverberation can be described by statistical models. Therefore, probabilistic approaches have been developed for dereverberation. Attias proposed Bayes-optimal signal estimation for dereverberation [40]. The algorithm uses a speech model that is pre-trained on the large clean speech database. Bayes estimation reconstructs source signal from observed microphone signal with respect to pre-trained model in probabilistic manner.

Different statistical approach is presented by Nakatani [41]. In this algorithm, harmonic structure components approach is reformulated as a maximum likelihood (ML) problem. Two types of pdf related to speech features and inverse filters are used in ML estimation. These pdfs are used to optimize inverse filters in the algorithm. In this way, the optimized inverse filters are estimated by taking account of room acoustic conditions and source speech features.

Palomäki uses Bayesian approach to classify speech features [42]. Since reverberation effects of speech can be seen in long time interval, the longtime context representation of observed speech is estimated. The patterns are classified with Bayesian approach, this algorithm is used to map reverberant speech features to clean speech features.

Nicolas proposed Lasso prediction algorithm to estimate late reverberant part of speech [43]. Late reverberant part is assumed as a linear combination of the previous frames in time-frequency domain. When Lasso algorithm is applied to predict late reverberation in frequency domain, residuals of the predictions are direct part of speech. Magnitude ratio of late reverberant to direct part is used to generate a filter for dereverberation.

Reverberation is smearing of energy of the previous samples over time. Therefore, spectral subtraction would be very useful method if smeared energy was known exactly. Based on this idea, a novel approach for dereverberation is presented by Lebart [44]. He proposed to combine Polack statistical model for reverberant room and spectral subtraction approaches. Simple structure of the approach can be shown in Figure 2.12:

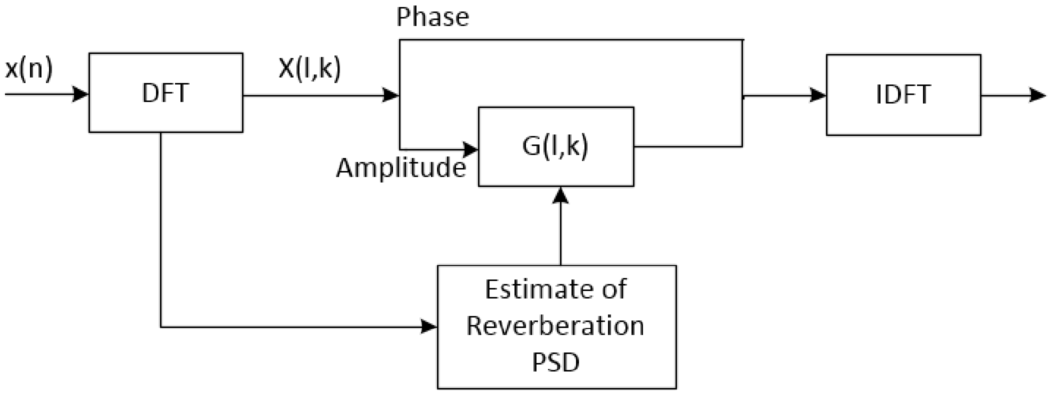


Figure 2.12. Dereverberation algorithm estimates the PSD of reverberation and generate a filter for reverberation.

Polack statistical model is used to estimate the PSD of the reverberant part. Then, these components are removed by spectral subtraction method.

2.6. Beamforming Method

Multi-microphone solutions have superiority over single microphone techniques in speech enhancement studies. Therefore, multi-microphones are used to obtain better performance in speech processing methods. In multi-microphone speech processing approaches, beamformers are frequently used. Beamformers can also be used in dereverberation algorithms. In this section, detailed background of beamformers is given.

Beamformers can be formulated as spatial filters which operate on inputs of microphone array in order to generate directivity. The main feature of a beamformer is to provide directional signal transmission, and they are studied in different signal processing areas such as direction of arrival estimation and enhancing the signal from specific direction.

Beamformers implement steering function by a weighted sum and this function provides the sensor array to rotate towards a specific direction algorithmically. In this way, sensors can pick up the desired signal more accurately, and the noise which comes from other directions can be suppressed. Steering function can be generated by shifting sensor signals appropriately in time domain. In frequency domain, it is implemented by applying exponential weights to STFT coefficients of the sensor signals. The weighted sum is controlled according to priori constraints to further improve the performance of beamformers [45]. Simplified diagram of a beamformer can be shown in Figure 2.13.

General equation of a beamformer output is:

$$z[n] = \sum_{m=1}^{M} w_m\,[n] y_m[n] \tag{2. 24}$$

where M is sensor numbers, $y_m[n]$ is the m^{th}sensor signal, $w_m[n]$ is the weight of m^{th}sensor.

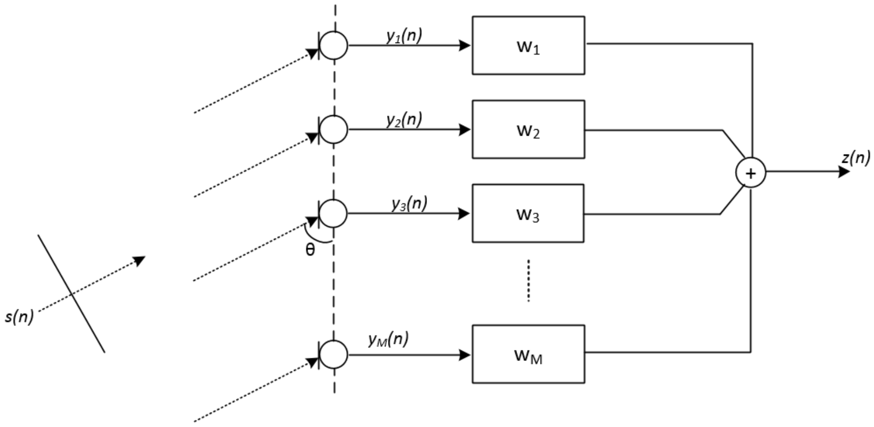


Figure 2.13. The beamformer synchronizes the input signals. The weighted coefficients are adjusted with respect to a constraint.

Beamformers work with narrow band signals because their transfer functions change with frequency of signal. Low-pass filters or subband decompositions can be used to process rich frequency content speech signals. The calculations can be made for broadband signals by processing its narrowband components separately.

2.6.1. Signals in Beamforming

Beamformers work with propagation path differences of received sound signals. Received signals at different microphones are out phase or in phase because of unequal sound wave paths. The spatial filters use these phase relations to amplify or attenuate the signals. The main object is to use these phase differences as constructive for the desired sounds, destructive for interferences. Therefore, it is important to describe the phase relations between microphones mathematically in order to adjust filter coefficients.

Each propagation channel in microphone array can be described by attenuation and delay parameters. At discrete time instant n, the signal at m^{th} microphone is described as [45]:

$$y_m[n] = \alpha_m s[n - t - F_m(t)] + v_m[n] \tag{2. 25}$$

where $s[n]$ is the desired signal, α_m is attenuation constant due to propagation, $F_m(t)$ is the relative time delay of the microphones with respect to reference, t is the time delay due to propagation of sound wave between source and reference microphone, $v[n]$ is noise parameter. (In the frequency domain, the delay parameters are written as exponential functions.)

$F_m(t)$ is significant time delay in beamforming method. It originates from microphone array structure. Sound waves reach the microphones at different time instants with respect to the direction of arrival (DOA). $F_m(t)$ is used to describe time difference between reference microphone and related microphone. As in Figure 2.14, first microphone can be accepted as reference microphone and relative delay depends on DOA as well as the microphone number.

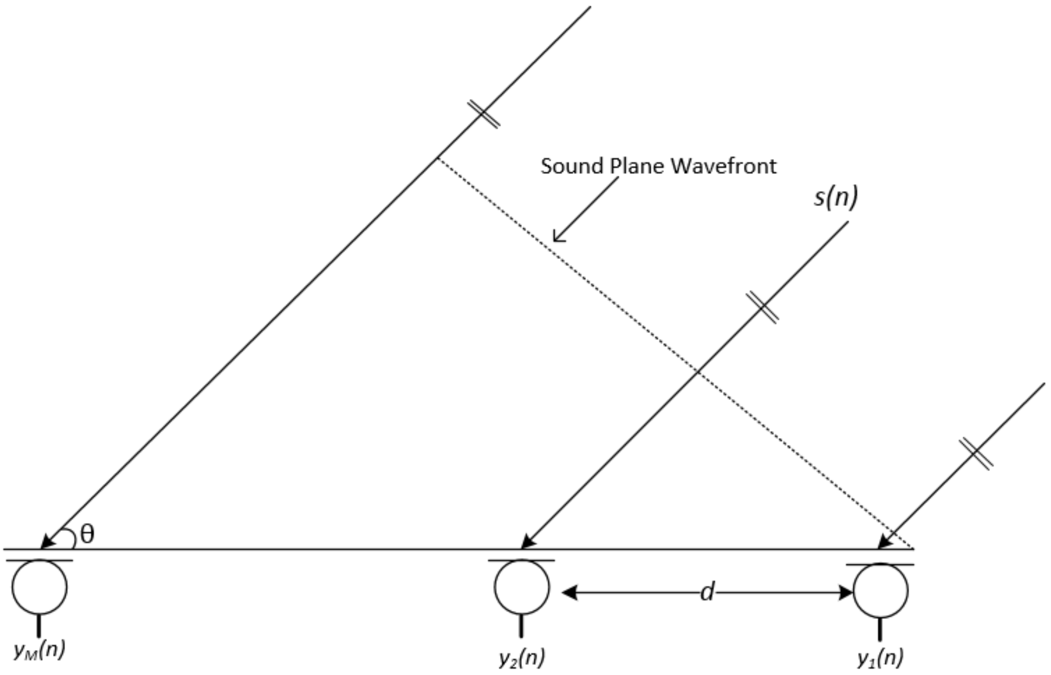


Figure 2.14. Sound wave propagation difference.

Especially in reverberation problems, sound waves are modeled by plane waves since they propagate along the distance larger than Eq. (2.7) for most of the frequencies due to reflections. This assumption is useful to specify the relative delay parameters.

As it can be seen in Figure 2.14, each relative time delay is described as:

$$F_m(\tau) = (m-1)\tau = \frac{(m-1)dcos\theta}{c} \qquad (2.\ 26)$$

where d is the distance between microphones, θ is the DOA, c is propagation velocity. Note that Eq. (2.26) is continuous time domain representation and it must be converted to discrete time domain initially; however, this time relation is assumed to be in discrete time domain in the next equations.

The aim of beamformer is reducing effect of $v[n]$ relative to the desired signal i.e. improving SNR in observed signal. In dereverberation case, $v[n]$ is replaced by reverberation, in this way the system can suppress reverberation.

2.6.2. Array Gain of Beamformer

The main object of a beamformer is to improve SNR of input signal at each subband, and it is achieved by array geometry and sensor weights. These parameters must be adjusted with respect to specific metric which is named *'array gain'*. Array gain is obtained by ratio of beamformer output SNR to the reference microphone SNR. The performance of a beamformer is evaluated by this metric. Mathematical expression is shown as:

$$\begin{aligned} Array\ Gain\ = G_a &= \frac{SNR_{output}}{SNR_{input}} \\ &= \frac{\boldsymbol{w}^H SNR_{input} \boldsymbol{w}}{SNR_{input}} \end{aligned} \qquad (2.\ 27)$$

where $\boldsymbol{w}$ is vector of microphone weights. In order to maximize array gain of a beamformer, derivative of this expression for each subband is used.

2.6.3. Beampattern of a Beamformer

Response of a beamformer inherently depends on the direction of sound signals. This response can be shown by beampatterns. A beampattern is the graph of beamformer output which shows output response versus DOA. Besides, frequency affects beamformer response; therefore, beampatterns are evaluated for a specified frequency. Typical beampattern can be shown as in Figure 2.15:

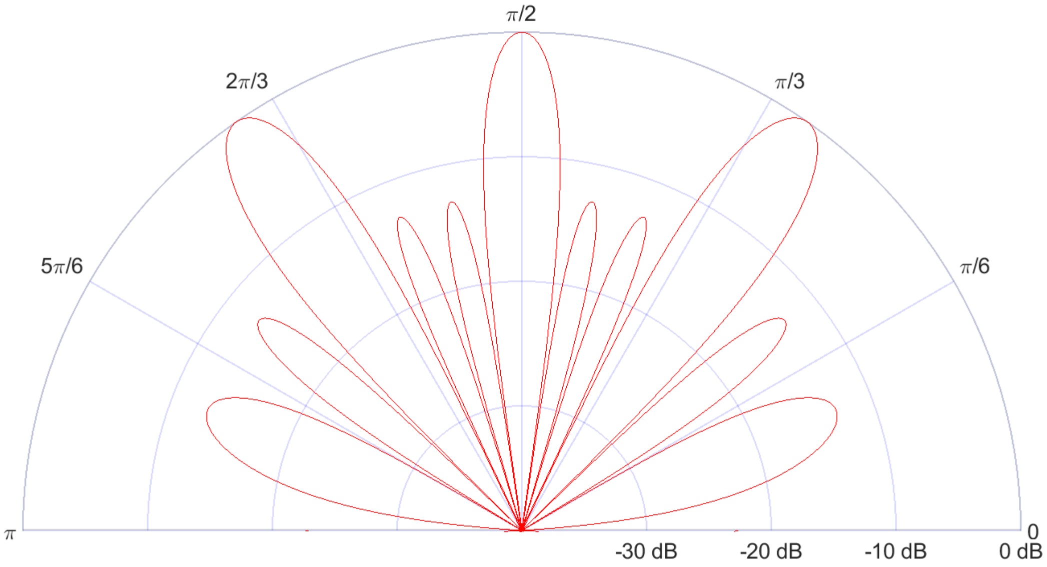


Figure 2.15. Beampattern of DSB (2000 Hz signal in 4 microphone case).

This graph shows magnitude of beamformer response for a specific frequency. As it can be seen in Figure 2.15, beamformers provide spatial selectivity, i.e. reinforce some signals from particular directions while attenuate other arrival directions. By setting the beamformer weights, strength and direction of attenuation can be adjusted.

The output response at the desired angle in a beampattern is main lobe while the other attenuated responses are side lobe. The height of side lobes represents attenuation for the unwanted signals. For ideal beamformer, side lobes magnitude should be very small compared to that of main lobe. The width of main lobe should be as small as possible in order to provide resolution in directions.

In Figure 2.15, the spatial selectivity does not work very well because there are three main lobes at different directions, so corresponding arrival directions are not distinguished from each other. In the beamforming literature, it is defined as spatial aliasing. It occurs when input signals are sampled too slowly at the sensors to observe different phases of sound waves. In order to prevent spatial aliasing, spatial sampling theorem must be satisfied [46]. According to this theorem, the distance between sensors must be lower than half of the minimum wavelength in the input signal content:

$$d_{sensors} < \frac{\lambda}{2} \quad . \tag{2. 28}$$

In order to analyze beamformers and their beampatterns mathematically, it is useful to study a simple beamformer structure.

2.6.4. Classical Beamformer

The simplest form of beamformers is DSB. This beamformer just steers the main lobe of beampattern to specific direction by compensating path differences of sound waves [45]. This can be done by shifting each sensor signal with respect to the reference microphone signal. The phase relations are found by DOA of the desired signal. The first microphone is assumed as the reference microphone in the equations.

Noise signal has not a deterministic structure like the desired signal, and their phases do not depend on the source direction. Therefore, there is no perfect matching of noise signals when the phases are shifted according to DOA. In this derivation, noise signals are assumed as stationary incoherent parts of the recorded signals.

For DSB, the simplest form of beamformer, all weights can be taken as $\frac{1}{M}$ where M is the microphone number. This procedure turns into just taking mean of the synchronized sensor outputs.

$$z_{DSB}[n] = \sum_{n=1}^{M} \mathrm{w_m} y_m[n + F_n(\tau)] \quad . \tag{2. 29}$$

Take all coefficients as $\frac{1}{M}$:

$$\begin{aligned} z_{DSB}[n] &= \frac{1}{M}\sum_{m=1}^{M} y_m[n + F_m(\tau)] \\ &= \frac{1}{M}\sum_{m=1}^{M} y_{shifted}^{m}\,[n] \\ &= \alpha_s s[n] + \ v_s[n] \end{aligned} \tag{2. 30}$$

$$\alpha_s = \frac{1}{M}\sum_{m=1}^{M} \alpha_m \tag{2. 31}$$

$$v_s[n] = \frac{1}{M}\sum_{m=1}^{M} v_m[n] \tag{2. 32}$$

where $s[n]$ is the desired signal; $\alpha_s, v_s[n]$ are the average of attenuation constants and noise signals respectively. Now, assume that noise signal is white noise for simplicity, analyze input SNR and output SNR:

$$\begin{aligned} SNR_{output} &= \frac{1}{M^2}\left(\sum_{m=1}^{M} \alpha_m\right)^2 \frac{E(s^2[n])}{E(v_s^2[n])} \\ &= \frac{(\sum_{m=1}^{M} \alpha_m)^2}{M}\frac{\sigma_s^2}{\sigma_v^2} \end{aligned} \tag{2. 33}$$

where $\frac{\sigma_s^2}{\sigma_v^2}$ is the input SNR. α is the attenuation constant of the desired signal. In normal room conditions, the attenuation of a direct sound wave can be neglected. (α_m=1) Therefore, $SNR_{output} = M\ .SNR_{input}$ and $SNR_{output} > SNR_{input}$.

Another way of analyzing performance of a beamformer is frequency response of the spatial filter, i.e. beampattern. In order to obtain output response of a beamformer, its transfer function is investigated in frequency domain. Frequency response of the filter is generated by delay parameters, and they are described by exponential terms in frequency domain.

All conditions in time domain calculations are same for frequency domain analysis. The delay functions are written according to Figure 2.14. The frequency response of DSB beamformer:

$$\mathrm{H_{DS}}(\varphi,\theta) = \frac{1}{M}\sum_{m=1}^{M}\left[\mathrm{e}^{\frac{\mathrm{j\omega d(m-1)cos}\theta}{c}}\right]\left[\mathrm{e}^{\frac{\mathrm{-j\omega d(m-1)cos}\varphi}{c}}\right]$$

$$= \frac{1}{M}\sum_{m=1}^{M}[\mathrm{e}^{\frac{\mathrm{j\omega d(m-1)(cos}\theta\mathrm{-cos}\varphi)}{c}}] \qquad (2.\,34)$$

where φ is steering direction angle, θ shows DOA with horizontal axis. By solving the equation, frequency response of a beamformer is obtained as:

$$A_{DS} \ = |H_{DS}(\varphi,\theta)|$$

$$A_{DS} \ = |\frac{\sin(Nwd(cos\varphi - cos\theta)/2)\ /c}{Nsin(wd(cos\varphi - cos\theta)/2)/c}| \qquad (2.\,35)$$

In frequency response equation (2.35), the significant parameters of beamformers can be easily seen. Also, beampattern of microphone array can be drawn according to this equation. Typical frequency response of beamformer with $\theta = 90°$ $(DOA = \ 90\,°)$ with respect to steering angle can be shown as:

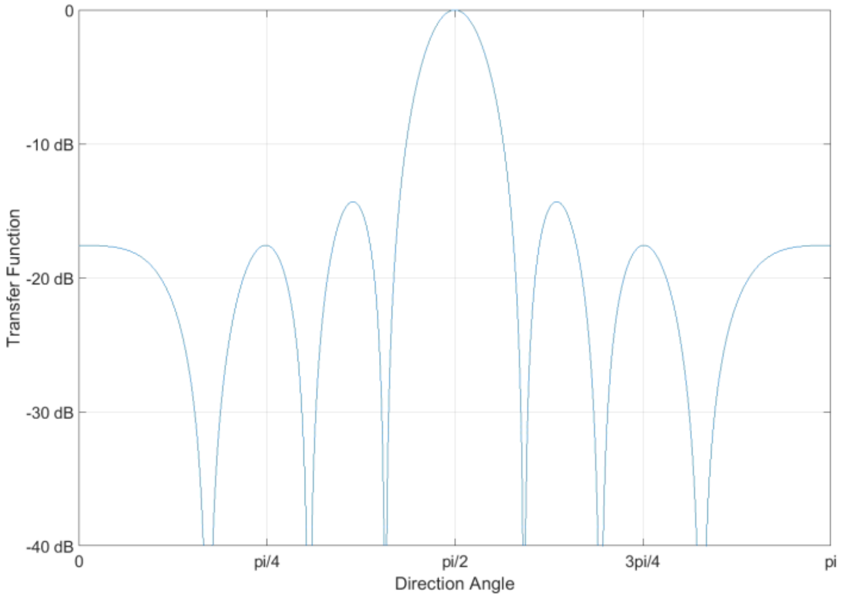


Figure 2.16. Frequency response of a beamformer in terms of steering angle when $\mathrm{DOA} = 90\,°$.

As it is shown in Figure 2.16, the maximum response can be obtained when the desired signal is in looking direction of microphone array. Also, it can be concluded that the microphone number and the distance between microphones are the parameters of beamformer. Beamwidth decreases when the number of sensors, the interval between sensors and the frequency of signal increase. The height of sidelobes also can be adjusted by same parameters.

In Figure 2.16, there are nulls in the transfer function of the beamformer. The purpose is to adjust these directions with respect to interference signals to obtain maximum attenuation. However, frequency response would be fixed unless physical changes are made in DSB beamformers. The problem is whether it is possible to adjust frequency response without physical changes. This procedure can be possible by varying weights. Therefore, more comprehensive beamformer structures are developed to adjust beampattern characteristics with respect to the desired and interference signals adaptively.

2.6.5. Adaptive Beamformers

In DSB, array processing parameters do not change dynamically. Delay parameters change with varying DOA, however microphone coefficients do not change with respect to noise characteristics. Therefore, the main characteristics of beampattern such as nulls, beamwidth, and sidelobes are fixed. Frost proposed a beamformer algorithm based on LMS method [23]. In this method, the weights of array can be adjusted dynamically frame by frame with respect to observed signals. This is the basis of adaptive beamformers.

The comparison of fixed and adaptive beamformer can be shown in Figure 2.17. Red and blue lines show beampattern of fixed and adaptive beamformers respectively. The null directions of fixed beamformer are independent of interference directions while adaptive beamformer adjusts null directions with respect to interference field dynamically. Therefore, adaptive beamformers can be considered as data dependent microphone array while fixed beamformers are data independent.

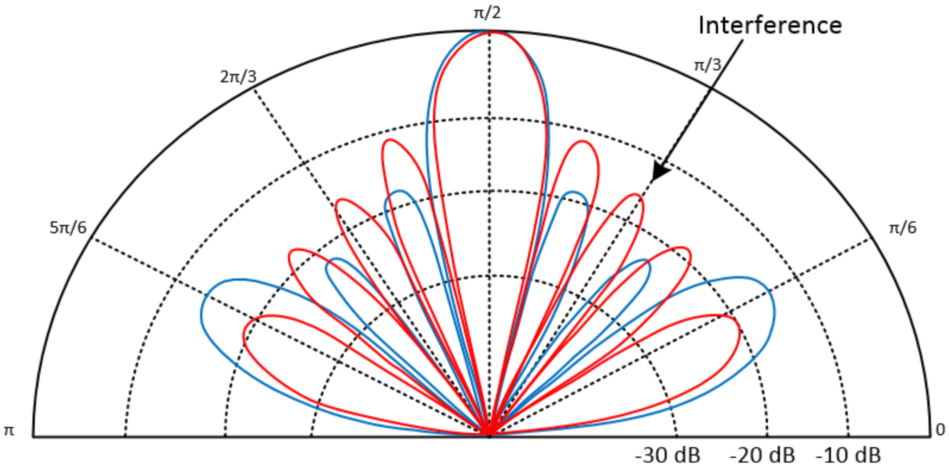


Figure 2.17. Frequency response of a fixed beamformer (DSB) and an adaptive beamformer (MVDR) in an environment where DOA of interference is 60 °.

In this thesis, MVDR beamformer is used to deal with reverberant parts. Therefore, more detailed knowledge about how adaptive beamformers work is presented in Chapter 3.

2.7. Statistical Model of Speech Signal

Statistical model of speech signals plays important role in the speech processing techniques. Due to unknown source signal and channel responses, speech modeling is very common in dereverberation techniques. Especially, if there is no prior knowledge about the system in single channel dereverberation approach, modeling of speech is crucial.

There is no deterministic structure to model speech signals due to various and non-stationary characteristics of sound waves. Therefore, statistical modelling is popular for speech signals. This approach considers each speech sample as a realization of a stochastic process and relies on this process to describe entire speech.

Accuracy of the statistical model is essential because the model directly affects performance of speech processing method. However, accuracy is not adequate to determine the appropriate model. The selected model must be also mathematically tractable to process speech signals in a reasonable time.

Various stochastic models are proposed to model speech. First, Davenport studied distribution of speech samples in time domain and stated that the amplitude distribution varies exponentially around the mean [47]. Therefore, in many applications, speech signals have been assumed Gaussian process in order to simplify the calculations [48] . Besides, Laplace and Gamma distributions have been also suggested with respect to the length of speech segment [49].

Distributions of DFT coefficients have been also considered since many speech processing algorithms are in frequency domain instead of time domain. Each DFT coefficient is weighted sum of the speech samples, indeed. If the speech samples are assumed as independent random variables, according to Central Limit Theorem DFT coefficients have roughly Gaussian shape regardless of distribution of speech samples in time domain. However, there are also different proposed distributions, and Laplacian distribution is used in some contributions for DFT coefficients [50].

Hendriks made an objective comparison of accuracy of the distributions of clean-anechoic speech signals in both time and frequency domain in different conditions [51]. He found that speech segment size had an effect on accuracy of the distributions. Therefore, it is necessary to investigate the distributions according to segment duration. It is found that the distribution of speech samples is Laplacian in 30 – 200 msec segments while Gaussian distribution fits best for less than 20 msec. Furthermore, the distribution of DFT coefficients is also studied in similar way. It is found that the coefficients are classified as Gaussian for 30 msec speech segments.

As a result, since speech depends on both speaker and environment, it is not an easy work to generate a model which always fits speech signals. However, an exponential distribution generally can be used to represent signals. According to mathematical tractability, accuracy and speech segment size in the algorithms; Gaussian distribution seems reasonable.

CHAPTER 3

DEREVERBERATION METHOD

3.1. Introduction

Dereverberation is a blind deconvolution process, since there is no prior knowledge about room impulse response and source signal. It means that there are parameters to be estimated before processing; however, it is difficult to estimate all of these parameters accurately. This issue leads to use statistical models in dereverberation approaches since the probabilistic models can be used to derive parameters of the system. These models also reduce computation complexity significantly, so even real time dereverberation becomes possible. In Chapter 2, the statistical models of room acoustics and sound waves are explained. The dereverberation methods in this chapter are based on these statistical models.

Dereverberation methods generally use well known speech processing algorithms. However, these algorithms have low performance in reverberant environments. Therefore, it is necessary to adapt these algorithms to the nature of reverberation. The statistical models of reverberation are used for this purpose. Besides, each algorithm suppresses reverberation in different approach. Therefore, combining various algorithms is popular in order to improve performance of dereverberation.

In this work, two main aspects of reverberant signals are combined to improve the performance. First, the reverberant part can be seen as a replica of the direct speech due to reflections of sound waves, so it causes long time correlations in recorded speech. It is possible to use these correlations to suppress the replicas in recorded speech signals. The second point of the view is to treat reverberation as independent components, so a directional microphone array algorithm can be applied to capture the desired speech signal while suppressing reverberation.

In this chapter, two different methods corresponding to these two different aspects of reverberation are presented. Both of these methods rely on the statistical models of room impulse response and source signal. Therefore, there is no need of prior knowledge of the system. This makes both methods useful in practical applications. Furthermore, combining of the methods and advantages of the combined system are explained in this chapter.

3.2. Problem Statement

In time domain, recorded speech is convolution of anechoic source signal and RIR. Let $s[n]$ be the desired speech source, i.e. anechoic speech in a room and n denote the discrete time index of the signal. The observed speech in a room can be modeled as:

$$y_m[n] = s[n] * h_m[n] + v_m[n] \tag{3. 1}$$

$$y_m[n] = \sum_{k=0}^{L-1} h_m[k]\, s[n-k] + v_m[n]$$

where m denotes microphone index, $h_m[n]$ is impulse response between source and related microphone, $v_m[n]$ denotes microphone observation noise. In Chapter 2, it has been stated that reverberation can be divided into two components as ‘early reverberation’ and ‘late reverberation’. Therefore, the observed speech $y[n]$ at the m^{th} microphone can be described as:

$$y_m[n] = d_m[n] + r_m[n] + v_m[n]. \tag{3. 2}$$

where d_m, r_m denotes early reverberation and late reverberation respectively.

In this thesis, it is assumed that the room is a quiet environment in order to simplify dereverberation calculations. Therefore, $v_m[n]$ is neglected, and $x_m[n]$ is used to denote the observation signal without noise:

$$x_m[n] = d_m[n] + r_m[n]. \tag{3. 3}$$

Human ears cannot distinguish the direct sound from early reverberation [10] . There is no need to evaluate the direct part and early reverberation separately, so $d_m[n]$ represents the sum of early reverberant and direct parts (in the rest of the document, early reverberation part contains the direct part of speech) while $r_m[n]$ represents late reverberation.

$$d_m[n] = \sum_{k=0}^{T} h_m[k]\, s[n-k] \tag{3. 4}$$

$$r_m[n] = \sum_{k=T+1}^{L-1} h_m[n]\, s[n-k].$$

T is the time index of transition from early reverberant to late reverberant part in RIR. It is ‘mixing time’ that is the elapsed time for diffuse field to exist in a room [6]. The mixing time depends on the volume of enclosed space. However, various estimation methods of the time exist in the literature [52] .

In frequency domain, speech signal can be processed frame by frame. Convolutive transfer function (CTF) model is used to describe STFT coefficients of the observed speech [53]. According to this model STFT coefficients of an observed speech frame can be written in terms of previous frames of source signal. By using this model, STFT coefficients of the observed speech, $Y_m(l,k)$ can be described as:

$$Y_m(l,k) = \sum_{n=0}^{L-1} H_m^*(n,k)\, S(l-n,k) + V_m(l,k) \tag{3. 5}$$

where l is the frame index, k is the frequency bin number, L is length of the transfer function. Also $S(l,k), V_m(l,k)$ denote STFT coefficients of the anechoic source signal $s[n]$ and noise signal $v_m[n]$. $H_m(l,k)$ relates the observation to the past elements of $S(l,k)$.

CTF representation is a useful model in reverberant environments. The time domain convolution between source signal and RIR can be approximated as convolution between STFT coefficients of speech and convolutive transfer function of room for each frequency bin independently. In this way, reverberant signals can be described in terms of previous frames in the subbands. Also, it can explain the effects of previous frames on the recent frame due to reverberation mathematically, so this representation is useful for dereverberation methods based on linear prediction. If the observation noise is neglected, STFT coefficients of observed speech frame have contributions from both early reverberation and late reverberation parts:

$$X_m(l,k) = D_m(l,k) + R_m(l,k) \tag{3. 6}$$

where $D_m(l,k)$ and $R_m(l,k)$ show frequency domain representations of $d_m[n]$ and $r_m[n]$ in Eq. (3.3). Early reverberant and late reverberant components are written in frequency domain by CTF model:

$$D_m(l,k) = \sum_{n=0}^{D-1} H_m^*(n,k)\, S(l-n,k) \tag{3. 7}$$

$$R_m(l,k) = \sum_{n=D}^{L-1} H_m^*(n,k)\, S(l-n,k)$$

where D is the frame number related to the mixing time.

The matrix form of CTF model:

$$X_m(l,k) = \boldsymbol{H}_m^H(k)\, \boldsymbol{S}(l,k) \tag{3. 8}$$

where $\boldsymbol{H}_m(k) = [\,H_m(0,k) H_m(1,k)\; H_m(2,k)\; \ldots\, ..\; H_m(L-1,k)]^{\mathrm{T}}$ is vector form of transfer function, $\boldsymbol{S}(l,k) = [S(l,k)\; S(l-1,k)\; S(l-2,k) \ldots .\; S(l-L+1,k)]^{\mathrm{T}}$ is the source vector.

3.3. Proposed Approach

The proposed dereverberation algorithm consists of two different techniques, adaptive beamforming and linear prediction, one after the other. Both techniques are adapted to the nature of reverberation. At the first stage, adaptive beamforming is applied by a microphone array as a multi-microphone dereverberation then, linear prediction is applied to the output of the array as a single channel dereverberation. Estimate of late reverberation PSD (ϕ_r) in the microphone array stage is used in the second stage as well. The overview of the proposed dereverberation algorithm is shown in Figure 3.1. Initially, the methods will be presented individually and then the combined approach will be explained at the end of this chapter.

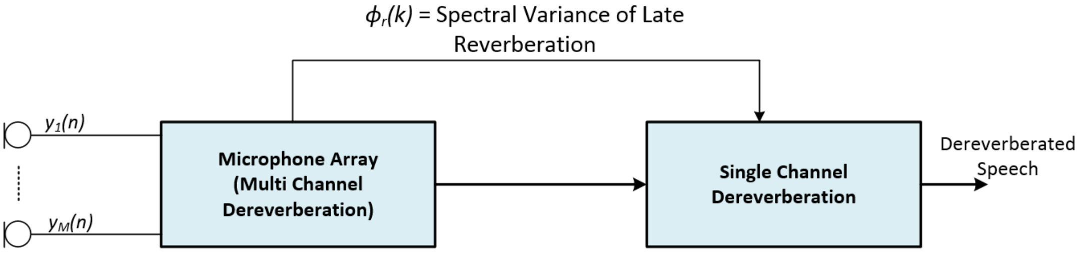

$\phi_r(k)$ = Spectral Variance of Late Reverberation
$y_1(n)$
$y_M(n)$
Microphone Array (Multi Channel Dereverberation)
Single Channel Dereverberation
Dereverberated Speech

Figure 3.1. After the observed reverberant signal is processed in the microphone array, single channel dereverberation algorithm is applied the output of microphone array.

3.4. Microphone Array

One approach for dereverberation is to treat reverberant part as an unwanted diffuse sound field which is independent of the direct components, so it is possible to design a beamformer which provides directivity towards the direct path of the incoming speech signal while nulling out the directions of reverberant components. Thus, the effect of unwanted components in observed speech signal can be reduced. In a reverberant environment, the direction and magnitude of the interferences are not fixed. Therefore, a fixed beamformer structure would not give substantial improvements to dereverberation performance. Adaptive beamformers can be used to deal with the time-varying nature of reverberation directions and magnitudes.

Adaptive beamformers adjust weights of the microphones to satisfy two criteria. First one is to maximize SNR at the output of beamformer for each subband. This can be achieved by maximum suppression to the interferences in microphone signal. Second one is to have no distortion of the source signal in the absence of interferences [23]. It is expected that the main lobe of a beampattern has a constant gain in the source signal's direct path direction.

3.4.1. Minimum Variance Distortionless Response (MVDR) Beamformer

One of the most widely used adaptive beamformer technique is MVDR beamformer. MVDR formulation aims a constant gain at the looking direction and nulling the interference directions. In this work, MVDR is implemented to provide high performance at filtering late reverberation while obtain unity gain to the direction of speech. MVDR adjusts microphone coefficients to minimize output power and not to disturb the desired signal at the looking direction [45].

In frequency domain, received microphone signals can be written in vector form

$$\boldsymbol{Y}(l,k) = \boldsymbol{X}(l,k) + \boldsymbol{V}(l,k) \tag{3. 9}$$

$$= S(l,k)\,\boldsymbol{g}(k) + \boldsymbol{V}(l,k)$$

where $\boldsymbol{X}(l,k) = [X_1(l,k)\; X_2(l,k)\; X_3(l,k)\; \ldots.\; X_M(l,k)]^T$

$\boldsymbol{g}(k) = [g_1(k)\; g_2(k)\; g_3(k)\; \ldots.\; g_M(k)]^T$

$\boldsymbol{V}(l,k) = [V_1(l,k)\; V_2(l,k)\; V_3(l,k)\; \ldots.\; V_M(l,k)]^T$ are M dimensional vectors.

Subscripts in these equations show the microphone number in the array. $S(l,k), \boldsymbol{X}(l,k), \boldsymbol{g}(k), \boldsymbol{V}(l,k)$ denote STFT coefficients of the unknown source signal, the desired parts of observed signals, transfer functions between the source and related microphones, and ambient noise, respectively. The following equations do not depend on the frame number. Therefore, frame number $'l'$ is not shown in the equations.

Let $\boldsymbol{w}(k)$ be the vector of beamformer coefficients,

$$\boldsymbol{w}(k) = [w_1(k)\, w_2(k)\, w_3(k)\ \ldots..w_M(k)]^T$$

and $z(k)$ be the beamformer output:

$$z(k) = \boldsymbol{w}^H(k)\, \boldsymbol{Y}(k) \tag{3. 10}$$

$$= \boldsymbol{w}^{\mathrm{H}}(k)\, [\boldsymbol{g}(k)s(k) + \boldsymbol{V}(k)]$$

If noise signal is independent of the desired signal, the power spectral density of the beamformer output can be written as:

$$\phi_z(k) = \boldsymbol{w}^H(k)\boldsymbol{\phi}_x(k)\boldsymbol{w}(k) + \boldsymbol{w}^H(k)\boldsymbol{\phi}_v(k)\boldsymbol{w}(k) \tag{3. 11}$$

where $\boldsymbol{\phi}_x(k), \boldsymbol{\phi}_v(k)$ are the PSD of desired signal and noise at input respectively, $\boldsymbol{w}^H(k)\boldsymbol{\phi}_v(k)\boldsymbol{w}(k)$ is the PSD of noise at the output. One aim of MVDR is the minimization of the noise PSD at the output and the other is fixed gain in the desired signal direction [45]. In our case, the gain is accepted as $g_1(k)$ that is the transfer function of the reference microphone. As a result, the MVDR problem is stated as:

$$\boldsymbol{w}_{mvdr} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \boldsymbol{w}^H(k)\boldsymbol{\phi}_v(k)\boldsymbol{w}(k)\ \ subject\ to\ \boldsymbol{w}^H(k)\, \boldsymbol{g}(k) \tag{3. 12}$$

$$= g_1(k).$$

Lagrange multiplier is used to solve these constraints. The cost function with Lagrange multiplier can be defined as:

$$L(k) = \boldsymbol{w}^H(k)\boldsymbol{\phi}_v(k)\boldsymbol{w}(k) + \left[\lambda\left(\boldsymbol{w}^H(k) * \boldsymbol{g}(k) - g_1(k)\right)\right]. \tag{3. 13}$$

The solution of MVDR algorithm is given as:

$$\boldsymbol{w}^H(k) = \frac{\boldsymbol{g}^{\mathrm{H}}(k)\boldsymbol{\phi}_v(k)^{-1}g_1^*(k)}{\boldsymbol{g}^{\mathrm{H}}(\mathrm{k})\boldsymbol{\phi}_v(k)^{-1}\boldsymbol{g}(k)}. \tag{3. 14}$$

3.4.2. MVDR in Reverberation

Now, we will present the MVDR formulation for a particular model of reverberant speech in a room under reasonable assumptions. In a reverberant room, the observed signal at the m^{th} microphone can be written in frequency domain as:

$$Y_m(l,k) = D_m(l\,,k) + R_m(l,k) + V_m(l,k) \qquad (3.\ 15)$$

where $D_m(l,k)$, $R_m(l,k)$, $V_m(l,k)$ represent STFT coefficients of early reverberant part, late reverberant part and observation noise signals, respectively. Observation noise will be neglected due to simplicity of calculations.

The purpose of MVDR is to suppress late reverberant part while maintaining directivity towards early reverberant part because early reflections cannot be distinguished by human ears while late reverberation is harmful for speech intelligibility [10].

Transfer function ratios of early reverberant part must be known in order to provide directivity. Since early reverberant part consists of direct part and multiple early reflections as in Figure 3.2, the exact transfer function of the microphone signals cannot be measured accurately. Therefore, dividing $D_m(l\,,k)$ component into two smaller components is useful:

$$D_m(l,k) = G_d^m(k)\,S(l,k) + D_l^m(l,k) \qquad (3.\ 16)$$

where m shows the microphone index, $G_d^m(k)\,S(l,k)$ represents the earliest components, i.e. direct part and the nearest early reflections generated by only current frame of the source signal in Eq.(3.7), $G_d^m(k)$ shows transfer function of these components, $D_l^m(l,k)$ shows the lagged early reflections generated by previous frames of the source signal. Direct part of speech is nearly instantaneous, and speech signals are processed frame by frame in frequency domain; therefore, the direct part of speech cannot be described individually in frequency domain. $G_d^m(k)\,S(l,k)$ represents the frame of direct part; however, it contains also a few following early reflections due to non-zero frame length. (This part is named ‘the earliest components’ in this work) Steering direction of MVDR is adjusted with respect to the earliest components because this part contains fewer reverberation, and it is dominated by direct part. Since the reflection number is low, it is easier to find a deterministic transfer function for this part. However, steering towards just the earliest components causes some of the lagged early reverberant components not to be processed properly within the MVDR. In Chapter 2, it is stated that early reverberation is not distinguished from the direct part by human ears, so these unprocessed do not reduce speech quality. As a result, although lagged early reverberation is not taken in consideration in steering process properly; steering towards the earliest components strengthens direct components with respect to the late reverberation.

Vector form of the observed reverberant signal can be written as

$$\boldsymbol{x}(l,k) = \boldsymbol{g_d}(k)S(l,k) + \boldsymbol{D_l}(l,k) + \boldsymbol{r}(l,k) \qquad (3.\ 17)$$

where $\boldsymbol{g}_d(k) = [G_d^1(k)\ \ G_d^2(k) \ldots\ldots G_d^M(k)]$ is vector form of transfer functions of the earliest components. Thus, if the lagged early reverberation is ignored, the observed reverberant speech signals are formulated similarly to Eq. (3.9). The source signals in MVDR algorithm are replaced by $\boldsymbol{g_d}(k)\ S(l,k)$ while the noise is replaced by $\boldsymbol{r}(l,k)$ in Eq. (3.11). In this way, standard MVDR solution in Eq. (3.14) provides dereverberation. In this situation, looking direction is adjusted with respect to $\boldsymbol{g_d}(k)$, while array gain is optimized with respect to $\boldsymbol{r}(l,k)$ vector. $\boldsymbol{D_l}(l,k)$ is the residual unprocessed early reverberation component.

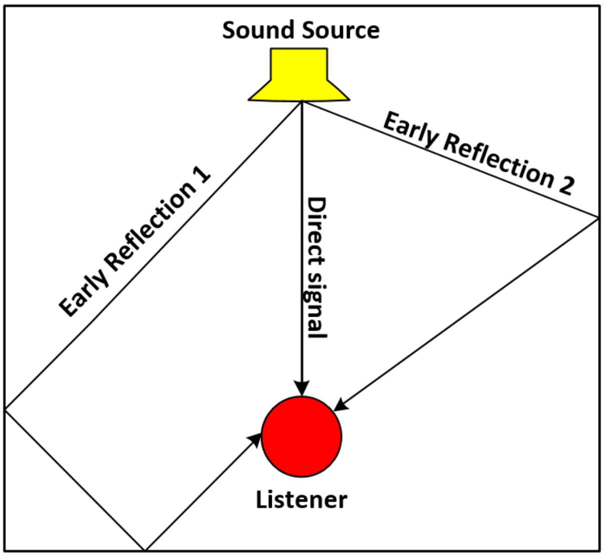


Figure 3.2. Early reverberation and the direct signal reach the microphone at the same time nearly.

In an enclosed space, the elements of $\boldsymbol{g}_d(k)$ show attenuation and phase shift between source and related microphone. It is obvious that $\boldsymbol{g}_d(k)$ vector cannot be estimated exactly. However, the relative transfer functions are sufficient to extract coefficients of MVDR. In fact, relative transfer function vector $\boldsymbol{g}'_d(k)$ is the normalized version of $\boldsymbol{g_d}(k)$ by $G_d^1(k)$ which is the transfer function of the earliest speech components at the reference microphone. The earliest components vector can be written in terms of $\boldsymbol{g}'_d(k)$:

$$\boldsymbol{g}_d(k)\, S(l,k) = \boldsymbol{g}'_d(k)\, G_d^1(k)\, S(l,k). \tag{3. 18}$$

Steering constraint of MVDR with respect to the earliest speech components is shown as:

$$\boldsymbol{w}(k)^H \boldsymbol{g}_d(k)\, S(l,k) - G_d^1(k)\, S(l,k) = 0. \tag{3. 19}$$

Substitute Eq. (3.18) in Eq. (3.19) gives:

$$\boldsymbol{w}(k)^H \boldsymbol{g}'_d(k) G_d^1(k)\, S(l,k) - G_d^1(k)\, S(l,k) = 0 \qquad (3.\,20)$$

$$\boldsymbol{w}(k)^H \boldsymbol{g}'_d(k) = 1.$$

Therefore, instead of finding $\boldsymbol{g}_d(k)$ vector exactly, $\boldsymbol{g}'_d(k)$ vector showing relative transfer functions is used to obtain directivity in this study.

The attenuations between source and each microphone can be accepted nearly equal for early reverberations. Therefore, the magnitudes of each transfer function are same. Relative phase differences are the results of path inequalities of the sound waves, and it depends on the direction of sound source. Therefore, the relative transfer function vector consists of only phase difference terms. $\boldsymbol{g}'_d(k)$ can be written as:

$$\boldsymbol{g}'_d(k) = \left[1 \;\; e^{-j\left(\frac{2\pi f_s k}{K}\right)\tau_1} \;\; e^{-j\left(\frac{2\pi f_s k}{K}\right)\tau_2} \;\; e^{-j\left(\frac{2\pi f_s k}{K}\right)\tau_3} \;\; \ldots\ldots \;\; e^{-j\left(\frac{2\pi f_s k}{K}\right)\tau_m}\right]^T \qquad (3.\,21)$$

where K is the FFT length and each exponential term represents phase difference of the corresponding microphone relative to the reference. When the filter is applied to the recorded input signals:

$$\boldsymbol{w}(k)^H \boldsymbol{x}(l,k) = \left\{\boldsymbol{w}(k)^H \boldsymbol{g'}_d(k) G_d^1(k) S(l,k) + \boldsymbol{W}(k)^H \boldsymbol{r}(l,k)\right\} + \boldsymbol{W}(k)^H \boldsymbol{D_l}(l,k). \qquad (3.\,22)$$

The last term represents lagged early reverberations. It is assumed that it is not related to the solution, so optimization is studied in the parenthesis in Eq. (3.22) and the Lagrangian can be written as

$$L(l,k) = \boldsymbol{w}^H(l,k) \boldsymbol{\phi}_r(l,k) \boldsymbol{w}(l,k) + \left[\lambda(\boldsymbol{w}^H(l,k) * \boldsymbol{g}'_d(k) - 1)\right] \qquad (3.\,23)$$

where k is the frequency bin number. The late reverberant part of the signal is assumed as independent of the earliest components due to diffuse sound field, so the standardized solution of beamforming Eq. (3.14) becomes valid for the reverberant signals in the parenthesis Eq. (3.22) and Lagrangian in Eq. (3.23). However, late reverberant correlation matrix ($\boldsymbol{\phi}_r$) and relative transfer functions ($\boldsymbol{g}'_d$) of the earliest components must be estimated for each subband to find the filter coefficients.

Relative transfer functions are derived from phase differences in this work. Thus, there is no need to estimate transfer functions and ratios of early reverberations blindly. However, the disturbances due to lagged early reverberations are neglected when the transfer functions are written in this way. Gannot used a single channel blind dereverberation method to generate $\boldsymbol{g}_d(k)$ vector more appropriately at the expense of computational complexity [28].

3.4.3. Estimating Phase Difference of the Earliest Components

In order to form the relative transfer function vector, $\boldsymbol{g}'_d(k)$ the phase difference of the earliest components at each microphone relative to the reference microphone must be known. The phase difference in classical MVDR algorithm can be found by DOA estimation.

However, in reverberant environments, there is no direct method to find DOA accurately because of the effects of multipath propagation. The sound waves arrive at the microphones from all the directions; therefore, estimating DOA is a very troublesome issue. At some instances, there can be strong reflections from a direction while at some other instances the reflection can be weak due to the statistical nature of reverberation. Because of this uncertain situation, DOA estimates in short frames change over time. High variability of estimates introduces a significant challenge. Besides, the reflected source signals may be added to the direct sound path with a delay. This situation makes the estimation of the source signal direction at each microphone much more difficult. As a result, reverberation has large destructive effects on DOA estimation [54].

The most commonly used method to find DOA is MUSIC algorithm [55]. This algorithm uses eigenvalues and eigenvectors of the correlation matrix of observed speech at microphone array. It generates the relation between eigenvectors and the desired signal or interference subspaces. However, the number of arrival directions of sound waves is larger than microphone number in reverberant environments. Therefore, classical MUSIC algorithm fails at reverberant environments. In this thesis, phase difference estimation is based on correlation of the microphone signals. In the estimation procedure, it is crucial to avoid the destructive effects of reverberation due to the reasons stated above.

The cross-correlation of sound signals at different microphones is useful to estimate relative phase difference. In microphone arrays, there is a deterministic time delay between microphone signals for the direct sounds. In anechoic environments correlation of the recorded signals between microphones gives time delay directly because the signal at a microphone is the phase shifted version of that at another microphone. Example of two microphone signals in an anechoic room is shown as in Figure 3.3. The difference between two recorded signals results from propagation delay and this quantity is obtained easily from the correlation graph as in Figure 3.4.

In reverberant environments, the cross-correlation graph is expanded across horizontal axis due to smearing effects. Additional correlations occur between the signals, so it is not possible to extract phase difference exactly by using the cross-correlation function. The cross-correlation function of reverberant speech signals is shown in Figure 3.5.

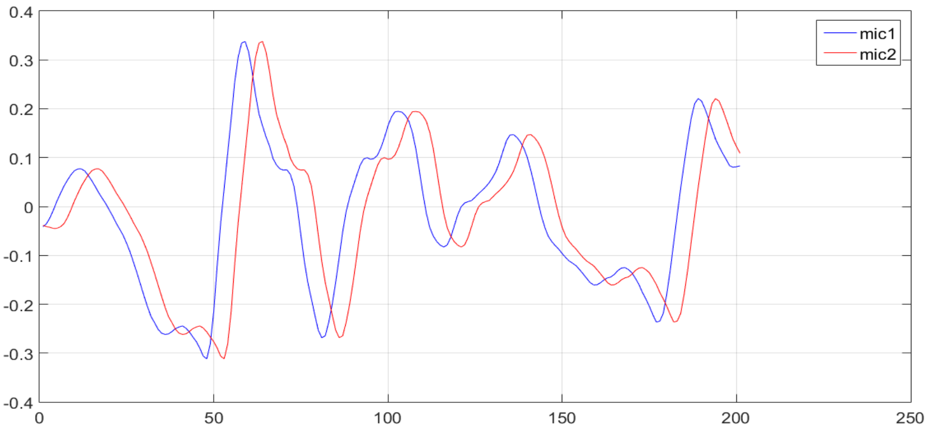


Figure 3.3. Two anechoic speech in a microphone array.

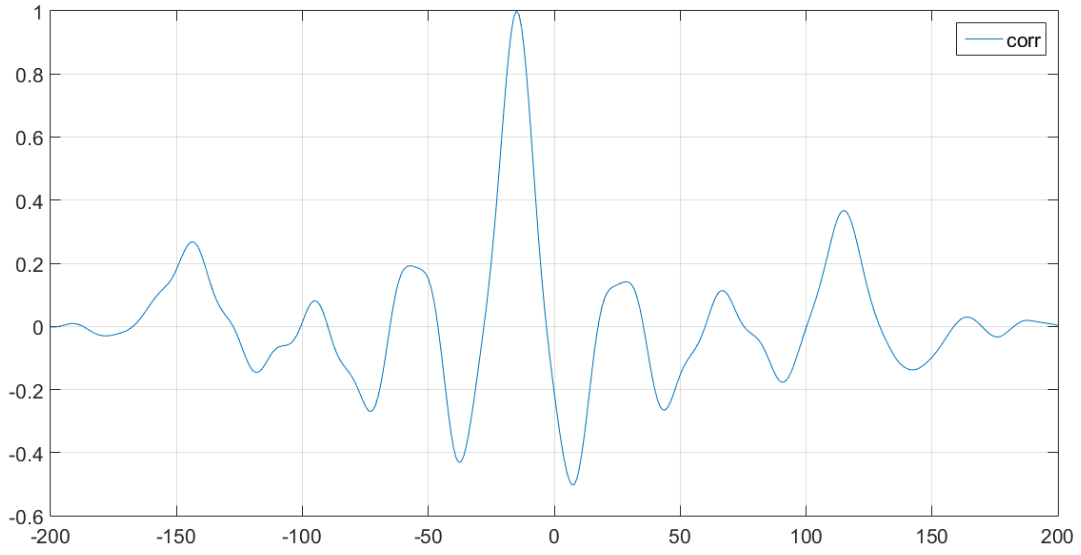


Figure 3.4. Correlation graph of anechoic speech.

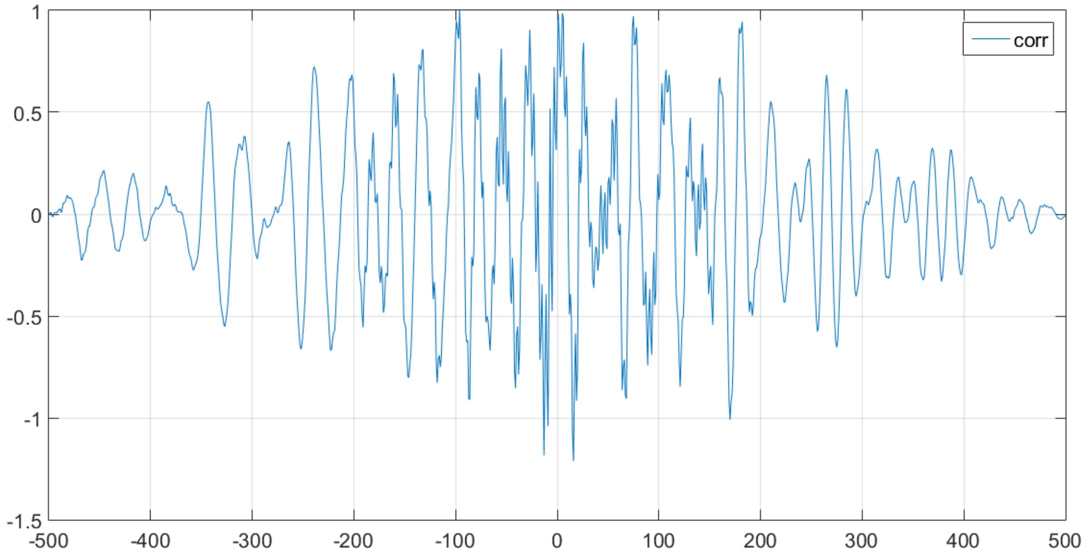


Figure 3.5. Correlation graph of reverberant speech.

The effects of late reverberation cannot be reduced without dereverberation of the observed speech. However, if the speech components which do not contain any late reverberation can be detected, the cross-correlation function of these parts can be used to estimate the phase difference.

In Chapter 2, it has been stated that late reverberation starts after a mixing time [6]. The average value of the mixing time is 10 – 15 msec in normal room conditions. Therefore, uncertain reverberant parts are added by a delay. If the speech onset is detected, the sound components without reverberation can be extracted and the phase difference can be estimated by using these frames properly.

The voice activity detection (VAD) algorithm is used to detect the onset of speech [7]. Two features are used in VAD. The first one is short time energy in a speech frame. Using this feature, speech can be detected when the energy of a frame is larger than a threshold. The other feature is the dominant frequency of the sound. It uses the magnitude of the most dominant frequency coefficient. These features for a typical speech signal are shown in Figure 3.6. If the values of these features are larger than their thresholds at the same frame, it is assumed that the related frame corresponds to speech, otherwise it corresponds to silence. In order to increase robustness of the system, these conditions must be satisfied successively at least few frames before making a decision as silence or speech.

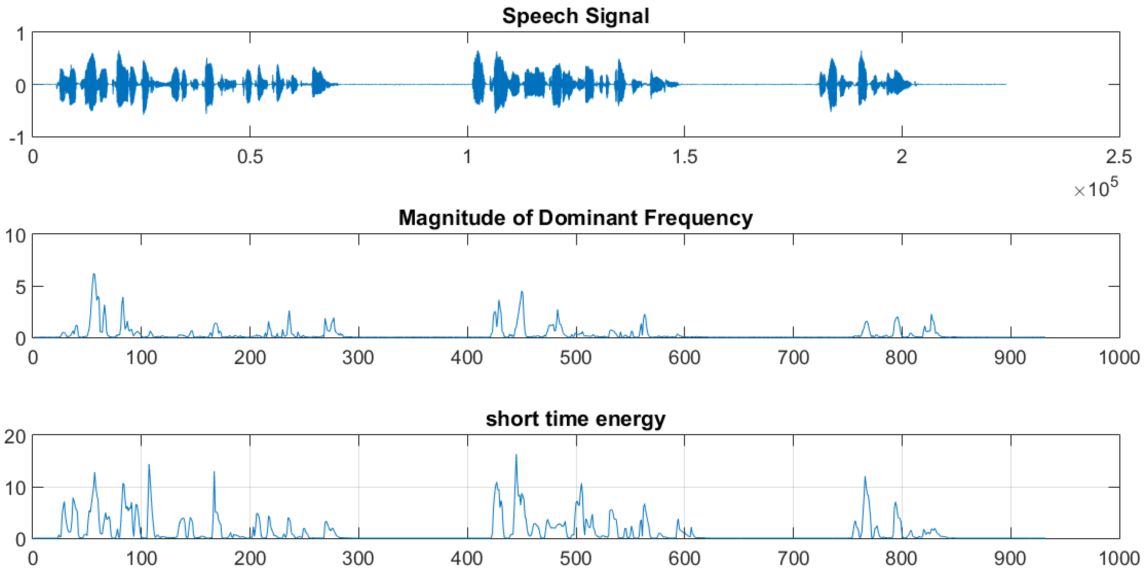


Figure 3.6. The algorithm decides whether the relative segment is silence or speech by magnitude of the dominant frequency and short time energy.

Initially, the transitions from silence to speech are detected. Then, the frames which are recorded during beginning of the speech by two different microphones are chosen to compute the cross-correlation function. The length of chosen signals must be less than the mixing time to avoid the effects of late reverberation. The cross-correlation function of reverberant speech signals at a speech onset frame is shown in Figure 3.7. The correlation function in Figure 3.7 is calculated for the same speech recording in Figure 3.5; however, since effects of late reverberation are less at speech onsets denser correlation graph is obtained. The phase difference can be seen at this graph easily. Furthermore, the average of the phase differences for all speech onsets in a recording can be taken in order to achieve more accurate estimation.

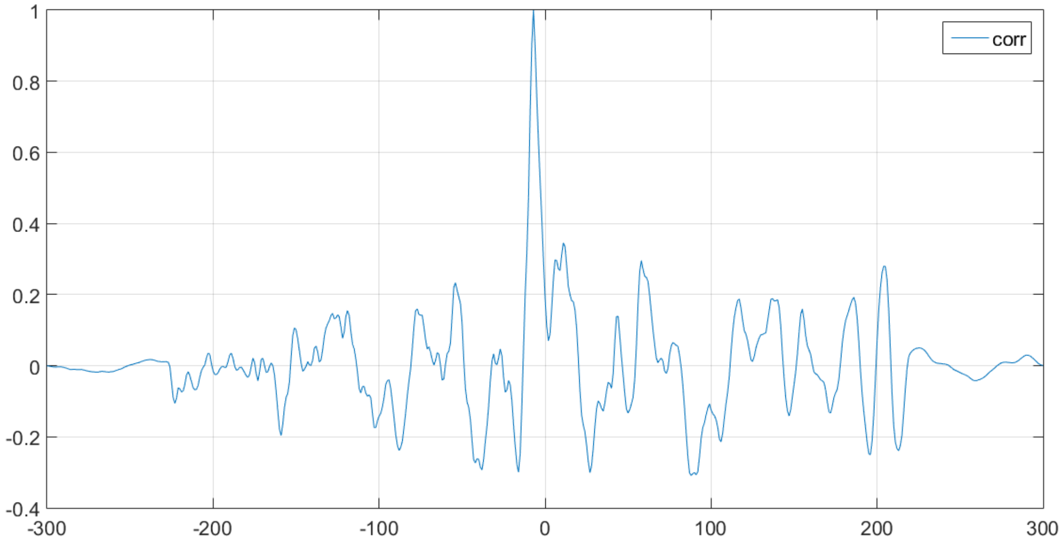


Figure 3.7. Correlation graph of the reverberant signals in Figure 3.5. However, correlation is taken from the first 15 msec speech onset.

3.4.4. Correlation Matrix of Reverberant Sound Field

Beamforming method needs the correlation matrix of interference signals according to Eq. (3.14). Therefore, it is necessary to express correlations of late reverberant part of speech mathematically in order to adjust MVDR filter.

Before considering reverberant conditions, it is useful to evaluate single sound wave situation for two microphones. This gives a better understanding about the correlation of reverberations in microphone arrays. Single plane wave propagation through the microphone array is shown as in Figure 3.8.

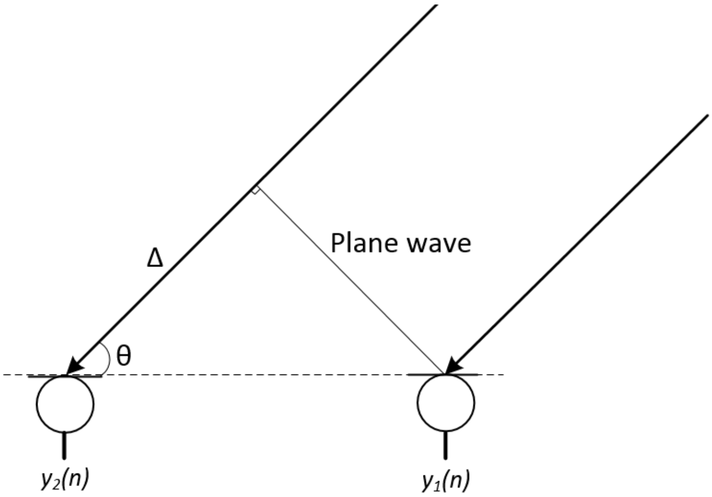


Figure 3.8. Plane wave propagation towards a microphone array.

The relation between the microphone signals in time domain can be written as:

$$x_2[n] = x_1\left[n - \frac{\Delta}{c}\right]. \tag{3. 24}$$

Δ is path length difference in m, c is sound wave propagation velocity in m/sec. In diffuse fields it is assumed that the PSD is distributed uniformly and independent of position. Therefore, PSDs of two microphone signals are assumed to be the same:

$$S_{x1}(k) = S_{x2}(k) \tag{3. 25}$$

where k is the frequency index. The cross-power density is defined as,

$$S_{x1x2}(k) = S_{x1}(k)e^{\frac{-j\left(\frac{2\pi f_s k}{K}\right)dcos\emptyset}{c}} \tag{3. 26}$$

where K is the FFT length. The correlation coefficient for a single plane wave can be written as,

$$\propto = \frac{S_{x1x2}(k)}{\sqrt{S_{x1}(k)S_{x2}(k)}} = e^{\frac{-j\left(\frac{2\pi f_s k}{K}\right)dcos\emptyset}{c}}. \qquad (3.\ 27)$$

Late reverberation is assumed to start with approximately 15 msec delay after speech onsets. Therefore, late reverberant sound waves travel an additional distance of at least 5 m compared to the waves following the direct path. This distance is enough for far field assumption in the majority of the speech frequency content. Therefore, sound waves can be considered as plane waves in late reverberation, and Eq. (3.27) is valid for the rest of the derivation.

Since late reverberation can be modeled by a diffuse field, the characteristic properties of the field are useful for correlation derivations. In diffuse field assumption, it is equally probable for sound energy to flow in each direction. Also, the sound waves are distributed uniformly in the room, so energy density of the sound is equal at every position in the room [17]. As a result, infinite number of identical sound sources which are distributed homogenously on the surface of a sphere can represent diffuse sound field as in Figure 3.9 [56]. Cross correlation of the microphone signals can be described by integration of the contributions of the sound sources.

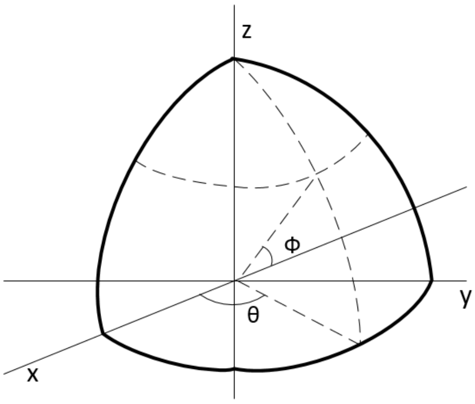


Figure 3.9. In diffuse sound field, it is assumed that each infinitesimal area on the surface of sphere is an identical sound source.

Assume that the microphones are located on x axis. Each infinitesimal area represents identical plane wave sound source. Cross-correlation of single plane wave was shown in Eq. (3.27). Integrating the cross-correlation over the surface,

$$\gamma = \frac{1}{4\pi r^2}\int_0^{2\pi}\int_0^{\pi} e^{-\frac{j\left(\frac{2\pi f_s k}{K}\right) d cos\phi}{c}} r^2 sin\phi \; d\phi \; d\theta \tag{3. 28}$$

$$= \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} e^{-\frac{j\left(\frac{2\pi f_s k}{K}\right) d cos\phi}{c}} sin\phi \; d\phi \; d\theta \, , \qquad substitute \; u = cos\phi$$

$$= \frac{\sin\left(\frac{\left(\frac{2\pi f_s k}{K}\right) d}{c}\right)}{\frac{\left(\frac{2\pi f_s k}{K}\right) d}{c}}$$

where d is the distance between related microphones in m, K is the FFT length. The result is used to form the correlation matrix of late reverberation signals between two microphones. The correlation matrix of late reverberation for two microphones is written as:

$$\boldsymbol{\phi}_r(l,k) = \phi_r(l,k)\begin{bmatrix} 1 & \frac{\sin\left(\frac{\left(\frac{2\pi f_s k}{K}\right) d_{2,1}}{c}\right)}{\frac{\left(\frac{2\pi f_s k}{K}\right) d_{2,1}}{c}} \\ \frac{\sin\left(\frac{\left(\frac{2\pi f_s k}{K}\right) d_{2,1}}{c}\right)}{\frac{\left(\frac{2\pi f_s k}{K}\right) d_{2,1}}{c}} & 1 \end{bmatrix} \tag{3. 29}$$

where $\phi_r(l,k)$ is the PSD of late reverberation. The larger correlation matrices can be written in the same manner.

After cross-correlations of late reverberant signals are estimated, LRSV i.e. PSD of late reverberation, $\phi_r(l,k)$ must be estimated as well.

3.4.5. Estimation of Late Reverberation PSD

Reverberation depends on the acoustic characteristics of room. Therefore, reverberant part of speech can be modelled with respect to the room acoustics. In Chapter 2, the statistical time domain reverberation model is derived by taking room acoustics into account.

The exponential decaying stochastic process is used to describe late reverberation [6]. However, this stochastic model becomes valid after a time interval. Therefore, the characteristics of the entire impulse response could be evaluated in two regions. The first part of the response corresponds to the direct sound signals and a few deterministic reflections. The second part of the impulse response involves the last part of the early reverberation and late reverberation. This part can be represented by a stochastic process complemented with exponential decay.

The causal RIR can be described as:

$$h[n] = \begin{cases} h_e[n] & where\ 0 \leq n \leq T \\ h_r[n] & where\ n \geq T \\ 0 & where\ n < 0 \end{cases} \tag{3. 30}$$

where $h_e[n]$ represents direct and early reverberant parts of RIR, $h_r[n]$ represents the rest of the reverberant part, T is related to mixing time. Time domain Polack model can represent the second part, so $h_r[n] = a[n]e^{-\alpha n}$ where $a[n]$ is zero mean Gaussian process, α is a constant related to the room acoustics.

In frequency domain, the second part of the transfer function can also be represented by a Gaussian process due to overlapping of the room modes, so transfer function is defined as:

$$H(k,l) = \begin{cases} B_d(k) & where\ l = 0 \\ B_r(l,k)\ e^{-\alpha(k)lR} & where\ l \geq 1 \end{cases} \tag{3. 31}$$

where R is speech frame length, $B_r(l,k)$ is zero mean Gaussian process, α is decaying constant [57]. Assume that there is no dependence between the coefficients of different frequency bands. STFT coefficients from different frames are also assumed to be independent. These assumptions are written as:

$$E\{H(l,k_1)H^*(l,k_2)\} = 0 \text{ for } k_1 \neq k_2, \forall l \tag{3. 32}$$

$$E\{H(l_1,k)H^*(l_2,k)\} = 0 \text{ for } l_1 \neq l_2, \forall k.$$

In addition, anechoic speech signal coefficients (zero mean, Gaussian) from different frames are independent.

Observed reverberant signal can be written as:

$$X(l,k) = D(l,k) + R(l,k) \tag{3. 33}$$

where $D(l,k)$ denotes early reverberation components, $R(l,k)$ represents the following reverberant signals. They are independent of each other, so PSD of the observed speech is:

$$\phi_X(l,k) = \phi_D(l,k) + \phi_R(l,k) \tag{3. 34}$$

where ϕ_D, ϕ_R denotes PSDs of early components and the following reverberant components. From another perspective, observed speech can be expressed by the CTF model, so the PSD can also be obtained by the square of CTF model [57]:

$$X(l,k) = \sum_{n=0}^{\infty} S(l-n,k)H(n,k)$$

$$\phi_X(l,k) = \left(\sum_{n=0}^{\infty} S(l-n,k)H(n,k)\right)^2. \tag{3. 35}$$

Variance of the STFT coefficients of convolutive transfer function in Eq. (3.31) is used to express the observed speech's PSD. It can be described as:

$$\phi_H(l,k) = \begin{cases} \beta_D(l,k) & l = 0 \\ \beta_R(l,k)e^{-2\alpha(k)Rl} & l \geq 1 \end{cases} \tag{3. 36}$$

where β_D and β_R denote the variances of the early components and the following reverberant parts in the transfer function, respectively. PSD of the observed speech can be written in terms of CTF:

$$\phi_{\mathrm{X}}(l,k) = \sum_{n=0}^{\infty} \phi_S(l-n,k)\phi_H(n,k) \tag{3. 37}$$

PSD of the observed speech can be divided into two terms due to the structure of RIR. By Eq. (3.36) and Eq. (3.37):

$$\phi_D(l,k) = \beta_D(k)\phi_s(l,k) \tag{3. 38}$$

$$\phi_R(l,k) = \sum_{m=1}^{\infty} \beta_R(k)e^{-2\alpha(k)Rm}\phi_s(l-m,k) \tag{3. 39}$$

$$= e^{-2\alpha(k)R}[\phi_R(l-1,k) + \beta_R(k)\phi_s(l-1,k)]\,.$$

PSD of reverberation is given by Eq. (3.39). This result corresponds to the sum of the late reverberation and lagged early reverberation PSDs. The first term in parenthesis can be expressed in terms of previous samples iteratively, so it can be considered as PSD of late reverberation:

$$\phi_l(l,k) = e^{-2\alpha(k)R}\,\phi_R(l-1,k). \tag{3. 40}$$

If $\beta_d(k) = \beta_R(k)$, Eq. (3.40) gives the PSD of late reverberant as:

$$\phi_l(l,k) = e^{-2\alpha(k)RN}\,\phi_X(l-N,k). \tag{3. 41}$$

Spectral power of late reverberation is calculated for each microphone in the microphone array. Since each microphone can sample stochastic reverberation process at different positions, the more robust result can be obtained by averaging the spectral powers [58]:

$$\phi_l(l,k) = \frac{1}{M}\sum_{m} \phi_l^m(l,k). \tag{3. 42}$$

As a result, PSD of late reverberant part can be estimated by the recorded speech signals and exponential decaying reverberation modeling. There is only one unknown in the equation; the exponential decaying term.

Exponential decaying term is a time constant related to reverberation time of room. Mathematical expression of the relation between the time constant '$\alpha(k)'$ and reverberation time is useful for calculations. Reverberation time is described as a time interval for the initial level of the sound signal to decay 60 dB. For a narrowband signal $\alpha(k) = \alpha$, T_r is reverberation time, the exponential decaying is defined as:

$$e^{-2\alpha T_r f_s} = \frac{1}{10^6}. \tag{3. 43}$$

Take the logarithm of each side and extract the time constant:

$$\alpha = \frac{3ln10}{T_r f_s} \ . \tag{3. 44}$$

The relation of reverberation time and exponential time constant is described as in Eq. (3.44). If reverberation time is known, PSD of late reverberation part of the speech can be calculated.

3.4.6. Reverberation Time Estimation

Reverberation time estimation is very common in the literature since it is a significant property of the room acoustics. Normally, reverberation time is measured by tone burst response of room. Tone burst is a short time excited sound which generates a decay curve for the response. Typical tone burst response is shown in Figure 3.10:

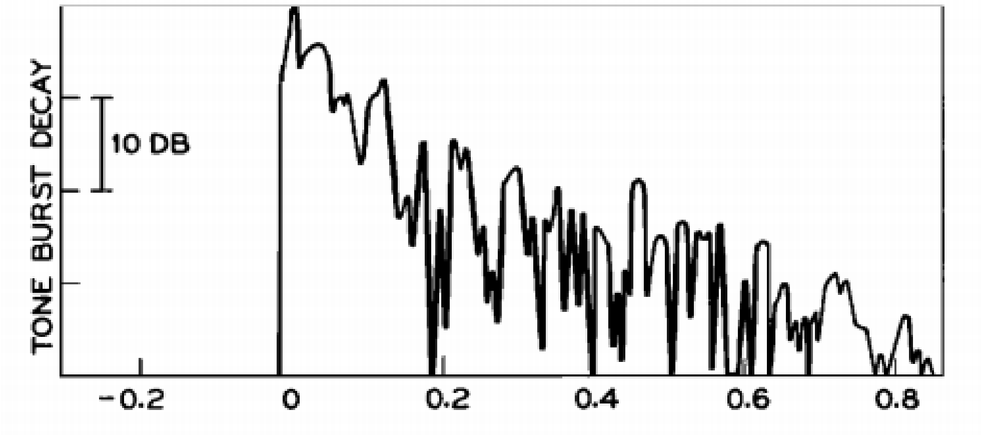


Figure 3.10. Tone burst response of a sound. Reverberation time is estimated by the level of sound [18].

Decay curve of a room has a lot of fluctuations due to the statistical characteristics of reverberation. Therefore, reverberation time measurement by a single graph is not reliable due to non-deterministic behavior of the response. In order to minimize the effects of randomness, the estimation experiment should be repeated as much as possible.

In a practical dereverberation task, it is not possible to estimate the reverberation time with an excitation signal. Reverberation time must be estimated by the recorded speech signals blindly. Therefore, an appropriate approach should be developed for the estimation.

The statistical model of reverberation can be used to estimate reverberation time blindly. For this purpose, Polack time domain reverberation model is used [6]. Speech signals can be divided into two different components:

$$x[n] = d[n] + r[n]. \tag{3. 45}$$

The decaying curve begins when the speech source signal $d[n]$ stop abruptly, and the decaying part of the speech signal equals late reverberant components. This part can be modeled as a stochastic process:

$$r[n] = A_r v[n] e^{\frac{-\alpha n}{f_s}} \tag{3. 46}$$

where $v[n]$ is normal random process, A_r^2 relates to the variance of the process. According to this model, probability distribution of the decaying signals can be written as:

$$p_{r[n]}(x) = \frac{1}{\sqrt{2\pi}\sigma(n)} \exp\left(-\frac{x^2}{2\sigma^2(n)}\right)^{\frac{1}{2}} \tag{3. 47}$$

$$\sigma(n, \alpha) = A_r e^{\frac{-\alpha n}{f_s}}$$

where $\sigma^2(n, \alpha)$ is the time varying variance of late reverberation. If reverberation samples are represented by independent Gaussian process, joint probability of the decaying part can be found as the product of individual pdfs of the samples. Then, the reverberation time constant (α) can be found by maximizing the likelihood function:

$$\alpha^{ML} = \max_{\alpha}\big(L(\alpha)\big) \tag{3. 48}$$

where $L(\alpha)$ is the likelihood function of the decaying part depending on the reverberation time constant. It is written as:

$$L(\alpha) = \prod_n \frac{1}{\sqrt{2\pi}\sigma(n, \alpha)} \exp\left(-\frac{x^2}{2\sigma^2(n, \alpha)}\right). \tag{3. 49}$$

Although, Eq. (3.49) is valid for the decaying part of observed speech, the algorithm cannot distinguish the decaying part of speech blindly. Therefore, E. Yılmaz proposed an approach to detect the decaying parts of recorded speech [59]. Before maximization of the likelihood function, E. Yılmaz proposes to divide recorded speech into segments, and then look for the decaying part in the segments. Some constraints are proposed to detect these decays. If all the constraints are satisfied in a segment, the segment is declared to contain a decaying part. Then, the reverberation time constant is estimated by maximum likelihood estimation in that segment. If recorded speech has sufficient length, it is possible to observe a lot of decaying speech segments. Thus, the fluctuation effects of the probabilistic signals can be reduced by taking average of the estimations.

The speech signals are divided into segments and then sub-segments. The segments can be written as,

$$x_s[\lambda, m] = x[\lambda M_\Delta + m] \qquad (3.\,50)$$

where M_Δ shows the segment length, m denotes individual sample in the segments, λ is the segment number. Then, these segments are divided into subsegments. The constraints are evaluated by these subsegments. They can be written as,

$$x_{ss}[\lambda, l, k] = x_S[\lambda, lP + k] \qquad (3.\,51)$$

where k is the sample index inside subsegments, P is the length of the subsegments, l is subsegment number.

There are three constraints to decide whether there is a decaying pattern in the related segment or not:

$$\sum_{k=0}^{P-1} x_{ss}^2[\lambda, l, k] > \sum_{k=0}^{P-1} x_{ss}^2[\lambda, l+1, k] \qquad (3.\,52)$$

$$\max_k(x_{ss}[(\lambda, l, k)]) > \max_k(x_{ss}[\lambda, l+1, k])$$

$$\min_k(x_{ss}[\lambda, l, k]) < \min_k(x_{ss}[\lambda, l+1, k]).$$

In fact, all of these constraints are based on the fact that the time varying variance of speech signal reduces at decaying parts.

As it can be seen in Figure 3.11, these constraints are satisfied at the decaying parts of the speech signal. Decaying parts lead variance of speech signals to reduce, so first constraint is satisfied. Also, drop in the variance means that magnitude of the samples tends to get closer to zero. The second and third constraint is also satisfied in this situation.

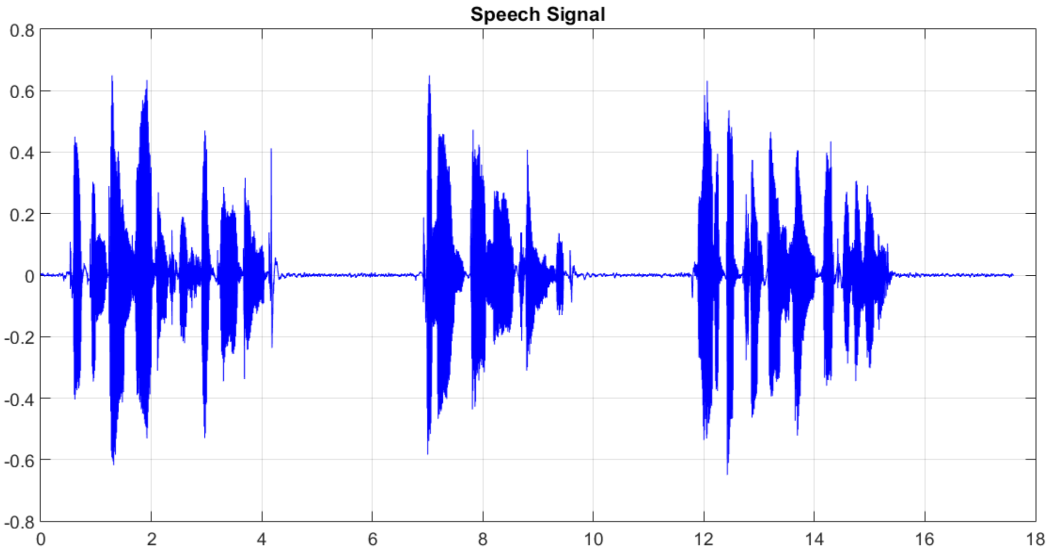


Figure 3.11. The variance of decaying parts in a typical speech reduces to zero gradually.

Reverberation time depends on frequency content of the signal [60]. It is longer at lower frequencies, and reverberation time should be estimated by narrow band signals for precise result. Therefore, bandpass filters can be used before the estimation procedure. However, this process brings a serious computational complexity. Since reverberation time is estimated by all decaying parts of speech in time domain, this estimation is acceptable for majority of the speech frequencies. Although reverberation time is not estimated in all subbands in this thesis, the effect of frequency is taken consideration roughly in the experiments part.

3.4.7. MVDR Solution for Dereverberation

In Chapter 3, MVDR algorithm has been presented. Initially, standard MVDR equations have been derived in section (3.4.1). Later, MVDR method has been adapted to reverberant environment by statistical model of reverberation in section (3.4.2). For this purpose, the observations are divided into early reverberation and late reverberation. Late reverberation is assumed as independent interference signals, and the recorded reverberant signals are written similar to general noisy observations of MVDR. In this way, the solution of MVDR algorithm for reverberation is generated in similar way of the standard MVDR solution.

In section (3.4.3) the transfer function of early reverberation is generated by estimating phase differences. Also, correlation matrix of late reverberation is required for solution of MVDR. The statistical room modeling is used to estimate correlation matrix and PSD of late reverberation in sections (3.4.4) & (3.4.5).

In the dereverberation, late reverberation is treated as independent interferences. The early reverberation is considered as the desired signal and steering direction of the microphone array is adjusted to these components, thus MVDR algorithm passes the early speech components while suppressing late reverberation.

The summary of the proposed MVDR algorithm is shown in Figure 3.12.

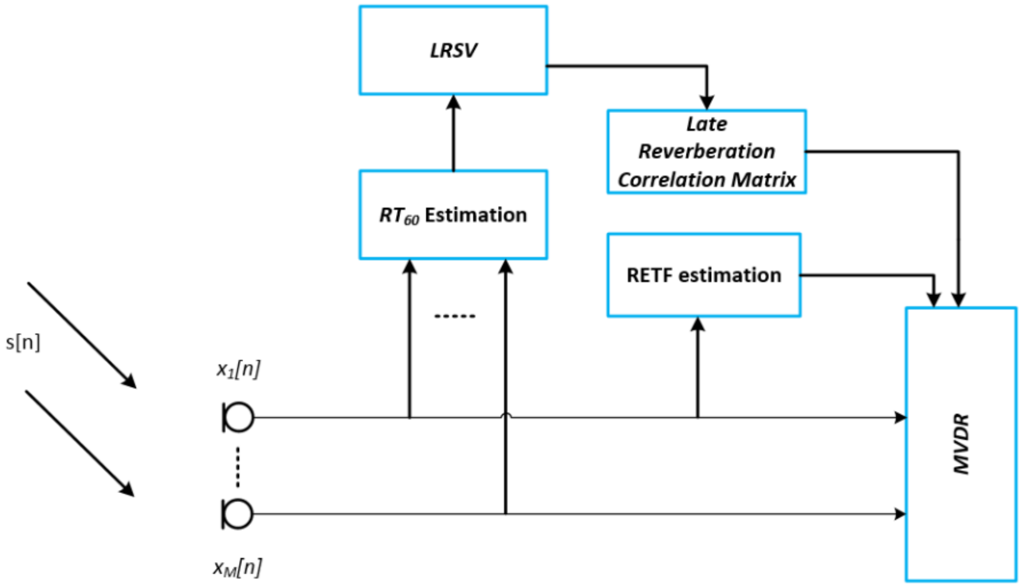


Figure 3.12. The graph of the first stage of the dereverberation algorithm.

3.5. Single Channel Dereverberation

In this thesis, a single channel blind dereverberation algorithm is used after MVDR processing to remove residual reverberant components. In this part, the single channel approach is studied separately. The combined system is presented at the end of the chapter.

Single channel dereverberation is a blind deconvolution process. There is no deterministic prior knowledge about RIR and source signals. Therefore, either the convolution or source signal parameters have to be known. However, it is impossible to estimate these parameters exactly in real time processing by a single microphone, the statistical models of reverberation and anechoic speech signals are very useful in single channel approaches. Similar statistical models have been used at MVDR based dereverberation in the first stage.

In this method, reverberation is considered as a temporal smearing of speech signals. Therefore, each sample affects the subsequent samples in reverberant environments.

The observed signal, $\bar{x}[n]$ is modelled as:

$$\bar{x}[n] = \bar{d}[n] + \bar{r}[n] \tag{3. 53}$$

where $\bar{d}[n]$ is the early reverberation at the output of MVDR, $\bar{r}[n]$ is the residual late reverberant part. Since late reverberant part can be considered as a replica of the previous samples in this algorithm, this part can be written as a function of the previous samples of observed speech signals:

$$\bar{r}[n] = f(\bar{\boldsymbol{x}}[n-T]) \tag{3. 54}$$

where $\bar{\boldsymbol{x}}[n-T] = \left[\bar{x}[n-T]\; \bar{x}[n-T-1]\; \bar{x}[n-T-2] \ldots.. \; \bar{x}[n-T-L+1]\right]^T$ is the vector of past samples.

Eq. (3.54) can be verified by Polack time domain model. The delay constant,'T' represents the discrete mixing time in Polack reverberation model. Late reverberant part of an observed signal $r[n]$is described as:

$$r[n] = \sum_{\tau=T}^{L} h[\tau]\, s[n-\tau] \tag{3. 55}$$

A previous sample can be written corresponding to mixing time:

$$x[n-T] = \sum_{\tau=0}^{T} h[\tau]\, s[n-T-\tau] + \sum_{\tau=T}^{L} h[\tau]\, s[n-T-\tau] \tag{3. 56}$$

According to Eq. (3.55) and Eq. (3.56), there is an obvious relation between late reverberant part of the recorded speech and the previous samples. Therefore, it is possible to describe residual late reverberation by a weighted sum of the previous samples at the output of MVDR which are at least mixing time before the current sample:

$$\bar{r}[n] = \sum_{\tau=T}^{L} \mathrm{c}_{\mathrm{k}}^{*}\, \bar{x}[n-\tau]. \tag{3. 57}$$

In frequency domain, the similar relation in Eq. (3.55) and Eq. (3.56) can be generated by using CTF model. The STFT coefficients of recorded speech are written as:

$$X(l,k) = \sum_{n=0}^{\infty} H(n,k)\, S(l-n,k). \tag{3. 58}$$

According to Eq. (3.7), Eq. (3.8) and Eq. (3.9), the coefficients of late reverberation in a frame can be explained by the coefficients of previous frames.

Presented dereverberation approach can be implemented either in frequency domain or in time domain. In time domain since speech signal is processed sample by sample, there are much more parameters than that of frequency domain. Therefore, frequency domain is more reasonable due to less computational complexity. Also, it is easier to combine different dereverberation approaches in frequency domain. This is a very tempting feature because two different dereverberation approaches are combined in this study. Therefore, frequency domain analysis is preferred in the rest of the work.

Late reverberation coefficients at MVDR output can be written as a function of the previous frames as in Eq. (3.57):

$$\overline{R}(l,k) = f\left(\overline{\boldsymbol{X}}(l-D,k)\right). \tag{3. 59}$$

The function is written as:

$$\overline{R}(\mathrm{l},\mathrm{k}) = \sum_{r=D}^{L} \mathrm{c}_{\mathrm{r}}^{*}\, \overline{X}(l-r,k) \tag{3. 60}$$

where $\overline{\boldsymbol{X}}(l-D,k) = \left[\overline{X}(l-D,k)\; \overline{X}(l-D-1,k) \ldots \overline{X}(l-L,k)\;\right]$, L is the filter length in terms of frame number, D is mixing time in the same manner. STFT coefficients of the early reverberation part can be written as,

$$\overline{D}(l,k) = \overline{X}(l,k) - \sum_{r=D}^{L} \mathrm{c}_{\mathrm{r}}^{*}\overline{X}(l-r,k). \tag{3. 61}$$

In matrix form,

$$\overline{D}(l,k) = \overline{X}(l,k) - \boldsymbol{c}^{H}\,\overline{\boldsymbol{X}}(l-D,k) \tag{3. 62}$$

where $\boldsymbol{c} = [c_1\; c_2 \;\ldots..\; c_{L-D+1}]^T$.

As it can be seen in Eq. (3.62), the dereverberation method is simply to estimate of filter coefficients (c_k^*). In order to estimate the filter coefficients, either prior knowledge about the coefficients or speech signal structure has to be known. Statistical modeling of speech is useful in this situation. Statistical distributions of coefficients can be used as a prior knowledge about the early reverberant part.

3.5.1. Statistical Distribution of Speech Signals

In Chapter 2, distributions of speech samples have been studied. Speech signal in a frame can be represented by an exponential distribution. Laplacian, Gamma or Gaussian distributions are used to represent STFT coefficients. Hendriks states that STFT coefficients can be represented by Gaussian distribution in approximately 30 msec frames [51]. Apart from the accuracy, Gaussian distribution is very useful with respect to mathematical tractability. Therefore, in this work, Gaussian distribution is used to represent STFT coefficients of early reverberation.

Speech is a nonstationary signal. Therefore, a single Gaussian distribution cannot represent frequency coefficients of different speech frames. Time varying Gaussian source model can be used to represent speech signals in frequency domain [8]. According to time varying Gaussian models, the STFT coefficient distributions in a frame is Gaussian; however, the variance varies through frames. The probability density function of the frequency coefficients of speech signal in the l^{th} frame for the k^{th} frequency bin can be written as:

$$p(\overline{D}_{l,k}) = N(\overline{D}_{l,k}; 0, \sigma_{l,k}^2) \qquad (3.63)$$

where $N(.)$shows zero mean Gaussian distribution, $\sigma_{l,k}^2$ denotes the time varying PSD in l^{th} frame, k^{th} frequency bin. In this method, STFT coefficients of early reverberation are represented by zero mean Gaussian distributions. The variance changes in accordance with temporal variations.

3.5.2. Estimation of Early Reverberant Part

Dereverberation method is reduced to estimation of the filter coefficients in Eq. (3.62) by the probability distributions of early reverberation, i.e. Gaussian distributions in Eq. (3.63). The filter coefficients can be estimated by maximizing log likelihood function for each subband.

In this estimation procedure, subbands of the speech signals are assumed independent. For each subband, consecutive STFT coefficients are also assumed independent. As a result, log likelihood function of the dereverberation process can be described as:

$$L(\theta_k) = \sum_l logp\Big(\overline{D}(l,k) = \overline{X}(l,k) - \boldsymbol{c}^H(k)\overline{\boldsymbol{X}}(l-D,k)\Big) \qquad (3.64)$$

$$= -\sum_l \frac{\left|\overline{X}(l,k) - \boldsymbol{c}^H\overline{\boldsymbol{X}}(l-D,k)\right|^2}{\sigma_{l,k}^2} - \sum_l log\,\sigma_{l,k}^2$$

where $\theta_k = \{\boldsymbol{c}_k^H, \sigma_{l,k}^2\}$. As in Eq. (3.64), the log likelihood function depends on the filter coefficients and spectral power of early speech components in related frame. In this chapter, PSD of late reverberant part has been estimated blindly in the microphone array. If late reverberant part is assumed independent of early speech components, spectral power of early reverberation can be found directly by subtraction.

Spectral power of late reverberant part at the output of the MVDR is given:

$$\sigma_{l,k}^{LRSV} = \phi_{l,k}\boldsymbol{w}^H(l,k)\Gamma(l,k)\boldsymbol{w}(l,k) \qquad (3.65)$$

where $\boldsymbol{w}(l\,,k)$ is the MVDR filter, $\Gamma(l,k)$ is the normalized correlation vector of late reverberant components $(\boldsymbol{\phi}_r(l,k))$ in room. Spectral power of early speech components at the MVDR output is written as:

$$\sigma_{l,k}^2 = \sigma_{l,k}^{\overline{X}} - \sigma_{l,k}^{LRSV}. \tag{3. 66}$$

The second term in the log likelihood function is related to PSD of early reverberation, so this term can be accepted constant. In fact, maximization of the function depends on the filter coefficients, i.e. first term in the likelihood. The first term of the log likelihood function can be seen as mean square error function which is normalized by the variance in linear prediction method. Therefore, the filter coefficients can be found by adapting the linear prediction solution where the solution is multiplication of inverse of the correlation matrix and the correlation vector. Solution of the log likelihood can be found as:

$$\boldsymbol{c}_k = \left(\sum_l \frac{\overline{\boldsymbol{X}}(l-D,k)\overline{\boldsymbol{X}}(l-D,k)^H}{\sigma_{l,k}^2}\right)^{-1} \left(\sum_l \frac{\overline{\boldsymbol{X}}(l-D,k)\overline{X}^*(l,k)}{\sigma_{l,k}^2}\right). \tag{3. 67}$$

In this method, subband decompositions are used in the processing. Likelihood estimation finds the filter coefficients for each frequency band. Similar method was proposed by Nakatani [8]. The approach is originally named ‘weighted prediction estimation’ (WPE) method. In WPE method, there is no estimate of the PSD of the desired speech. Instead, the method reaches solution iteratively. However, in the proposed approach the PSD of early reverberant is estimated by the statistical model of reverberation. In this way, there is no need of iterative calculations, so processing time of the method reduces drastically.

3.6. Entire Dereverberation Algorithm

In this chapter, MVDR algorithm for noisy environments has been derived and adapted to reverberant environments by using the statistical models of reverberation. The necessary parameters of reverberation signals have been estimated separately, then they have been used to generate appropriate MVDR algorithm in reverberant environments. In this work, it is expected that generated MVDR algorithm suppresses late reverberation; however, when some of the late reverberation reaches microphones in the same direction of the direct sound, MVDR algorithm cannot deal with these components. Therefore, using a post filter to suppress residual reverberant signal is required.

An appropriate single channel blind deconvolution algorithm is used as a post filter. In the first stage of the system, MVDR algorithm works as a spatial filter utilizing both room acoustics and the statistical models of reverberation. Therefore, there are some reverberation parameters which are already estimated in this stage. Choosing a single channel method which uses these parameters makes system more compact. In this way, computational complexity and processing time is reduced.

At the second stage, linear prediction and statistical nature of direct speech is combined. In this algorithm, ML estimation is used to estimate early reverberation in the observed speech signal. Solution of the likelihood function requires PSD of the early reverberation; therefore, reverberation time and LRSV have to be estimated initially. However, since they have been already estimated at the first part, there is no need of these extra computations. Although the second stage is completely different approach from the first one, there are lots of common parameters between them. In this way, two different aspects of reverberation are combined to filter reverberant signals without extra computation complexity.

Another advantage of the combined system is that LRSV estimation is made by multi microphones. According to the statistical properties of reverberation, energy density of sounds is same for every position in a room. However, since it can be represented by a random process, instant value of reverberation may not represent the process accurately. Therefore, estimation of LRSV at various positions gives different realizations of the same random process. In this way, estimation of LRSV is improved. When microphone numbers of the array increases, more reliable estimation is also possible. LRSV estimation affects directly the success of the second dereverberation method. Therefore, using MVDR before linear prediction algorithm improved its dereverberation performance.

Entire dereverberation algorithm is shown in Figure 3.13:

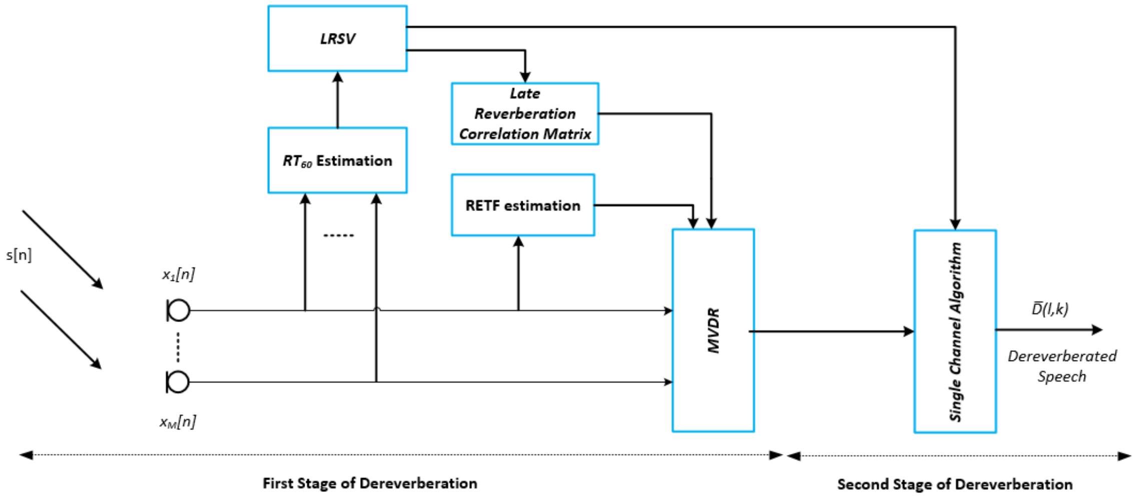


Figure 3.13. The graph of the entire dereverberation algorithm.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1. Introduction

The dereverberation approach through combined methods of microphone array and linear prediction has been proposed in the previous chapters. The theory of the system was given in Chapter 3 and it is applied to reverberant speech in this chapter. This chapter describes experimental study of the presented method.

In order to analyze the system, artificially reverberated dataset has been produced. In this way, it is easy to generate speech signals in various conditions, and dereverberation method is evaluated under these conditions.

Although there are lots of methods to evaluate general performance of speech processing algorithms, it is not a simple task to assess performance of dereverberation system. In this chapter, evaluation tools for reverberant environments will be explained. Then, the most accurate evaluation methods will be used to interpret the performance of the dereverberation.

Standard speech processing procedures such as segmentation and windowing are applied to recorded speech signal in the dereverberation. Therefore, these concepts are also explained in the experimental part.

4.2. Speech Quality Measures

There are two types of speech quality measures: subjective and objective methods. Subjective quality measure is a comparison of original and processed speech by listeners with respect to pre-determined scales. Listeners rank the quality of speech subjectively. Differences due to subjective evaluations can be reduced by averaging results from multiple listeners. However, this procedure is time-consuming and expensive. Therefore, objective quality measures are more reasonable to assess speech quality.

Objective quality measures are based on a particular feature of speech. They calculate ‘distance’ between reference feature and the processed one. Objective measures are expected to have high correlation with subjective ones. However, there are many objective assessment methods which do not correlate with subjective methods. Therefore, it is important to use proper objective measures to evaluate speech processing performance with respect to application.

When speech processing methods are applied, a wide variety of sources of speech distortions come up and their effects on speech signal are not similar. Therefore, measurement methods have to be chosen carefully. Early measurement methods were proposed to evaluate distortions due to codecs or network conditions [61]. However, it is not definite whether a measurement method which was developed to evaluate communication distortions can be successful at evaluating distortions due to reverberant conditions.

In this thesis, it is considered that dereverberation algorithm has some distortion effects on speech signal. The proposed dereverberation method consists of multiple algorithms. Since different algorithms introduce various distortions in the original speech signal, a group of objective measures have to be used. Loizou made a comprehensive study to assess correlation of existing objective measurement methods with subjective evaluations and he specified the most accurate measures [9]. He used speech signals which contain spectral subtraction and statistical model processing distortions in the experiments and compared performance of different objective measures with large number speech samples. Since the proposed dereverberation is based on both spectral subtraction and statistical models, objective measures are chosen according to Loizou’s study in this work.

The objective quality measures which are ‘frequency-weighted segmental SNR’ (FWSegSNR), ‘cepstrum distance’ (CD), ‘log likelihood ratio’ (LLR) and ‘perceptual evaluation of speech quality’ (PESQ) are described before the experiments.

4.2.1. Cepstrum Distance

CD is a measure of the log spectrum distance between clean and distorted speech signals [9]. It is used to show discrepancy between dereverberated and reference signals. Cepstrum is calculated by taking IDFT of logarithm of the spectrum. CD can be calculated as:

$$CD = {}^{10}\!/_{log10} \sqrt{2\sum_{i=1}^{p}\left\{c_x(i) - c_y(i)\right\}^2} \qquad (4.\ 1)$$

where c_x, c_y denote cepstrum coefficients of clean and processed speech signals respectively. Cepstrum coefficients can be used in different processing algorithms, so it is very efficient to use this measure in order to evaluate performance.

4.2.2. Log Likelihood Ratio

LLR is an LPC based objective measurement method [62]. It describes discrepancy between processed and clean speech signals. It is written as:

$$LLR = log\left(\frac{\boldsymbol{a}_p\boldsymbol{R}_c\boldsymbol{a}_p^T}{\boldsymbol{a}_c\boldsymbol{R}_c\boldsymbol{a}_c^T}\right) \qquad (4.\ 2)$$

where $\boldsymbol{a}_p, \boldsymbol{a}_c$ is LPC vectors of processed speech and clean speech signals respectively. $\boldsymbol{R}_c$ shows clean speech correlation matrix. Clean speech signals are used as reference signals, and this method measures how similar processed and reference signals are.

4.2.3. Perceptual Evaluation Speech Quality

PESQ is one of the most comprehensive objective measurement methods which represents human subjective test. PESQ measures distortions in a speech signal. It is used to estimate degradation effects of different network conditions and communication systems. However, since it assesses general speech signal quality, it can be used for dereverberated speech signals. PESQ score of a system can be described as:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{4. 3}$$

$$a_0 = 4.5 \quad a_1 = -0.1 \quad a_2 = -0.0309$$

where D_{ind}, A_{ind} represent average disturbances and asymmetrical disturbances respectively [9]. PESQ of a reference signal is 4.5. When distortions in speech signal increase, PESQ value reduces.

4.2.4. Frequency – Weighted Segmental SNR

SNR is the oldest and simplest speech assessment metric. SNR of speech recordings does not correlate very well with speech quality; however, it can provide benefits. After SNR of each speech segment is extracted, average of the results gives 'segmental SNR'. Segmental SNR can be generated through frequency domain. FWsegSNR is average of the SNR through short time segments in frequency domains [62]:

$$fwSNRSeg = \frac{10}{K} \sum_{k=0}^{K-1} \frac{\sum_{l=0}^{L-1} w(l,k) \log_{10} \frac{X(l,k)^2}{\{X(l,k) - \bar{X}(l,k)\}^2}}{\sum_{l=0}^{L-1} w(l,k)} \tag{4. 4}$$

where $X(l,k), \bar{X}(l,k)$ show reference and processed STFT coefficients respectively. $w(l,k)$ is weight coefficient of the frame in related subbands. K represents total subband number and L shows total frame number in recorded speech.

These objective assessment methods are selected because they correlate very well with subjective test results for processed speech signal. These measures give overall quality of a speech signal. In this way, the undesirable effects of the dereverberation algorithm on a speech signal can be deduced. However, there is still no performance measurement method for the dereverberation algorithm. Speech to reverberation (SRMR) metric was proposed to concentrate on measuring performance of dereverberation algorithms [63]. In this thesis, SRMR method will be used to measure reverberation level in speech signals.

4.2.5. Speech to Reverberation Ratio

SRMR metric is different from previous objective assessment methods. It does not need a reference signal to evaluate a processed speech. Speech signal can be represented by a carrier frequency and amplitude modulation. Information conveyed via speech signal can be identified by amplitude variation of the signal. Hilbert transform is used to generate an envelope of speech signal. The temporal envelope of a speech signal can be shown in Figure 4.1. It is proved that envelope spectral characteristics correlate with subjective evaluations of speech quality.

Hilbert envelope of a clean speech signal contains frequencies of 2 Hz – 20 Hz [63]. In chapter 2, the uncertain nature of reverberation was explained. This means that reverberation causes fluctuations in the envelope, so high frequency content of the envelope increases. Therefore, spectral characteristic of the envelope signals can be useful cues in order to distinguish reverberant signals from clean speech signals.

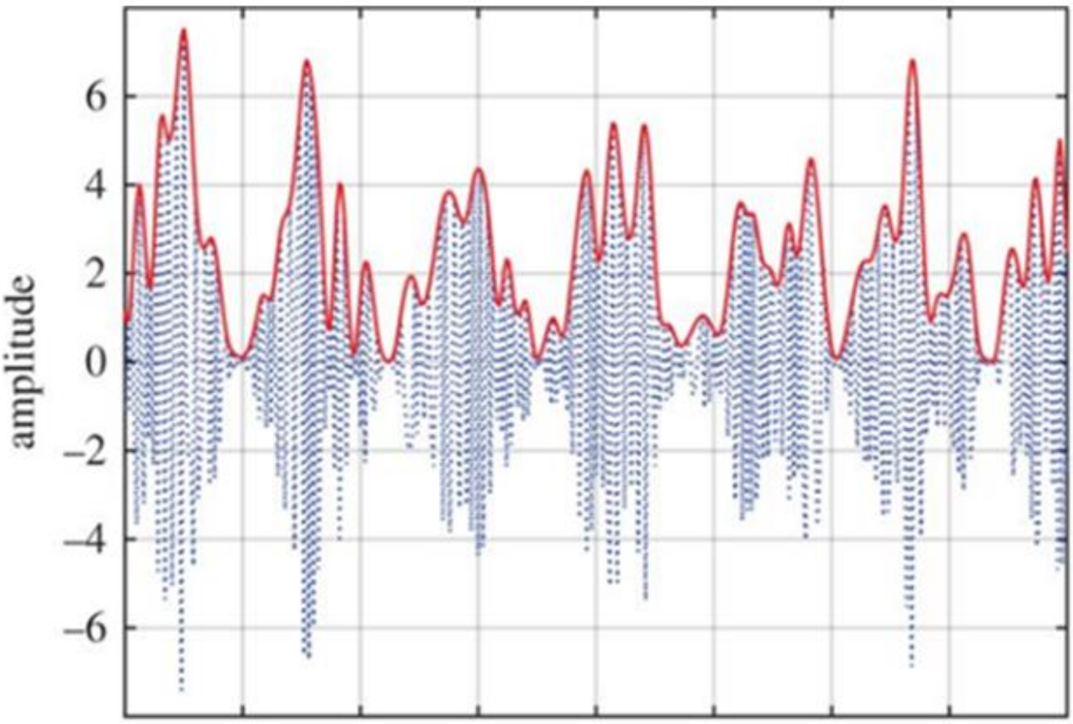


Figure 4.1. Hilbert envelope a typical speech signal.

Apart from distinguishing reverberant speech from clean speech, numeric SRMR score can be derived from Hilbert envelope spectral contents with respect to the density of reverberation. In this way, performance of a dereverberation method can be evaluated objectively.

Hilbert envelope is divided into K frequency bands, $e_1(l)$ ... $e_k(l)$. Spectral power of each frequency bin f in this band is obtained by taking square of Fourier transform,

$$\varepsilon_k(l,f) = \left|F\left(e_k(l;f)\right)\right|^2 \tag{4. 5}$$

where l denotes frame number in each band. Then, average spectral power of the frequency bin in k^{th} frequency band of modulation over the frames,

$$\bar{\varepsilon}_{f,k} = \frac{1}{L}\sum_{i=1}^{L} \varepsilon_k(i,f) \tag{4. 6}$$

where L is the frame number. Take summation of each frequency bin in related subband,

$$\bar{\varepsilon_k} = \sum_{f} \bar{\varepsilon}_{f,k}. \tag{4. 7}$$

K value can be optimized with respect to application. $K = 8$ has superior performance to evaluate dereverberated speech, so SRMR is defined as:

$$SRMR = \frac{\sum_{k=1}^{4} \bar{\varepsilon_k}}{\sum_{k=5}^{8} \bar{\varepsilon_k}} \ . \tag{4. 8}$$

In conclusion, SRMR can be described as a ratio of spectral power of low frequencies to high frequencies in Hilbert envelope. Since reverberation has whitening effects on speech signals due to uncertainty, it causes envelopes of speech to fluctuate at higher modulation frequencies. Therefore, reverberation reduces SRMR value according to Eq. (4.8).

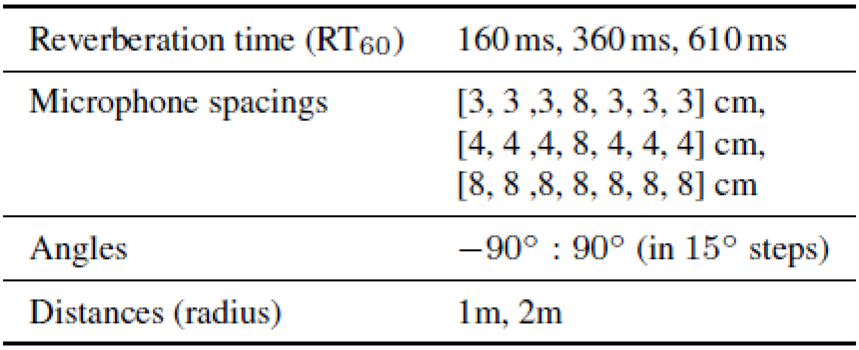
4.3. Database

A dataset should be provided to evaluate performance of the dereverberation method under different conditions. In this work, it is assumed that the utterances are spoken by a stationary speaker and they are captured by linear microphone array.

In the beginning of this chapter, objective assessment methods were explained. Most of them require a reference signal to evaluate performance of the system. Therefore, in this work, reverberant utterances are generated artificially by convolving RIR and anechoic speech signals. In this way, success of the dereverberation method on reverberant signals can be measured with respect to the anechoic signal.

BarIlan University (BIU) impulse response database is used to provide effects of a quiet reverberant room [64]. It contains transfer functions of the microphone arrays in reverberant room for different setup configurations. The room dimensions are 6 m × 6 m × 2.4 m. Impulse responses of the reverberant room were generated in different conditions. Detailed measurement conditions can be summarized as:

Table 4.1. The recording conditions of BIU database



Reverberation time (RT_{60})	160 ms, 360 ms, 610 ms
Microphone spacings	[3, 3 ,3, 8, 3, 3, 3] cm, [4, 4 ,4, 8, 4, 4, 4] cm, [8, 8 ,8, 8, 8, 8, 8] cm
Angles	$-90^\circ : 90^\circ$ (in 15° steps)
Distances (radius)	1m, 2m

As it can be seen in Table 4.1, there is a wide variety of conditions in the dataset. Each configuration of the impulse responses affects dereverberation algorithm from a different perspective:

- Various reverberation times mean different room acoustic properties. The dereverberation method can be performed and evaluated for different room acoustic conditions by using utterances in various reverberation times.
- Microphone array configuration and DOA affect MVDR algorithm drastically. They are the main parameters of beampattern function. Therefore, it is informative to observe the dereverberation method for different configurations.
- Source distance affects the plane wave assumption. Plane wave assumption suffers in short distances. In this work, correlation of reverberation and MVDR algorithm are based on plane wave modeling, so performance of the algorithm

should be observed in different source distances. In addition, in short distances direct sounds dominate recorded signals, and diffuse field modeling may suffer in short distances. Since the proposed approach depends on diffuse modeling heavily, it must be performed in different source distances.

The configuration of the setup is shown as:

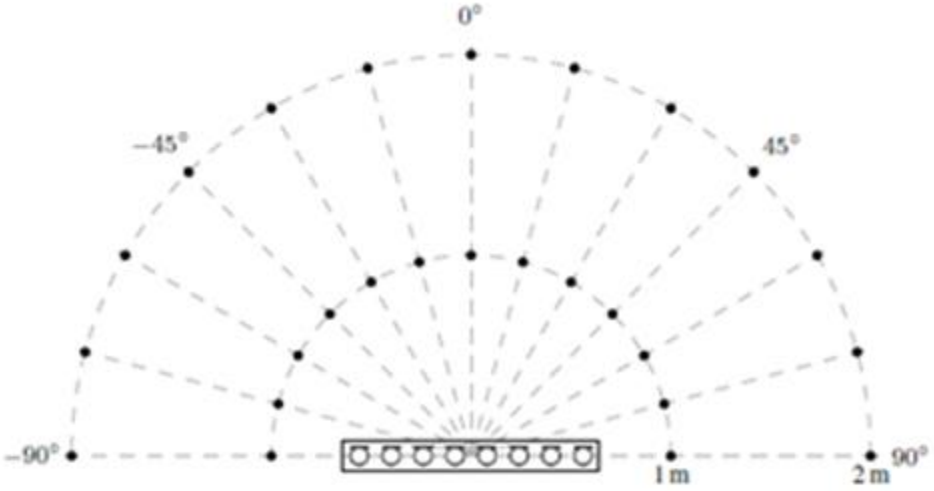


Figure 4.2. Setup of impulse response dataset.

Knowledge about the system to generate impulse response is given as:

- 8 Omni directional microphones of AKG CK32
- Fostex 6301 BX as loudspeakers
- 48 kHz sampling frequency
- 24 bit resolution.

The impulse responses for different reverberation times can be seen in Figure 4.3.

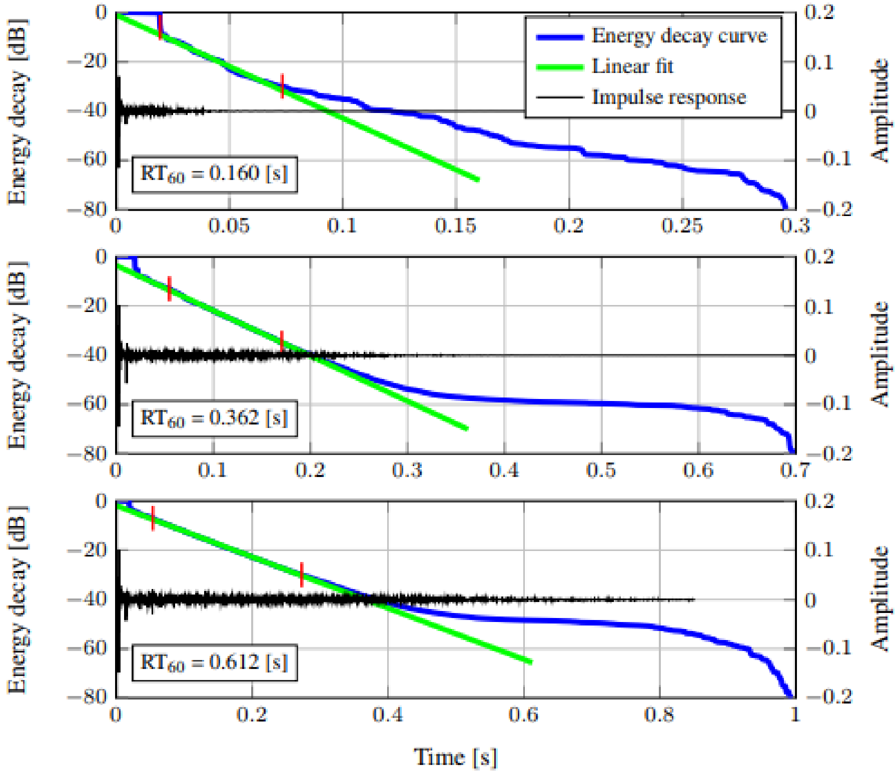


Figure 4.3. The impulse responses of different reverberation times.

By using this impulse response dataset, any reverberant utterance can be obtained artificially in MATLAB. Source-microphone distance, DOA effects on the dereverberation method can be analyzed in different reverberation times.

4.4. Implementation Details and Parameter Settings

In this section, experimental work of the dereverberation algorithm will be explained. The proposed algorithm consists of two stages. Each stage consists of smaller successive parts which work together, so experimental work will be analyzed part by part.

- **Downsampling**

Sampling frequency of the generated speech is 48 kHz, which is beyond the required rate of speech processing. Large number of samples increases computational complexity of the system. Frequencies of a human speech signal can be shown in Figure 4.4:

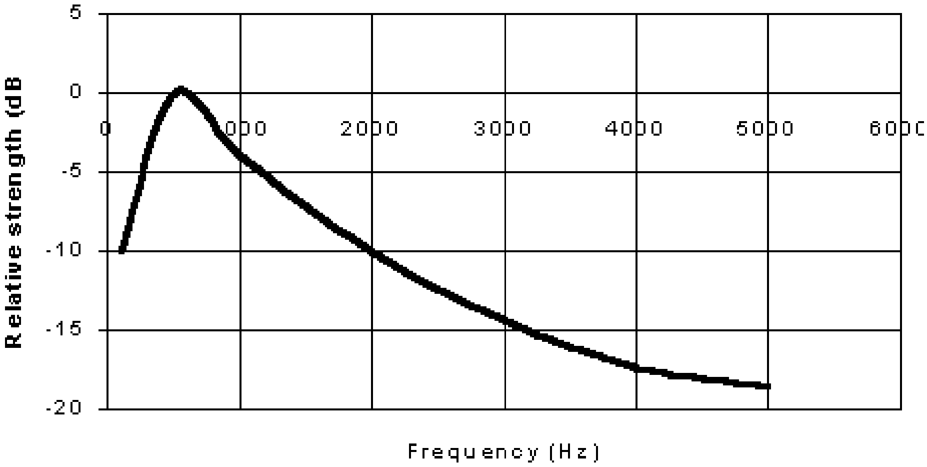


Figure 4.4. Frequency content of a speech signal.

As it can be seen in Figure 4.4, the majority of speech frequencies are in 300 Hz – 2000 Hz. If high frequencies of speech are taken in consideration, $fs = 16000$ Hz is sufficient for speech processing algorithms. Therefore, downsampling is performed initially in order to reduce computational complexity of the system.

- **Estimation of MVDR Parameters**

MVDR algorithm needs correlation of late reverberation. Necessary parameter for estimating LRSV is reverberation time of the room. Theoretical background of reverberation time estimation is given in Chapter 3. Reverberation time can be found by sound decays in reference microphone signal. The sound decays are searched in segments by subsegment constraints. Speech signals are stationary in short time-scales, so subsegment length is chosen 20 msec. There are 10 subframes in a single segment. If the constraints are satisfied for at least 5 subsegments successively, ML estimation is applied to corresponding part to find decay constant. There must be an overlapping at least 5 subsegments between shifted segments in order not to miss any decaying part.

In reference microphone signal, ML estimation is applied for each decaying part of speech. In order to obtain a more robust system, it should be paid attention to variance of the estimations. There can be a deceptive decay in speech signal due to its own characteristics. These parts may mislead the result. In order to reduce variance, all estimations should be taken in account by a forgetting factor:

$$RT_{60}(\lambda) = (1-\alpha).\mathrm{RT}_{60}^{*}(\lambda) + \alpha.RT_{60}(\lambda-1) \qquad (4.\,9)$$

where λ is segment number, $\mathrm{RT}_{60}^{*}(\lambda)$ is current segment estimation, $\alpha = 0.8$ is forgetting factor. Various forgetting factors can be used according to the effect of new estimation. In addition, a compulsory interval for the estimation can be used:

$$|RT^{*}{}_{60}(\lambda) - RT_{60}(\lambda-1)| < \beta \qquad (4.\,10)$$

where β value is related to estimation variance. If the recent estimation is not inside a specified interval with respect to previous ones, it can be neglected. Thus, the result is constituted by just low variance estimates.

Reverberation time constant can be found by Eq. (3.44), and it is used at calculating PSD of late reverberation as in Eq. (3.41). LRSV is used as coefficient of the correlation matrix, $\boldsymbol{\phi}$. In addition, previous samples of speech are used for calculating LRSV. Value of delay parameter (N_e) should be specified with respect to the mixing time of the room. Mixing time in the experiment room is approximately 10 msec. In order to ensure existing diffuse field in the room, mixing time is considered between 20 – 25 msec, so $N_e = 3$. This additional margin reduces performance of late reverberation suppression; however, distortions in early speech components are avoided.

In order to reduce the effects of probabilistic nature of reverberant signals, LRSV should be estimated by multiple microphones. Average of the calculated PSDs is:

$$\sigma_{LRSV} = \frac{1}{M} \sum_{m=1}^{M} \sigma_{LRSV}^{m} \tag{4. 11}$$

where m is the microphone number. Averaged spectral power is used in the calculations to improve results. Correlation matrix of reverberant part can be written as in Eq. (3.29).

After correlation of reverberation, DOA is required to describe the direct transfer function vector ($\boldsymbol{g'}_d(k)$). In DOA estimation, speech signals are divided into 10 msec frames. Maximum dominant frequency and energy of the frames are estimated, and they are classified as 'silence' or 'speech' according to the measurements. If these values are higher than the threshold for 5 frames successively, the frames are classified as speech. Silence classification is made just in the opposite manner. The speech parts which come after silence parts are detected. Correlation of the microphone signals during these transition regions directly gives time delays. In order to reduce the effect of reverberation, initial 15 msec parts of the transitions are used for correlation calculation. This process continues all through the speech. The average of numerous time delay estimations gives more robust results. Finally, transfer function of direct part can be obtained by time delays of the microphones with respect to the reference microphone.

By using these necessary parameters, reverberant signals can be processed. Both MVDR and single channel algorithm operate in frequency bands. Therefore, it is necessary to write signals in frequency domain. Although speech signals are non-stationary, they are stationary in short frames and the frequency content does not change in these short time scales [65]. Speech segmentation is applied with respect to this feature of speech signals. Hamming window is chosen for segmentation to obtain satisfactory frequency resolution. The length of windows is 32 msec, and %75 overlapping is used between speech segments, i.e., there are 8 msec frame shifts. STFT coefficients of each segment are extracted.

MVDR filter coefficients can be calculated by direct speech transfer function vector and correlation of late reverberation as in Eq. (3.14). The filter coefficients are updated for each frame in the subbands.

- **Single Channel Algorithm**

In the single channel dereverberation stage, the most important term is PSD of early speech components, because it directly affects filter estimation in Eq. (3.67). Assume that late reverberation is independent of early speech components, so spectral power of early speech can be described by subtraction:

$$\sigma_{l,k}^{2} = \sigma_{l,k}^{\overline{X}} - \sigma_{l,k}^{LRSV} \qquad (4.\ 12)$$

where $\sigma_{l,k}^{\overline{X}}$ is spectral power of MVDR output signals, $\sigma_{l,k}^{LRSV}$ is spectral power of late reverberation at the output of MVDR, $\sigma_{l,k}^{2}$ is spectral power of early reverberation at the output of MVDR, k is frequency bin, l is frame number.

Coefficients of the single channel algorithm are estimated by Eq. (3.67). Prediction delay is specified with respect to mixing time. 25 msec is reasonable delay according to mixing time. Since segments are shifted %25 (8 msec) in the processing and N_e =3:

$$\overline{D}_{k}^{l} = \bar{X}_{k}^{l} - \boldsymbol{w}_{k}^{H}\, \overline{\boldsymbol{X}}_{k}^{l-3} \qquad (4.\ 13)$$

where $\overline{\boldsymbol{X}}_{k}^{l-3} = \left[\, \bar{X}_{k}^{l-3}\ \bar{X}_{k}^{l-4}\ \bar{X}_{k}^{l-5} \ldots\ldots\ \bar{X}_{k}^{l-L+1}\right]^{T}$.

Reverberation time changes with frequency, and it affects the length of $\overline{\mathbf{X}}_{k}^{l-\,N_e}$ vector in Eq. (4.13). Reverberation time is not estimated for each frequency in this work for the sake of simplicity. However, L can be adjusted roughly with respect to frequency.

In this situation, frequency content is divided into three regions. L_1, L_2, L_3 show filter length of low, dominant and high frequency content, respectively. Assume that dominant frequency region is 600 Hz –1000 Hz interval and L_2 is used to show this region. The relation between reverberation time and filter length is given as:

$$L_2 = \frac{RT_{60}}{T_{shift}} \tag{4. 14}$$

where $RT_{60}$is estimated reverberation time. $T_{shift} = 8$ msec is frame shifts. Reverberation time is higher at lower frequencies; therefore, it is assumed that $(L_1, L_2, L_3) = (40, 35, 30)$ in $RT_{60} = 310$ msec case. In this way, the effect of frequency on reverberation time is modeled roughly.

The summary of the whole algorithm in this thesis can be given as:

```
1.  Downsample the speech signal to $f_s = 16\ kHz$.
2.  Estimate DOA, $(\theta)$ and reverberation time, $RT_{60}$ blindly.
3.  Find STFT of each recorded signal $X_k^{l,m}$.
4.  Estimate $\sigma_{LRSV,k}^{l,m}$ of each recording, take average $\frac{1}{M}\sum_{m=1}^{M} \sigma_{LRSV}^{l,m}$.
5.  Calculate $\boldsymbol{\phi}_r(k,l)$ and $\boldsymbol{g}'(\theta)$, then $\boldsymbol{w}_{MVDR}(k,l)$.
6.  for $k = 1: number\ of\ frequency\ bands$
7.    for $l = 1: frame\ numbers$
8.      $\bar{\mathrm{X}}_k^{\mathrm{l}}(k,l) = \boldsymbol{w}_{MVDR}^H(k,l)\boldsymbol{X}(k,l)$
9.    end
10. end
    The output of the MVDR is fed to the single channel algorithm.
11. $\sigma_{l,k}^{LRSV} = \boldsymbol{w}_{MVDR}^H(k,l)\boldsymbol{\phi}_r(k,l)\boldsymbol{w}_{MVDR}(k,l)$ and $\sigma_{l,k}^2 = \sigma_{l,k}^{\bar{X}} - \sigma_{l,k}^{LRSV}$
12. Calculate the second stage filter, $\boldsymbol{w}_k$ for each band by ML estimation.
13. for $k = 1: frequency\ bands$
14.   $\bar{\mathrm{D}}_{l,k} = \bar{\mathrm{X}}_{l,k} - \boldsymbol{w}_k^H \bar{\mathbf{X}}_{l-N_e,k}$
15.   $k = k + 1$
16. end
```

Implementation sequence is given item by item:

1. RT_{60} is estimated by recording at the reference microphone. In order to detect the decaying curves recording signal is divided into 200 msec segments, then segments are divided into 20 msec subsegments. The decaying curves are detected in the segments by subsegments constraints. If a segment contains a decaying curve MLE is applied to estimate RT_{60}. Segments are shifted with %50 in order not to miss any decaying curve. After estimating RT_{60}'s from all decaying curves average is taken to get more accurate result.

2. All speech onsets are detected by VAD algorithm at all microphones. 15 msec frames are taken from onsets and the correlation of the frames between microphones are calculated. The phase differences of the microphones with respect to the reference microphone are found by correlations. Relative early transfer function vector is generated by the phase differences.

3. STFT of the recordings is taken. In this process %75 overlapping 32 msec Hamming window is used (Shifting = 8 msec). Since the mixing time is approximately 25 msec, $N_e = 3.$ LRSV is found as:

   $$\phi_r(l,k) = e^{-2\alpha R N_e}\,\phi_X(l-N_e,k) \qquad (4.\ 15)$$

   where ϕ_X, ϕ_r are spectral variance of observation and late reverberation.

4. Correlation matrix of reverberation is calculated as:

   $$\boldsymbol{\phi}_r(l,k) = \phi_r(l,k)\begin{bmatrix} 1 & sinc(\gamma) \\ sinc(\gamma) & 1 \end{bmatrix} \qquad (4.\ 16)$$

   where $\gamma = \frac{(\frac{2\pi f_s k}{K})d}{c}$, k is frequency bin, K is FFT number, d is microphone distance in m, c is sound wave speed m/sec, f_s is sampling frequency.

5. Apply MVDR to the recordings.

$$\overline{X}(l,k) = \boldsymbol{w}^{H}(k)\,\boldsymbol{X}(l,k) \qquad (4.\ 17)$$

$$\boldsymbol{w}^{H}(k) = \frac{\boldsymbol{g}'^{H}(k)\boldsymbol{\phi}_{r}^{-1}(l,k)}{\boldsymbol{g}'^{H}(k)\boldsymbol{\phi}_{r}^{-1}(l,k)\boldsymbol{g}'(k)}$$

   where $\boldsymbol{\phi}_{r}, \boldsymbol{g}'$ are correlation matrix of late reverberation and relative early transfer function vector.

6. Spectral power of early reverberation is calculated at the output of MVDR.

$$\sigma_{l,k}^{LRSV} = \phi_{l,k}\boldsymbol{w}^{H}(l,k)\Gamma(l,k)\boldsymbol{w}(l,k) \qquad (4.\ 18)$$

$$\sigma_{l,k}^{2} = \sigma_{l,k}^{\overline{X}} - \sigma_{l,k}^{LRSV}$$

   where $\sigma_{l,k}^{2}, \sigma_{l,k}^{\overline{X}}, \sigma_{l,k}^{LRSV}$ are spectral power of early reverberation, observed speech and late reverberation at the output of MVDR algorithm.

7. The filter vector of the single channel dereverberation is estimated by MLE.

$$\boldsymbol{c}_{k} = \left(\sum_{l}\frac{\overline{\boldsymbol{X}}(l-D,k)\overline{\boldsymbol{X}}(l-D,k)^{H}}{\sigma_{l,k}^{2}}\right)^{-1}\left(\sum_{l}\frac{\overline{\boldsymbol{X}}(l-D,k)\overline{X}^{*}(l,k)}{\sigma_{l,k}^{2}}\right) \qquad (4.\ 19)$$

   where $\boldsymbol{c}_{k}$ is the dereverberation filter. Output of the single channel dereverberation is calculated:

$$\overline{D}(l,k) = \overline{X}(l,k) - \boldsymbol{c}^{H}(k)\overline{\boldsymbol{X}}(l-N_{e},k) \qquad (4.\ 20)$$

   where $N_{e} = 3$, the length of $\overline{\boldsymbol{X}}$ vector is reverberation time in terms of frame lengths as Eq. (4.14).

8. Take inverse Fourier transform to obtain dereverberated speech signal with 32 msec %75 overlapping Hamming window.

4.5. Results

In this section, the proposed dereverberation algorithm is evaluated by objective quality measures which are explained in the beginning of this chapter. In order to evaluate the algorithm extensively, it is tested under various conditions. The results of the objective assessments are written before and after the dereverberation process to show relative improvements in reverberant speech signal for each different condition. The evaluations are also performed after the first stage to show effects of each stage in the dereverberation.

Reverberation effects can be seen in spectrograms. Therefore, spectrogram of a processed signal is included to show dereverberation performance at the end of this section.

4.5.1. Performance of the Dereverberation

Initially, the proposed method is tested under different room acoustics. The utterances are conveyed to omni-directional microphone array as $DOA\ (\theta) = 90^\circ$ in Figure 4.2. There are two microphones in the array and the distance between microphones is 8 cm. Speaker position and configuration of the microphones are fixed during recordings.

The measurements are taken under three different reverberation times. Also, two distinct source distances are used in the recordings. Table 4.2 shows the results obtained with objective assessment methods in the corresponding conditions. Measures with 'Input' subscript show evaluations of the observed speech. 'MVDR' and 'Output' subscripts mean that the measures are taken in output of the first and the second stages respectively.

In order to interpret Table 4.2, it is necessary to give information about measure values. SRMR shows ratio of speech to reverberation shortly. Therefore, higher value means more successful dereverberation. CD and LLR illustrate discrepancy between anechoic and processed speech signals, so lower value means lower discrepancy. PESQ is a common measure and the higher PESQ is, the better quality of speech becomes. Lastly, higher FwSNRSeg shows better SNR i.e. more successful dereverberation.

The proposed model generally provides dereverberation without a significant distortion in anechoic speech. SRMR is directly related to reverberation level, while other measures are related to processing distortions in speech signal. The algorithm provides a gain in SRMR except in the room which has 160 msec reverberation time. Distance has a positive effect on the proposed method. It provides better dereverberation in larger distances. PESQ, CD, LLR results behave similar to SRMR. However, FwSNRSeg shows that the first stage always reduces segmental SNR. This can be related to estimation of late reverberation correlation matrix in MVDR stage. Estimation errors may cause SNR to decrease. However, the second stage recovers SNR of the whole algorithm. In short, the proposed algorithm suppresses reverberation without causing a significant distortion in the original speech with respect to the objective quality measures except for $RT_{60} = 160$ msec.

Table 4.2. The results for different conditions at input, output of the first and the second stages.

Results	***Simulated Data***					
	$RT_{60} = 160\ ms$		$RT_{60} = 360\ ms$		$RT_{60} = 610\ ms$	
	$1m$	$2m$	$1m$	$2m$	$1m$	$2m$
$SRMR_{INPUT}$	5.98	6.12	5.29	5.24	4.42	4.48
$SRMR_{MVDR}$	5.60	5.90	5.23	5.48	4.57	4.82
$SRMR_{OUTPUT}$	5.78	6.08	5.58	6.24	5.44	6.03
$PESQ_{INPUT}$	3.67	3.23	3.08	2.69	2.71	2.42
$PESQ_{MVDR}$	3.70	3.51	3.29	2.84	2.90	2.48
$PESQ_{RESULT}$	3.78	3.54	3.64	3.09	3.37	2.66
CD_{INPUT}	1.65	1.65	2.17	2.51	2.78	3.25
CD_{MVDR}	1.57	1.62	2.15	2.41	2.76	3.18
CD_{OUTPUT}	1.57	1.61	1.76	1.99	2.05	2.70
LLR_{INPUT}	0.39	0.46	0.47	0.63	0.65	0.83
LLR_{MVDR}	0.38	0.46	0.48	0.65	0.64	0.81
LLR_{OUTPUT}	0.35	0.44	0.39	0.50	0.51	0.75
$FwSNRSeg_{INPUT}$	15.38	12.73	13.98	10.79	11.77	9.63
$FwSNRSeg_{MVDR}$	14.87	12.65	12.47	10.72	10.65	9.08
$FwSNRSeg_{RESULT}$	15.63	12.90	14.19	11.57	12.12	9.98

The most comprehensive evaluation methods are ‘SRMR’ and ‘PESQ’. Therefore, these results are shown in the graphs individually. As it can be seen in the graphs, the proposed method improves quality of speech signal in most of the conditions apart from $RT_{60} = 160$ msec. Performance of each algorithm is better in higher reverberation time environments. It is expected because early components can be dominant in lower reverberation time. Diffuse field assumption suffers in low reverberation time. Distance is also another significant parameter in the algorithm. In larger distance, reverberation becomes dominant and diffuse field conditions are satisfied, so the probabilistic models in the algorithm get more accurate. Also, plane wave model works in large distances more accurately. Therefore, it is expected that each stage of the algorithm works better in larger distances and reverberation time.

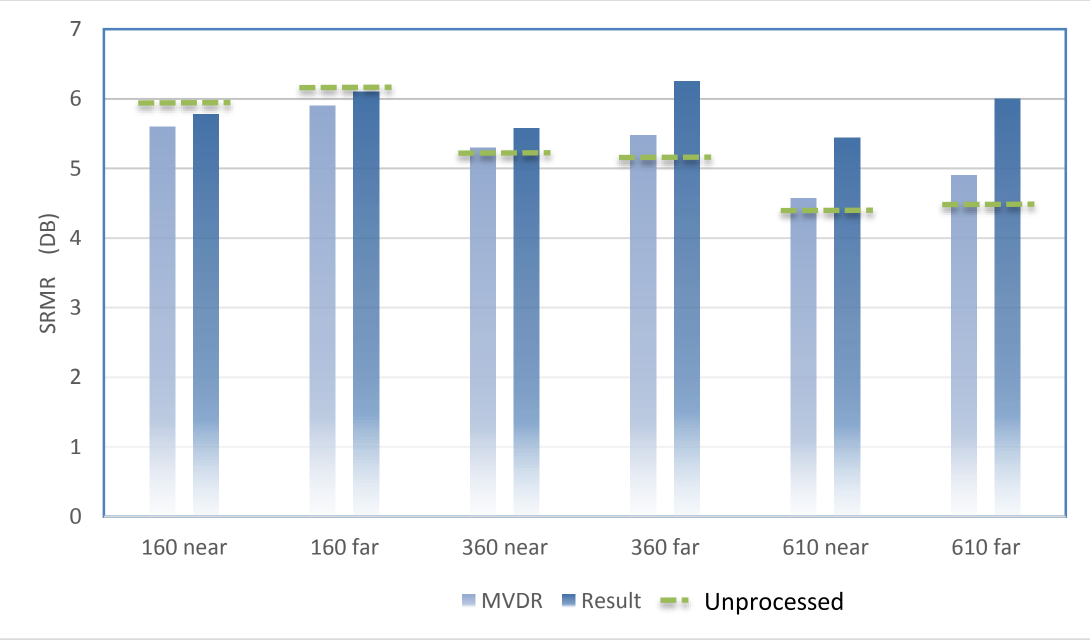


Figure 4.5. Performance comparison of the algorithm for different conditions with respect to SRMR metric.

PESQ score also verifies SRMR results. Dereverberation algorithm does not have detrimental effect on the desired signal while suppressing the reverberation according to PESQ.

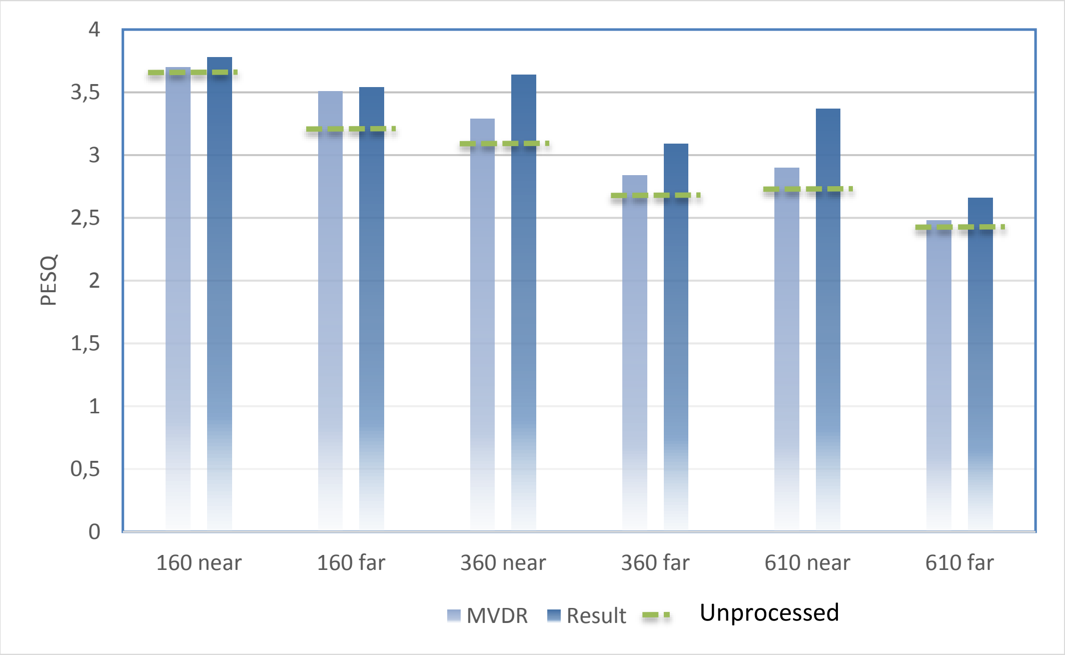


Figure 4.6. Performance comparison of the algorithm for different conditions with respect to PESQ.

One of the parameters which affects performance is the distance between microphones. The distance must be lower than that of half of speech signal wavelength. However, when the distance increases in this limit, beamwidth of MVDR filter reduces. This provides more precise filter for reverberation. The results are summarized in Table 4.3.

Table 4.3. The results of objective quality measures for different microphone distances.

Results	***Simulated Data***	
	$RT_{60} = 360\ ms$	
	$d = 0.08\ m$	$d = 0.16\ m$
$SRMR_{INPUT}$	5.31	5.31
$SRMR_{MVDR}$	5.41	5.91
$SRMR_{OUTPUT}$	6.49	7.06
$PESQ_{INPUT}$	2.77	2.77
$PESQ_{MVDR}$	2.81	2.83
$PESQ_{RESULT}$	3.02	3.05
CD_{INPUT}	2.31	2.31
CD_{MVDR}	2.39	237
CD_{OUTPUT}	2.04	2.21
LLR_{INPUT}	0.62	0.62
LLR_{MVDR}	0.65	0.65
LLR_{OUTPUT}	0.70	0.71
$FwSNRSeg_{INPUT}$	9.16	9.16
$FwSNRSeg_{MVDR}$	8.89	9.04
$FwSNRSeg_{RESULT}$	9.72	9.92

SRMR and PESQ results are given in the following graphs. It is seen that performance of the algorithm improves when the distance between microphones increases.

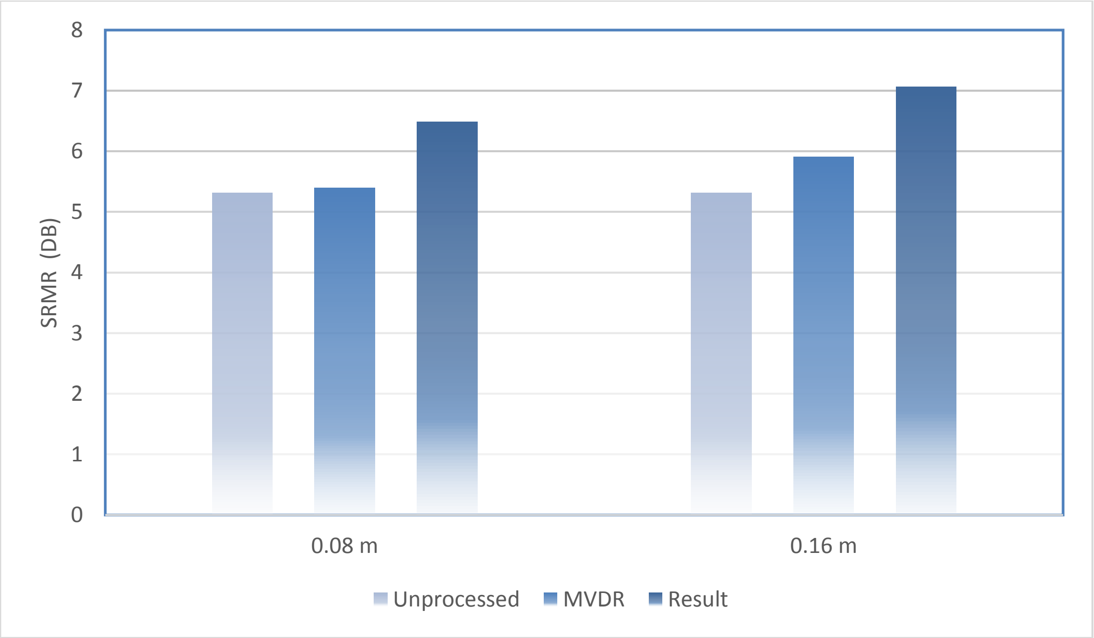


Figure 4.7. The performance of the algorithm is better at 0.16 m distance according to SRMR metric.

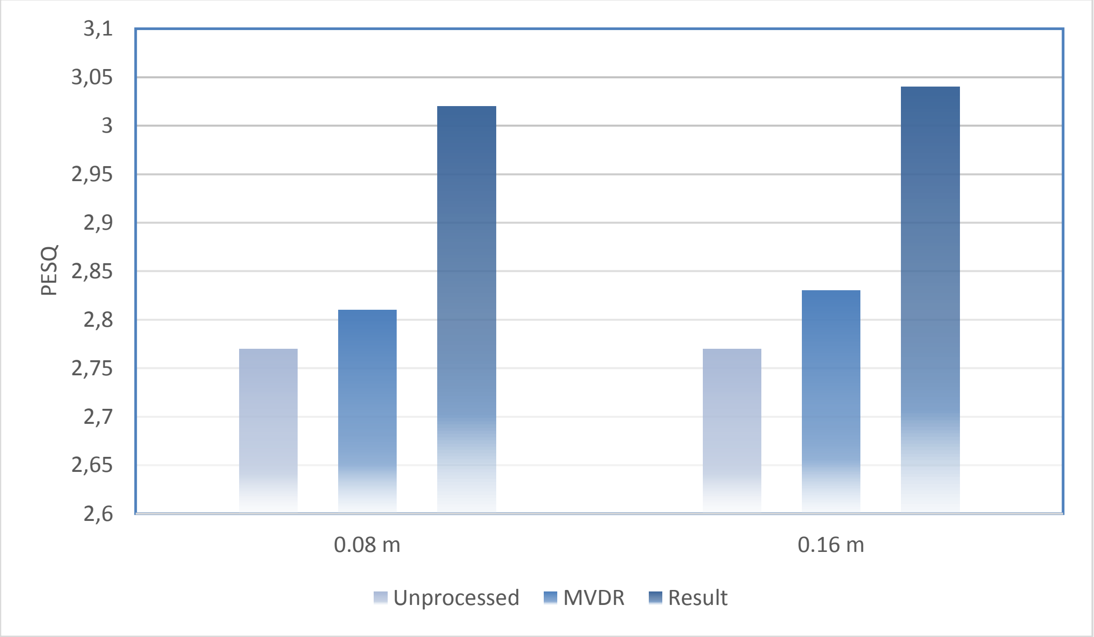


Figure 4.8. PESQ of the algorithm enhances when the distance between the microphones increases.

Table 4.4. The results of objective quality measures for different DOAs.

Results	***Simulated Data***	
	$\boldsymbol{RT_{60} = 610\ ms}$	
	$\boldsymbol{\theta = 0°}$	$\boldsymbol{\theta = 45°}$
$SRMR_{INPUT}$	3.78	3.80
$SRMR_{MVDR}$	4.20	3.86
$SRMR_{OUTPUT}$	5.57	4.56
$PESQ_{INPUT}$	2.34	2.33
$PESQ_{MVDR}$	2.41	2.48
$PESQ_{RESULT}$	2.59	2.64
CD_{INPUT}	3.19	3.26
CD_{MVDR}	3.18	3.25
CD_{OUTPUT}	2.86	2.99
LLR_{INPUT}	0.82	0.79
LLR_{MVDR}	0.80	0.79
LLR_{OUTPUT}	0.75	0.78
$FwSNRSeg_{INPUT}$	8.91	9.57
$FwSNRSeg_{MVDR}$	8.53	9.12
$FwSNRSeg_{RESULT}$	9.20	10.41

It is wise to analyze the results for different DOAs because it may change performance of MVDR filter. As it can be seen in Table 4.4, the whole algorithm provides similar performance at different DOA conditions.

The performance comparison between two algorithms may not be reliable when the datasets are different for each algorithm. However, in order to understand the performance of the proposed topology, the objective quality measures in Table 4.2 can be evaluated with respect to other proposed topologies. In this part, the results are compared with the proposed algorithms in *'Reverb Challenge 2014'* [66]. In this workshop, the utterances are conveyed to a microphone array under various experimental conditions. The experiments are conducted in two different microphone to speaker distances (50 cm, 200 cm) and there are three RT_{60} conditions (0.3 sec, 0.6 sec, 0.7 sec). Although the dataset in this workshop is completely different from the utterances in this work, the results under RT_{60} = 0.3 sec (Sim Room1) and RT_{60} = 0.6 sec (Sim Room2) conditions in far field (2 m) are compared with our results in Table 4.2 roughly.

In the workshop there are three different two-channel topologies [67, 68, 69]. The objective quality measures of the unprocessed and processed speech signals are given in Figure 4.9 – 11.

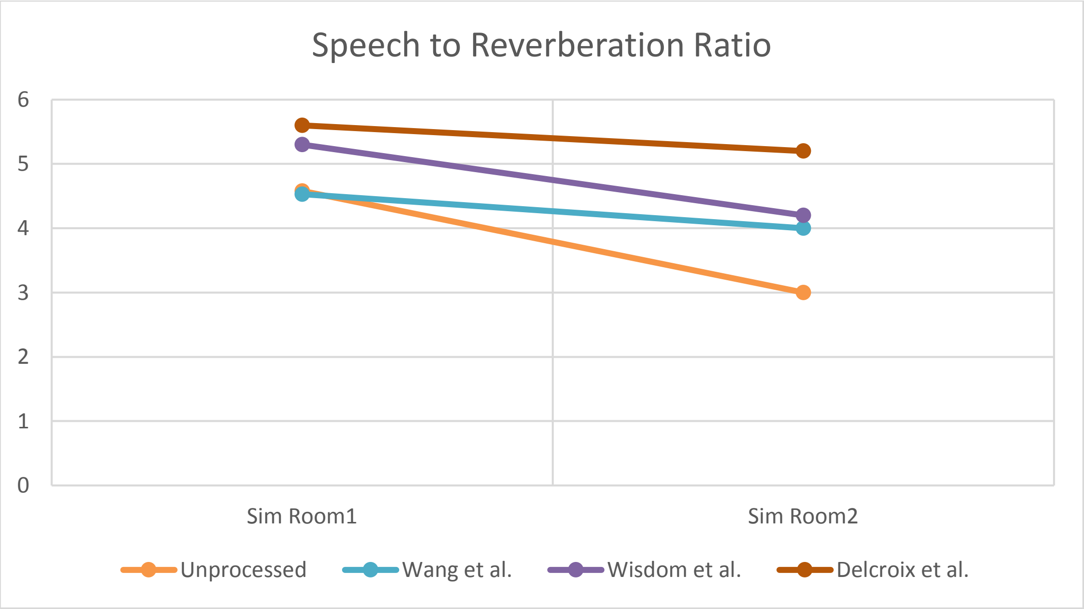


Figure 4.9. SRMR results of the two microphone solutions in the workshop.

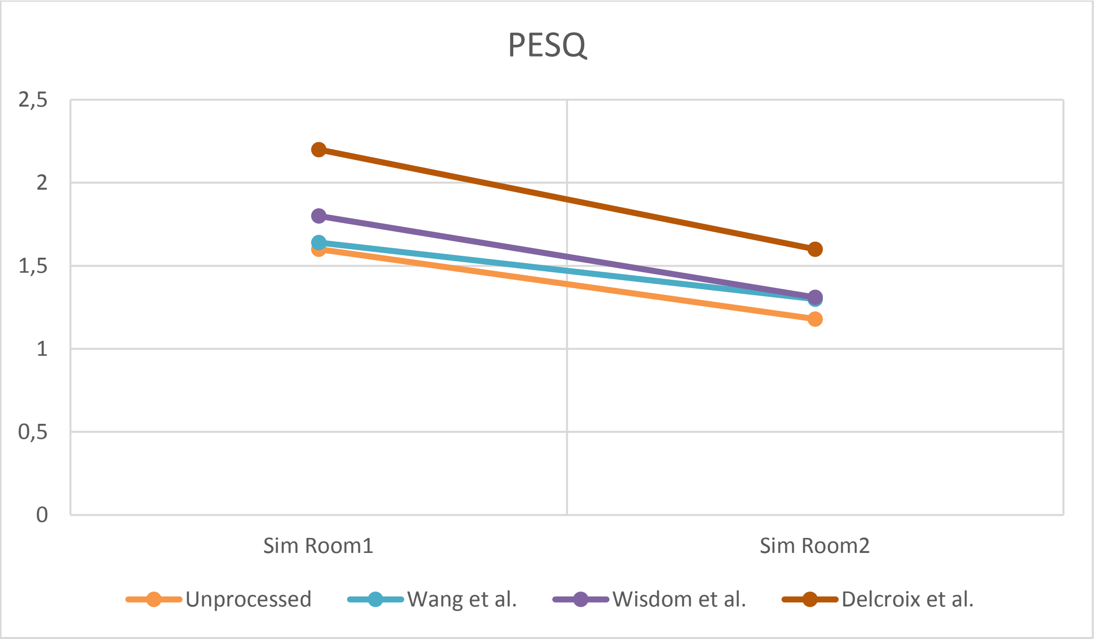


Figure 4.10. PESQ results of the two microphone solutions in the workshop.

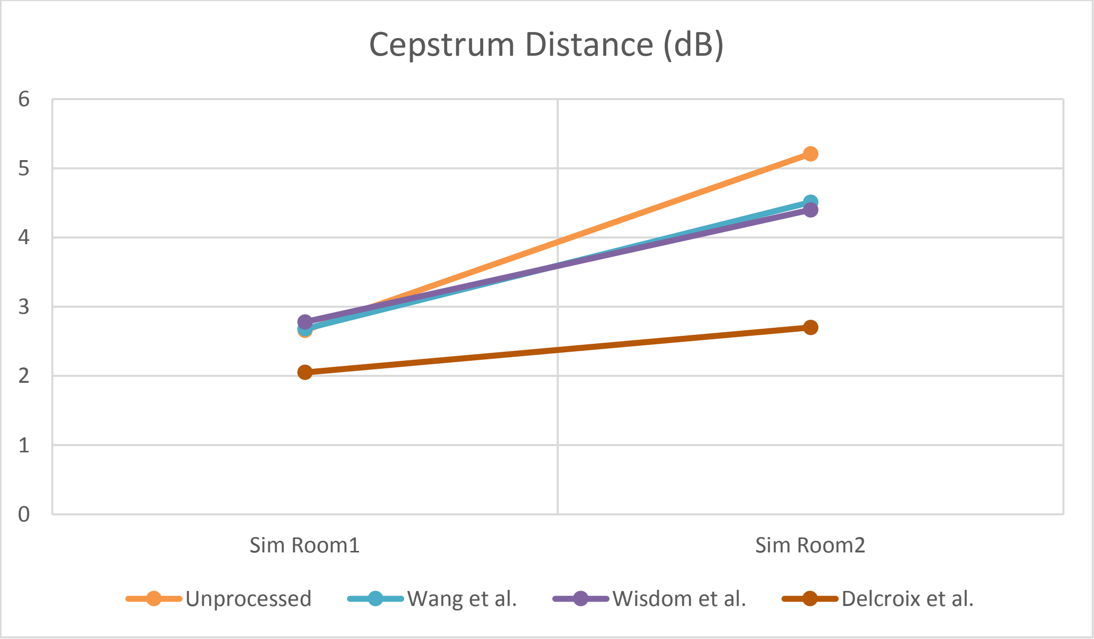


Figure 4.11. CD results of the two microphone solutions in the workshop.

Sim Room1 and Sim Room2 conditions are similar to RT_{60}= 360 msec and RT_{60}= 610 msec conditions respectively. In terms of SRMR results in Table 4.2, the proposed solution provides %20 and %36 enhancement respectively. According to Figure 4.9, %20 enhancement in Sim Room1 is better than the algorithms in the workshop in terms of SRMR. In the Sim Room2, although %36 enhancement is above the average, Delcroix's algorithm provides approximately %70 enhancement.

In terms of PESQ, the proposed algorithm provides an enhancement about %14 and %12 in Table 4.2. According to Figure 4.10, these results are better than two of the proposed algorithms. However, Delcroix provides an enhancement about %20 and %30 in terms of PESQ in Sim Room1 and Sim Room2 respectively.

Lastly, in terms of CD, the proposed solution provides an enhancement about 55 dB under similar conditions of both Sim Room1 and Sim Room2. In Sim Room1, the algorithms in the workshop apart from Delcroix's algorithm do not provide an enhancement. However, the performance of the proposed dereverberation is worse than the algorithms in the workshop in Sim Room2.

Another way to evaluate reverberation level in speech signal is spectrogram. Spectrogram of 10 sec reverberant speech ($RT_{60} = 610$ msec, θ = 45°) signal before and after the dereverberation process is shown in Figure 4.12. As it is seen, smearing effect that stems from reverberation is reduced and blurring of the speech phonemes decreases as a result of dereverberation.

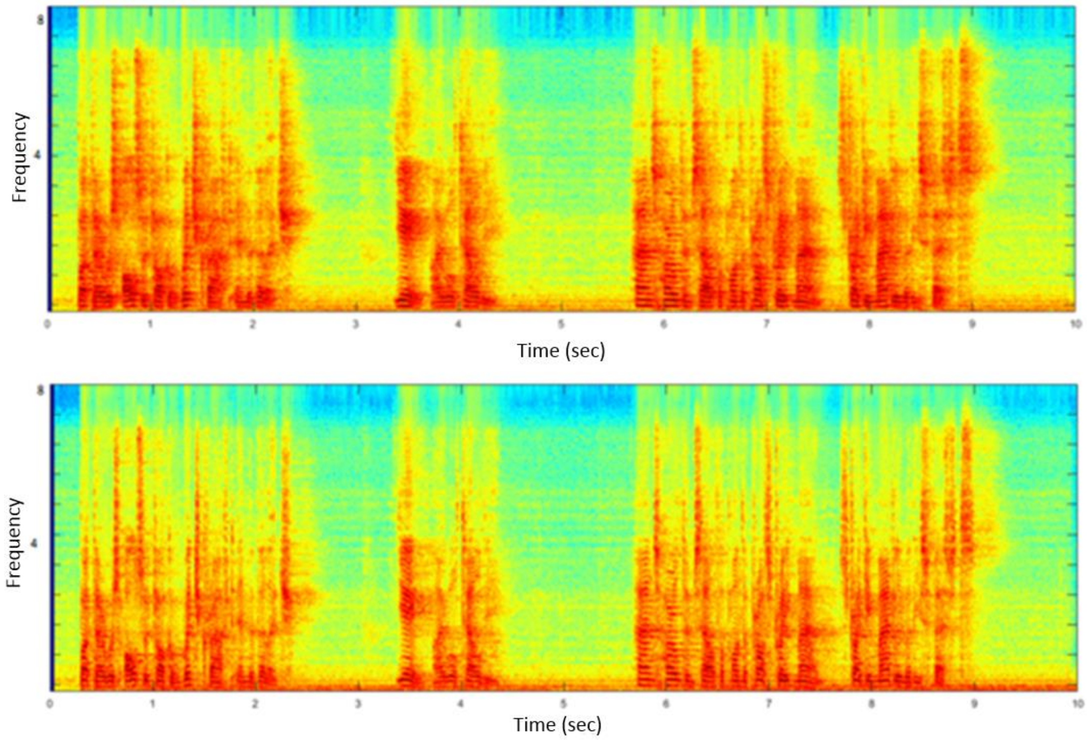


Figure 4.12. Spectrograms of unprocessed and processed speech signals respectively.

4.5.2. Discussions

This study proposes an algorithm which consists of two different stages. These stages process reverberation from two distinct approaches. Although these stages are completely different from each other, they can be combined efficiently because of shared parameters. The whole algorithm suppresses reverberant part of observed speech, and it does not cause any serious distortions in the desired speech under most of the experimental conditions according to objective measures. However, there are some shortages of the proposed model.

MVDR algorithm is not used in dereverberation normally; however, it can be used to filter late reverberation thanks to diffuse field assumption in this thesis. The assumption suffers in short distances and small reverberation time, because early components dominate observed signal. If late reverberation can be modeled in these conditions more accurately (instead of diffuse field), MVDR performance improves further. In addition, DOA cannot be found accurately in reverberant environments. This problem is solved by detecting ‘silence’ to ‘speech’ transitions. Simple voice activity detection algorithm is used to find DOA in MVDR.

Plane wave model of sounds suffers in short distances. This case causes performance of MVDR to decrease. If sound wave can be modelled more realistically in short distances, better results can be obtained. Since this brings a lot of computational complexity, plane wave model is preferred in this thesis.

Using MVDR before single channel dereverberation algorithm enhances its performance. Original WPE method is based on an iteration to find spectral power of early reverberant part [70]. However, estimating spectral power of this part in the first stage makes this iteration unnecessary. In addition, spectral power of early reverberant is found by multiple microphones thanks to this combined algorithm. In this way, more accurate estimation can be obtained.

The main shortage of the second part is mixing time. There are lots of studies about estimating mixing time. In this thesis, mixing time is found by volume of the room. Since the algorithm uses mixing time to obtain the result directly, more accurate mixing time provides better results. Also, frequency effect on reverberation time is taken consideration roughly. This reduces potential performance of the second part; however, this provides faster and more efficient algorithm. If frequency effect on reverberation time is modelled accurately without complexity, performance of the second stage improves.

In conclusion, independent algorithms are combined efficiently, and overall performance of the combined algorithm is better than each one except for RT_{60}= 160 msec. Although second algorithm provides dereverberation in this case, MVDR algorithm does not perform very well. Low reverberation time, short source distance lead worse results. Performance of the whole algorithm improves when late reverberation starts to dominate observed signal (in long distance and high reverberation time).

The samples of unprocessed and processed speech can be found at https://drive.google.com/drive/folders/1zFmjziuuP-aflWGBU5YymmYnUDv_hZKY for demonstration.

CHAPTER 5

CONCLUSION

5.1. Summary

The noise reduction techniques have been widely used in communication systems for many years. In the literature, there are a lot of state-of-the-art noise reduction techniques. Dereverberation has not been common due to lack of practical usage until a couple of years ago. However, it has attracted attention with hands-free speech interfaces because the received microphone signals are inevitably corrupted by room reverberation. Therefore, dereverberation techniques are of great interest to the tech industry recently. This thesis has addressed this problem and introduced the method which can be used in practical applications to deal with reverberation.

Deterministic prior knowledge about the system is used in most of dereverberation techniques. However, these techniques cannot be implemented in daily life applications appropriately. Therefore, the proposed method is based on suitable statistical assumptions instead of deterministic prior knowledge. Also, Gaussian distribution speech model significantly reduces computational complexity. Thus, processing time of the method is reduced. Thanks to short processing time, the proposed blind dereverberation technique can be implemented in practical applications.

The proposed algorithm consists of well-known techniques which are beamforming and linear prediction. However, they are not used directly because reverberation cannot be seen as a simple incoherent signal. Reverberation signals have to be modeled according to their characteristics and room acoustic properties.

The beamforming in the proposed approach needs transfer function ratios of early reverberation and correlation matrix of late reverberation. However, there is no established method to find these parameters. Therefore, it is necessary to investigate the underlying reasons for reverberation process. In this way, it is possible to derive the required parameters of the MVDR algorithm. In Chapter 3, these parameters are derived according to the fundamentals of reverberation.

The single channel approach is used as a post filter at the second stage of the proposed method. It suppresses the residual reverberation components at the output of MVDR algorithm. This approach is quite different from the first stage. However, the required parameters in this approach are already derived in the first stage.

In order to measure performance of the proposed algorithm, it must be tested under various conditions. The objective measures show that the proposed method provides considerable enhancement for reverberant speech signals. Furthermore, it is revealed that the algorithm does not cause any serious degradation in the desired speech with respect to objective measures in Chapter 4. The results which are taken in different directions and distances also demonstrate that the performance of method does not depend on speaker-microphone configurations.

5.2. Conclusion and Future Work

The proposed dereverberation method relies on the statistical models of reverberation and speech signal. These models are used instead of deterministic prior knowledge and they reduce computational complexity; however, these models may oversimplify the nature of signals.

STFT coefficients of the desired speech are modeled by independent Gaussian distributions. In an anechoic speech signal, there can be correlations between the frames. Therefore, a joint pdf which take account of inter-frame correlations can be more appropriate model for the STFT coefficients.

The statistical model of reverberation signals assumes late reverberation is completely independent of early reverberation. However, a statistical model which considers the correlation between these parts can be more accurate.

There is no consensus about the transition time in the literature. Transition time is significant parameter especially for the second stage. Estimating transition time accurately improve performance of dereverberation.

Although this thesis assumes that noise is ignored in the calculations, dereverberation algorithm should be robust in noisy environments. Noise may affect the estimations of reverberation parameters. Therefore, a significant effort can be devoted to dereverberation algorithms in noisy environments.

Finally, it is concluded that instead of deterministic prior knowledge about the system, statistical models can be used to process reverberation and a blind dereverberation method independent of speaker-microphone configuration can be generated with denoising algorithms. The most important thing in this goal is to find efficient models of sound signals. It is seen that when the diffuse sound model becomes more accurate (in longer distance and larger reverberation time) the performance of the dereverberation improves.

REFERENCES

[1] H. Sato and J. S. Bradley, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America,* vol. 113, no. 6, pp. 3233-44, 2003.

[2] E. A. P. Habets, "Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement," in *Ph.D. Thesis*, Technische Universiteit Eindhoven, 2007.

[3] I. Kodrasi and S. Doclo, "Signal-Dependent Penalty Functions for Robust Acoustic Multi-Channel Equalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 25, no. 7, pp. 1512-1525, 2017.

[4] J. Benesty, J. Chen, Y. Huang and I. A. Cohen, Noise Reduction in Speech Processing, Springer, 2009.

[5] E. A. P. Habets, "Speech Dereverberation Using Statistical Reverberation Models," in *Speech Dereverberation,* Springer, 2010, pp. 57-93.

[6] J. Polack, "Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics," *Applied Acoustics,* vol. 38, no. 2-4, pp. 235-244, 1993.

[7] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *17th European Signal Processing Conference*, Glasgow, Scotland, 2009.

[8] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita and M. Delcroix, "Speech dereverberation based on Maximum Likelihood Estimation with time varying Gaussian source model," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 8, pp. 1512 - 1527, 2008.

[9] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 16, no. 1, pp. 229 - 238, 2008.

[10] P. A. Naylor and N. D. Gaubitch, "Introduction," in *Speech Dereverberation*, Springer, 2010, pp. 1-21.

[11] H. Kuttruff, Room Acoustics, London: CRC Press, 2000.

[12] S. Doclo and M. Moonen, "Design of far-field and near-field broadband beamformers using eigenfilters," *Signal Processing*, vol. 83, no. 12, pp. 2641-2673, 2003.

[13] W. C. Sabine, "Collected papers on acoustics," Harvard University Press, Cambridge, 1922.

[14] J. Polack, "La transmission de l'énergie sonore dans les salles," Thesis, Université du Maine, 1988.

[15] M. R. Schroeder and K. Kuttruff, "On Frequency Response Curves in Rooms. Comparison of Experimental, Theoretical, and Monte Carlo Results for the Average Frequency Spacing between Maxima," *Acoustical Society of America*, vol. 34, no. 1, pp. 76-80, 1962.

[16] J.-M. Jot, O. Warusfel and L. Cerveau, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," *The Journal of the Acoustical Society of America*, vol. 99, no. 4, p. 2530, 1997.

[17] P. A. Naylor, E. A. Habets, J. Y.-C. Wen and N. D. Gaubitch, "Models, Measurement and Evaluation," in *Speech Dereverberation*, Springer, 2010, pp. 21-57.

[18] M. R. Schroeder, "New method of measuring reverberation time," *Acoustical Society of America*, vol. 37, no. 6, pp. 409-412, 1965.

[19] J. Allen, D. Berkley and J.Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Acoustical Society of America*, vol. 62, no. 4, pp. 912-915, 1977.

[20] M. Kajala and M.hamalainen, "Filter and sum beamformer with adjustable filter," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA, 2001.

[21] J. Flanagan, J. Johnston, R. Zahn and G. Elko, "Computer steered microphone arrays for sound transduction in large rooms," *Acoustical Society of America,* vol. 78, no. 5, p. 1508, 1985.

[22] Y. Yamamoto and Y. Haneda, "Spherical microphone array post-filtering for reverberation suppression using isotropic beamformings," in *IEEE International Workshop on Acoustic Speech Enhancement,* Xi'an, China, 2016.

[23] O. Frost, "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE,* vol. 60, no. 8, pp. 926-935, 1972.

[24] L. Griffiths and J. Charles, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation,* vol. 30, no. 1, pp. 27-34, 1982.

[25] S. Gannot, D. Burshtein and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing,* vol. 49, no. 8, pp. 1614 - 1626, 2001.

[26] M. Hoffman, M. Link and K. Buckley, "Desired speech signal cancellation by microphone arrays in reverberant rooms," in *Twenty Fifth Asilomar Conference on Signals, Systems & Computers,* Pacific Grove, CA, USA, 1991.

[27] T. Dietzen, N. Huleihel, S. Doclo, M. Moonen, T. v. Waterschoot, A. Spriet and W. Tirry, "Speech dereverberation by data dependent beamforming with signal pre-whitening," in *23rd European Signal Processing Conference,* Nice, France, 2015.

[28] O. Schwartz, S. Gannot and E. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 2, pp. 240 - 251, 2015.

[29] T. Dietzen, S. Doclo, M. Moonen and T. v. Waterschoot, "Joint multi-microphone seech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC),* Tokyo, Japan, 2018.

[30] D. Bees, M. Blostein and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing,* Toronto, Ontario, Canada, 1991.

[31] J. Flanagan and R. Lummis, "Signal processing to reduce multipath distortion in small rooms," *Acoustical Society of America,* vol. 47, no. 6A, p. 1475, 1970.

[32] J. Erkelens and R. Heusdens, "Correlation based and model based blind single channel late reverberation suppression in noisy time varying acoustical environments," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 7, pp. 1746 - 1765, 2010.

[33] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Speech and Audio Processing,* vol. 8, no. 3, pp. 267 - 281, 2000.

[34] B. Gillespie and H. Malvar, "Speech dereverberation via maximum kurtosis subband adaptive filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* Salt Lake City, UT, USA, 2001.

[35] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 36, no. 2, pp. 145 - 152, 1988.

[36] E. Moulines, P. Duhamel, J.-F. Cardoso and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Transactions on Signal Processing,* vol. 43, no. 2, pp. 516 - 525, 1995.

[37] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP Journal on Applied Signal Processing,* pp. 1074-1090, 2003.

[38] K. Furuya and A.Kataoka, "Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 5, pp. 1579 - 1591, 2007.

[39] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *2003 IEEE International Conference on*

Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, China, 2003.

[40] H. Attias, J. C. Platt, A. Acero and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Neural Information Processing Systems*, 2001.

[41] T. Nakatani, B.-H. Juang, K. Kinoshita and M. Miyoshi, "Speech dereverberation based on probabilistic models of source and room acoustics," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Toulouse, France, 2006.

[42] S.Keronen, K. Palomaki, H. Kallasjoki, G. Brown and J. Gemmeke, "Feature enhancement of reverberant speech by distribution matching and non-negative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, 2015.

[43] N. Lopez, Y. Grenier, G. Richard and I. Bourmeyster, "Single channel reverberation suppression based on sparse linear prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014.

[44] K. Lebart, J. M. Boucher and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359-366, 2001.

[45] J. Benesty, "Conventional beamforming techniques," in *Microphone Array Processing*, Springer, 2008, pp. 39-65.

[46] P. Wei, S. Liu and Z. Du, "Analysis of spatial sampling characteristics for the circular array," *Advances in Information Sciences and Service Sciences*, vol. 5, no. 7, pp. 666-676, 2013.

[47] W. Davenport, "A study of speech probability distributions," Research Laboratory of Electronics, Cambridge, Massachusetts, 1950.

[48] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443 - 445, 1985.

[49] S. Gazor and W. Zhang, "Speech probability distributions," *IEEE Signal Processing Letters,* vol. 10, no. 7, pp. 204 - 207, 2003.

[50] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using laplacian speech priors," in *International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, 2003.

[51] J. Jensen, C. Hendriks, I. Batina and H. Richard, "A study of the distribution of time domain speech samples and discrete fourier coefficients," in *IEEE BENELUX/DSP Valley Signal Processing Symposium*, 2005.

[52] K. Meesawat and D. Hammershøi, "An investigation on the transition from early reflections to a reverberation tail in a BRIR," in *International Conference on Auditory Display*, Kyoto, Japan, 2002.

[53] A. Jukić, T. v. Waterschoot, T. Gerkmann and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches," in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 2014.

[54] N. Kumar and A. Singh, "Effect of reverberation on different DOA estimation techniques using microphone array," *International Journal of Science Technology & Engineering,* vol. 3, no. 7, pp. 166-170, 2017.

[55] A. Vesa, "Direction of arrival estimation using MUSIC and root," in *18th Telecommunications Forum,* Hong Kong, 2010.

[56] S. Gannot and E. A. P. Habets, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America,* vol. 122, no. 6, pp. 3464-70, 2007.

[57] E. A. P. Habets, S. Gannot and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters,* vol. 16, no. 9, pp. 770-773, 2009.

[58] E. Habets, "Towards multi-microphone speech dereverberation using spectral enhancement and statistical reverberation models," in *42nd Asilomar Conference on Signals, Systems and Computers,* Pacific Grove, CA, USA, 2008.

[59] H. W. Lollmann, E. Yilmaz, M. Jeub and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, 2010.

[60] L. Beranek, Concert Halls and Opera Houses, Cambridge: Springer, 2004.

[61] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT, USA.

[62] K. Kondo, Subjective Quality Measurement of Speech, Springer, 2012.

[63] T. H. Falk, C. Zheng and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766 - 1774, 2010.

[64] S. Gannot, "Multichannel audio database in various acoustic environments," in *14th International Workshop on Acoustic Signal Enhancement*, Juan-les-Pins, France, 2014.

[65] L. Rabiner and R. Schafer, Digital Processing of Speech Signal, New Jersey: Prentice - Hall, 1978.

[66] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj and A. Sehr, "A summary of the REVERB challenge:state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal*, no. 1, pp. 1-19, 2016.

[67] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto and N. Ito, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the Reverb Challenge," in *Reverb Challenge Workshop 2014*.

[68] X. Wang, X. Yang, Y. Guo and Q. Fu, "Acoustic scene aware dereverberation using 2-channel spectral enhancement for Reverb Challenge," in *Reverb Challenge Workshop 2014*.

[69] S. Wisdom, T. Powers, L. Atlas and J. Pitton, "Enhancement of reverberant and noisy speech by extending its coherence," in *Reverb Challenge Workshop 2014*.

[70] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717 - 1731, 2010.