Master Thesis in

Electrical Engineering With Emphasis on Signal Processing

Microphone Array Wiener Beamforming with modeling of SRP- PHAT for Speaker Localization

SRISESHUKUMAR BASAVA

This thesis is presented as a part of Degree of Master of Science in Electrical Engineering with Emphasis on Signal Processing

Blekinge Institute of Technology

January-2012

Blekinge Institute of Technology

School of Engineering,

Department of Electrical Engineering,

Supervisor : Dr. Nedelko Grbić

Examiner : Dr. Benny Sällberg

BLEKINGE TEKNISKA HÖGSKOLA

SE-371 79 KARLSKRONA, SWEDEN

TEL. 0455-385000, FAX. 0455-385057.

Contact Information:

Author: Sriseshukumar Basava

Email: srba10@student.bth.se

Supervisor :

Dr. Nedelko Grbić

Department of Electrical Engineering

School of Engineering,

Blekinge Institute of Technology, Sweden

Email: nedelko.grbic@bth.se

Examiner :

Dr. Benny Sällberg

Department of Electrical Engineering

School of Engineering,

Blekinge Institute of Technology, Sweden

Email: benny.sallberg@bth.se

ABSTRACT

The use of microphone arrays to acquire and recognize speech in meetings (conference) poses several problems for speech processing as there exist many speakers within a small space, typically around a table. The necessity to design a suitable microphone array system with minimum noise and more efficient localization algorithms is drawing attention of researchers to work on it. Extensive research is being carried out on Microphone Array Beamforming to make the system, robust, viable and elegant for commercial use. This study is done with a similar objective.

A system consisting of 4 microphones arranged in linear array is setup in a simulated reverberant environment. Filter-and-sum beam forming is implemented both in time domain and frequency domain. A Wiener filter is chosen as post filtering technique. One of the main goals of the thesis is to improve the quality of the primary speech signal based on microphone array with Wiener beam forming (filter-and-sum beam forming with wiener post filtering). Weighted over lap add (WOLA) filter bank is also implemented as a part of frequency domain wiener beam forming to make use of subband beam forming. Also RLS algorithm is used to make the subband beamforming adaptive.

Speaker localization plays a pivotal role in the development of speech enhancement methods requiring information of the speaker position. Among many localization algorithms, Steered Response Power (SRP) with a combination of Phase Alignment Transform (PHAT) called SRP-PHAT has proved to be a robust one in many studies. Also as a part of this project, modeling of SRP-PHAT for detecting the speaker position for the above described system is done.

To evaluate the system performance, Signal-to-Noise-Ratio (SNR) is calculated for both original and beam formed signals. Perpetual Evaluation of Speech Quality (PESQ), an International Telecommunication Union (ITU-T) standard for evaluating quality in speech signals is used for determining the Mean opinion Score (MOS) for both the original and the beam formed signals.

ACKNOWLEDGEMENT

To begin with, I would like to thank BTH and JNTUK for their Double Degree Program to which I have been admitted.

I would like to express my gratitude to Dr. Nedelko Grbić for his inspiring support and guidance throughout this work. His constant encouragement has been a major role in successful completion of thesis.

I also thank my examiner Dr. Benny Sällberg for his constructive comments while evaluating this thesis.

I would like to express my appreciation for the endless hours of discussion, technical and otherwise, that I had with my friends Vamsy, Rajesh and Hemanth during this work. Their support has been a great advantage for me especially in learning MATLAB programming.

Finally I would like to express my gratitude to my parents who have always been there for me throughout my good and bad times, always encouraging me and for making me who I am. I also thank my family and friends for their affection and encouragement that they have provided during my studies at BTH.

Sriseshukumar Basava

Karlskrona, Jan 2012.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TDOA	Time Difference of Arrival
SRP	Steered Response Power
PHAT	Phase Alignment Transform
WOLA	Weighted Overlap Add
GCC	Generalized cross correlation
SNR	Signal to noise ratio
PESQ	Perpetual Evaluation of Speech Quality
LCMV	Linearly Constrained Minimum Variance
MVDR	Minimum Variance Distortion-less Response
GSC	Generalized Side lobe Canceller
RLS	Recursive Least Squares
SC-RLS	Soft-constrained Recursive Least Squares
IIR	Infinite Impulse Response
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
FIFO	First In First Out
BFP	Block-floating-point
LMS	Least Mean Square
PSQM	Perpetual Speech Quality Measure
ITU	International Telecom Union
MOS	Mean Opinion Score
VAD	Voice Activity Detection
AWN	Additive White Noise

CHAPTER 1 INTRODUCTION

1.1 Motivation

Signal processing has been a major part in the development of advanced technology in communications like teleconferencing, mobile communication etc. Speech enhancement is one of the methods of signal processing intended towards speech processing. The need for intelligible technology is growing along with the advancement of technology. For example, teleconferencing which involves speaker detection and localization, noise cancellation, speech enhancement, attenuating the low grade speech etc, is an intelligible technology because it has to decide on its own, what is necessary and also ensure robust performance.

Microphone array, in which many microphones are placed at different spatial locations, is a key tool in applications like teleconferencing, hands free communication, enhancement or suppression of the received signal, noise reduction, optimal filtering, source separation and speaker tracking using different methods. Using multiple microphones allows spatial sampling as the arrival time of signal will be different for each microphone. With the increased development in speech processing technologies, effective speech communication has drawn attention of many researchers who came up with newer methods to utilize various microphone array configurations to extract reliable and intelligible speech over the last three decades. The most typical methods are beamforming or blind source separation [1, 2].

The inherent ability of microphone arrays to exploit the spatial correlation of multiple received signals has enabled the development of combined temporal and spatial filtering algorithms known as beamforming [3]. Beamforming or spatial filtering has its roots in narrowband applications where, by adjusting the time-delays of each sensor for a particular direction of acoustic source in noisy environment and then summing them. The direction of interest is called look-direction. The signals from the desired source are added in constructive way while interfering signals are dealt with destructive manner. Hence, Beamforming is often used for removing noise and reverberation from speech signals by taking advantage of spatial information. Beamforming techniques can be broadly classified as being either data-independent, or data-dependent. Data-independent, or fixed, beamformers are so named because their parameters are fixed during operation. Conversely, data-dependent, or adaptive, beamforming techniques continuously update their parameters based on the received signals.

For these reasons, Microphone array beamforming became an integral part of far-field communication in which microphone array will be at considerable distance from the speaker and yet meant optimal performance. Due to their ability to provide hands-free acquisition and directional discrimination, microphone arrays present a potential alternative to close-talking microphones. The motivation for steerable microphones comes mainly from teleconferencing and car telephony applications. The difference in these two applications is that in the car environment, we usually deal with much lower Signal to Noise Ratio's, while in the teleconferencing environment, we usually have to change the beam direction more often, as well as requiring larger spatial coverage. Currently, there are ways to use many microphones to create beam patterns that will focus on one speaker in a room. For this purpose, we require source localization strategies to locate the speaker by steering the microphone towards the direction of speaker.

Sound source localization using microphone arrays has a wide variety of applications including talker tracking, human computer interaction (HCI) and robotics [4]. Different methods based on steered beamformers, high resolution spectral estimation and time difference of arrival (TDOA) are used for sound source localization [5]. Localization strategies based on one of these methods have limited applications as they are either computationally expensive or less robust to reverberant and noisy conditions. Steered response power (SRP) algorithm is a localization algorithm based on steered beamformers and TDOA methods. The algorithm uses filter and sum beamforming operation. The microphone signals received are time aligned by applying suitable time shifts and their correlation terms are summed together to obtain the steered response power. Performance of SRP algorithm under coherent noise conditions can be improved by using phase transform (PHAT) [6].

In a closed room, the sound at the microphone arrives not only directly from the source, but also because of multiple reflections from the walls of the room. This phenomenon, which is very common in conference rooms and classrooms, is called reverberation. The presence of a significant amount of reverberation can severely degrade the performance of TDOA estimation algorithms. The motivation for this thesis comes from the need to find reliable algorithm which can locate and track a single speaker in a reverberant room using an array of microphones with an enhanced speech output.

1.2 Objective

The problem with microphones is that they not only capture the intended speech signal but also capture all acoustic sounds that are in the range of the microphones. All the unwanted acoustics sounds are referred as noise and interferences. When more than one source is active, each microphone records an additive mixture of 1) uncorrelated background noise, 2) direct speech signals from the sources 3) correlated echoes of the sources. We have many methods to eliminate these undesired acoustic sounds and enhance the desired signals under the heading "Speech Enhancement Techniques".

The main objective of this thesis is to present different approaches of the Beamforming, verify the best algorithm in terms of the reverberant and noisy environment in both time domain and frequency domain and finally making it adaptive. Also to present source localization strategies like SRP-PHAT and locate the position of sound source along with direction of arrival (DOA) in the same reverberant and noisy environment.

1.3 Overview of thesis

This thesis work is carried out in 4 different stages. First the simulation of reverberant room is done. For this purpose, image method is used to get Room Impulse response (RIR). When a speech signal is convolved with this RIR, the resultant signal is the desired speech signal with reverberant effect.

Secondly, a time domain wiener beamformer is implemented with reverberated speech signal and noise signal as inputs. This is followed by a frequency domain wiener beamformer implementation which requires a filter bank. Weighted Overlap Add (WOLA) filter bank is chosen.

In the next stage, adaptive beamforming is implemented with the help of weighted Recursive Least Squares (RLS) algorithm. Finally, SRP-PHAT source localization strategy is implemented and tested.

1.4 Organization of thesis

In chapter 2, Room Impulse Response which is used for simulation of reverberant environment is discussed. Also acoustic signal modeling is explained briefly.

Chapter 3 describes the problem that motivated this thesis: the degradation caused by the use of far-field microphones in speech processing applications. Microphone array signal processing is presented as an alternative to solve these problems. Hence, an overview of the fundamentals of array signal processing theory and the main particularities of microphone array signal processing is provided as the basic background for the following chapters. Various beamforming techniques are also briefly explained.

Chapter 4 describes the efficient sub-band beamforming technique which requires implementation of filter bank. Hence WOLA filter bank is also discussed in this section. Also the implementation procedure of adaptive RLS approach for frequency domain is discussed. With this Adaptive beamforming, better results were obtained.

Chapter 5 introduces the source location strategies. Generalized cross correlation (GCC) algorithm with Phase Transform (PHAT) and SRP-PHAT are explained and implemented.

In chapter 6, Signal to noise ratio (SNR), Perpetual Evaluation of Speech Quality (PESQ) are discussed. These are the parameters used in this thesis to assess the speech quality after the processing of speech is done.

In Chapter 7, overall results are presented along with appropriate explanations followed by conclusion and potential for future work in chapter 8.

CHAPTER 2 ACOUSTIC ROOM MODELING

2.1 Introduction

Reverberation is one of the major factors that affect multi-channel equalization performance [7]. It is defined as the persistence of sound in a particular space after the original sound is removed. The conference room has many objects in its surrounding which becomes a cause of reverberation. These objects could be the furniture, white board and walls of the conference room. This reverberation can severely affect the performance of the speech processing algorithm used for speaker localization. This phenomenon can be observed when the sound source stops even as the reflections continue, decreasing in amplitude, until they can no longer be heard. The length of this sound decay, or reverberation time, receives special consideration in the architectural design of conference rooms which need to have specific reverberation times to achieve optimum performance.

To predict this energy decay, in 1979, Allen and Berkley proposed a method called image source method [8] to simulate the room acoustics. From then, this method has been used by many researchers to create ample virtual acoustic environment called Room Impulse Response (RIR). With the RIR function, one can choose parameters like size of the room, reflection coefficient, position of microphone, source position etc., depending upon their requirement.

The image source method can also be used to simulate the reverberation in a conference room for a given source and microphone location and to compute a Finite Impulse Response (FIR) that models an acoustic channel between the source and microphones in reverberant conditions. Remember that calculating FIR can only be done with discrete time impulses which can be achieved by impulse response function. In this thesis, First image method is used to create reverberant room and then the unit impulse of each echo is calculated accordingly with a fractional time delay. For this purpose an all pass Thiran filter is used [9]. Next the magnitude of each impulse is calculated and all the data is put together into a one dimensional function called Room impulse response.

In the next section, the image method is discussed followed by fractional delay filter to find accurate time delay of the virtual room.

2.2 Image model

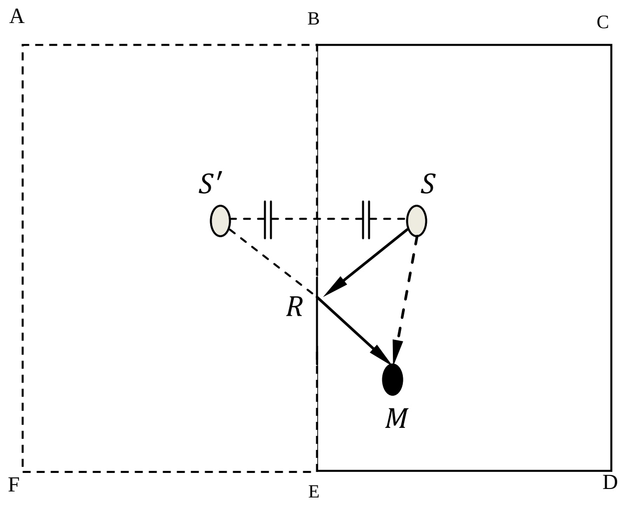

Fig. 2.1. Path involving one reflection with one image

A simple image model is explained in Fig. 2.1. The area under BCDE is the actual room and the area ABEF represents its mirror image. Let the source S be located at some position in the room and M be the microphone. Also assume that the source S and microphone M is separated by a distance d . The line SM represents the direct path between S and M and the path length can be calculated from the known locations of the source and the microphone. Now in image section ABEF, a source image S’ is formed at the same distance from the wall as that of Source S as shown in the figure. Let R be the reflection point. Because of symmetry in mirror image, the triangle SRS’ is isosceles and therefore the path length SR + RM is the same as S’D. Therefore to compute length of the reflected path, it is enough to compute the distance between microphone and source image. So, whenever we are calculating the distance using source image, it is implied that there is a reflection in the path.

In this way we can calculate the distance for any number of reflections. In Fig 2.2, a dual virtual source reflection model is presented. This can be extended to ‘n’ virtual sources with a relation (2*n+1) ^3. We can visualize this scenario by folding a piece of paper and making a hole in it and when you reopen you can see many holes on the same paper. If you assume one hole as a source, then the remaining will be images of the source.

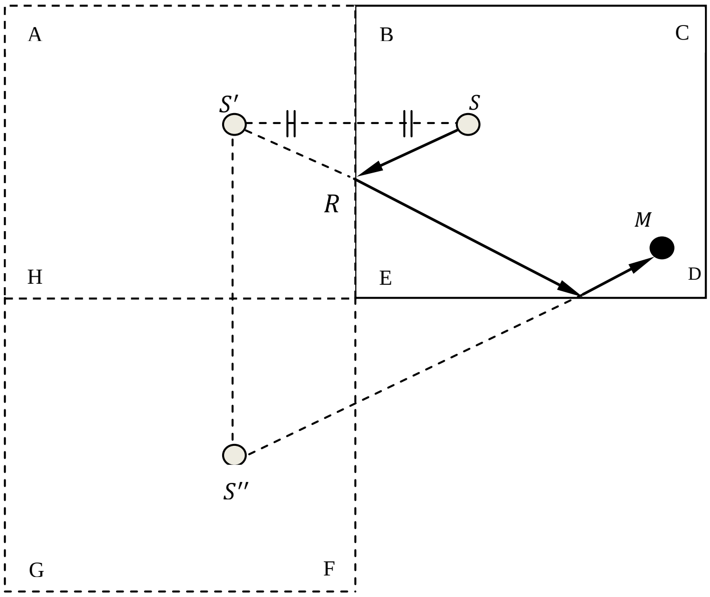

Fig. 2.2. Path involving two reflections with two virtual sources

2.3 Image Method

Consider a rectangular room which has dimensions like length (l), width (w) and height (h). Let x_s be the distance of sound source, x_m be the distance of the microphone and x_r is the length of the room with respect to origin O as shown in the figure below.

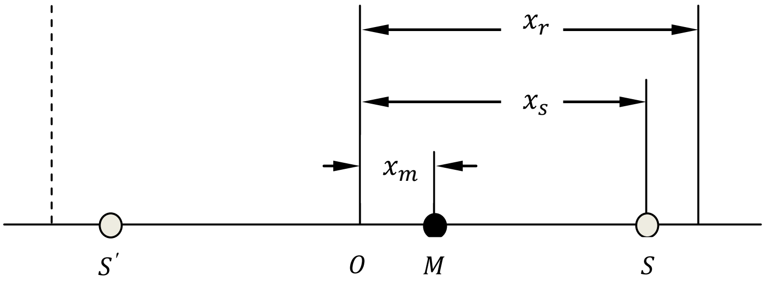

Fig. 2.3. Model of Image method in one dimension

The x-coordinate of i^{th} virtual source, x_i can be expressed using the following equation.

$$x_i = (-1)^i x_s + \left[i + \frac{1-(-1)^i}{2} \right] x_r \qquad (2.1)$$

When i is a negative number, then the virtual source will be located on the negative X-axis elsewhere virtual source will be on positive X-axis. The distance between the i^{th} virtual sound source and microphone is calculated by subtracting the microphone's x-coordinate x_m, fromx_i i.e.

$$x_i = (-1)^i x_s + \left[i + \frac{1-(-1)^i}{2}\right] x_r - x_m \qquad (2.2)$$

In a three dimensional setup, we have X, Y and Z axes. We can find distance of the virtual sources from the microphone with respect to the Y and Z axes using the equations 2.3 and 2.4 respectively.

$$y_j = (-1)^j y_s + \left[j + \frac{1-(-1)^j}{2}\right] y_r - y_m \qquad (2.3)$$

$$z_k = (-1)^k z_s + \left[k + \frac{1-(-1)^k}{2}\right] z_r - z_m \qquad (2.4)$$

The Euclidean distance of each virtual source x_i, y_j and z_k is calculated according to Pythagoras theorem and this will be a three dimensional matrix i.e.,

$$d_{ijk} = \sqrt{x_i^2 + y_j^2 + z_k^2} \qquad (2.5)$$

2.3.1 Unit impulse response function of each virtual source

Assume $a_{ijk}(t)$ be the desired impulse response and it is calculated using equation 2.6

$$a_{ijk}(t) = t - \frac{d_{ijk}}{c} \qquad (2.6)$$

Where t is the time, d_{ijk} is the Euclidean distance and c is the speed of the sound. The term $\frac{d_{ijk}}{c}$ is the time delay of each echo.

Therefore, the unit impulse response u_{ijk} can be expressed as a function of $a_{ijk}(t)$ i.e,

$$u_{ijk}(a) = \begin{cases} 1, & \text{if } a_{ijk}(t) = 0 \\ 0, & \text{otherwise} \end{cases} \qquad (2.7)$$

2.3.2 Magnitude of unit impulse response

The magnitude of unit impulses of virtual sources is affected mainly by the distance the sound wave travels from the source to the microphone and the number of reflections the sound wave makes while it is transmitted i.e. reflection coefficient of the room.

If the room has uniform reflection coefficient r_w, then reflection factor of virtual sources will be,

$$r_{ijk} = r_w^n \qquad (2.8)$$

Where n represents the total number of reflections that the sound wave has undergone and it is given by adding all the virtual sources $n = |i| + |j| + |k|$.

Now the total magnitude of each echo is calculated by multiplying the equations 2.6 and 2.8 together.

$$e_{ijk} = f_{ijk} r_{ijk} \qquad (2.9)$$

Where, f_{ijk} is a function that varies inversely with d_{ijk} i.e., $f_{ijk} \propto \frac{1}{d_{ijk}}$

2.4 Room Impulse Response (RIR)

The room impulse is obtained by multiplying equations 2.7, 2.9 and summing over with all the three indices i, j, k as shown in the equation here under

$$h_{RIR}(t) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} \sum_{k=-n}^{n} u_{ijk} e_{ijk} \qquad (2.10)$$

2.5 Fractional time delay

The manual delay made with Matlab by creating Deltas is very simple and useful to test the system. Unfortunately it only allows delays in integer values. For example, a signal can be delayed three samples or four samples, but not three and a half. Thus the simulations are not as reliable as they could be. In order to obtain higher accuracy, a fractional delay all pass filter is implemented. By using the fractional delay filter, the time delay in room can be assessed for fractional values instead of rounding to nearest integer. Thus it is possible to obtain greater accuracy by using all-pass filter. In this thesis, the RIR function has the advantage of having a fractional delay all-pass filter.

The design of fractional delay all pass filters is usually based on solving a set of linear equations. The maximally flat group delay method [5] that is based on Thiran's all pole filter design [10] is the only one Fractional Delay all-pass filter design method that can be implemented using closed-form formulas but it is limited to excellence only on a narrow band at low frequencies. The magnitude response of the ideal fractional delay element should be perfectly flat irrespective of reflection coefficients so we choose all-pass filter for this purpose [11].

A discrete time all pass filter has a transfer function as cited below.

$$A(z) = \frac{z^{-N}D(z^{-1})}{D(z)} = \frac{a_N + a_{N-1}z^{-1} + \cdots + a_1 z^{-(N-1)} + z^{-N}}{1 + a_1 z^{-1} + \cdots + a_{N-1}z^{-(N-1)} + a_N z^{-N}} \quad (2.11)$$

Where N is the order of the filter and the filter coefficients $a_k(\mathrm{k} = 1,2, \dots \dots \dots, \mathrm{N})$ are real.

Later Thiran (1971) proposed an analytic solution for the coefficients of an all-pole low pass filter with a maximally flat group delay response at the zero frequency.

$$a_k = (-1)^k \frac{N!}{k!\,(N-k)!} \prod_{n=0}^{N} \frac{D-N+n}{D-N+k+n} \quad for\ k = 0,1,2, \dots \dots, N \qquad (2.12)$$

Where D refers to the actual delay and N represents the number of samples. Thiran’s proof of stability implies that this all pass filter will be stable when D > N. If D > N, the poles are inside the unit circle in the complex plane. In this case, the filter is stable. Since the nominator is a mirrored version of the denominator, the zeroes lie outside the circle. For the same reason, the radii of the poles and the zeroes are inverse of each other. That makes the amplitude response flat.

2.5 Acoustic signal modelling

Reverberation

The impact of the reverberation and background noise on the speech signal at n^{th} microphone can be modeled as

$$x_n(t) = s(t) * h(d_s, t) + n_n(t) \qquad (2.13)$$

Where $s(t)$ is the source signal, $n_n(t)$ is the background and channel noise and $h(d_s, t)$ is the room impulse response. The room impulse response varies due to temperature and humidity but its characteristics remain same for a short period of time, which makes the response time-invariant. The signal $x_n(t)$ received by n^{th} microphone can be used to localize the speaker in a reverberant and noisy environment.

CHAPTER 3 MICROPHONE ARRAY PROCESSING

3.1 Introduction

As discussed in the previous chapter, speech signals captured by a microphone located away from the sound source can be corrupted by additive noise and reverberation. One method of reducing the signal distortion and improving the quality of the signal is to use multiple microphones rather than a single microphone. By using an array of microphones rather than a single microphone, we are able to achieve spatial selectivity, reinforcing sources propagating from a particular direction, while attenuating sources propagating from other directions.

Array processing refers to the joint processing of signals captured by multiple spatially-separated sensors such as microphones. More recently, the demand for hands-free speech communication and recognition has increased and as a result, newer techniques have been developed to address the specific issues involved in the enhancement of speech signals captured by a microphone array.

This “spatial selectivity” varies as a function of frequency. A linear array generally has a wide beam width at low frequencies, which narrows as the frequency increases. An array of microphones essentially samples the sound field at different points in space which results in spatial analog of temporal aliasing that occurs when signals are sampled too slowly. When spatial aliasing occurs, the array is unable to distinguish between multiple angle of arrivals for a given frequency.

Aliasing

Spatial sampling can produce aliasing in an analogous manner to temporal sampling of continuous-time signals [12].To prevent spatial aliasing in linear arrays, the spatial sampling theorem must be followed, which states that if λ_{min} is the minimum wave length of interest and d is the microphone spacing, then $d < \lambda_{min}/2$. Hence in the array of microphones the distance between them should meet above criteria to avoid aliasing.

3.2 Microphone array processing for speech enhancement

3.2.1 Beamforming

The concept of algorithmically steering the main lobe or beam of a directivity pattern in a desired direction is called beamforming. The direction the array is steered is called the look direction. Beamforming or spatial filtering is one the simplest method for discriminating between signals based on the physical location of source and is used for directional transmission or reception of signals. During the transmission, the beamformer controls the phase and amplitude of the signal at each transmitter in order to obtain the pattern of the constructive and destructive interference. At the receiving side, information from different sensors are combined together to obtain a desired radiation pattern. In a typical conference room, the desired signal originates from the source, and is corrupted by interfering noise signal before reaching the microphones. By exploiting beamforming technique, microphone array attempts to obtain a high-quality speech signal especially in the far field communication.

Beamforming is used in wide variety of array processing algorithms which require signal capturing ability in a particular direction. It also finds use in communication applications like radars, sonar and also in medical engineering. Beamforming consists of combining microphone output, convolved with optimal weights and added to get a "beam" in direction of interest. This beam makes the array a highly directive microphone. The arbitrarily placed sensors together work as a microphone array to spatially sample a sound wave targeted on them. All beamforming techniques depend on the directivity pattern of the desired signal. Various beamforming techniques were briefly described in next section.

3.2.2 Types of beamforming.

3.2.2.1 Classical beamforming.

Delay-sum Beamforming

The simplest of all microphone array beamforming techniques is delay-sum beamforming. In order to steer an array of arbitrary configuration and number of sensors, the signals received by the array are first delayed to compensate for the path length differences from the source to the various microphones and then the signals are combined together. Fig 3.1. shows the basic structure of delay and sum beamforming.

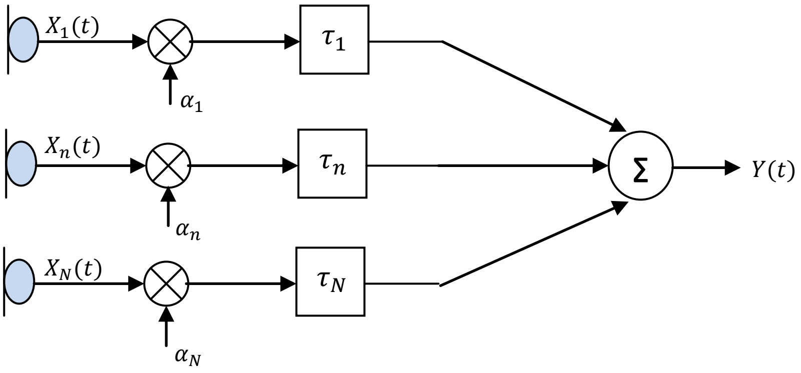

Fig 3.1. Structure of Delay and Sum Beamforming

By applying phase weights to the input channels, we can steer the main lobe of the directivity pattern to a desired direction. Considering the horizontal directivity pattern, if we use the phase weights

$$\varphi_n = \frac{-2\pi(n-1)d\cos\phi' f}{c} \tag{3.1}$$

Then the directivity pattern in this case becomes,

$$D(f,\phi) = \sum_{n=1}^{N} e^{j\frac{2\pi f(n-1)d(\cos\phi - \cos\phi')}{c}} \tag{3.2}$$

Such that an angular shift with angle ϕ' of the directivity pattern's main lobe is accomplished. Usually, each channel is given an equal amplitude weighting in the summation, so that the directivity pattern demonstrates unity gain in the desired direction. This leads to the complex channel weights

$$w_n(f) = \frac{1}{N} e^{j\frac{-2\pi f}{c}(n-1)\ d\ \cos\phi'} \tag{3.3}$$

Expressing the array output as the sum of the weighted channels we obtain

$$y(f) = \frac{1}{N}\sum_{n=1}^{N} x_n(f) e^{j\frac{-2\pi f}{c}(n-1)\ d\ \cos\phi'} \tag{3.4}$$

Where $x_n(f)$ is the frequency representation of sound wave received by n^{th} microphone. N is the total number of microphones in the array, *c* is the velocity of sound (340m/s), *d* is the distance between microphone and source. The negative phase shift in the frequency domain can effectively be implemented by applying time delays to the sensor inputs. Equivalently, in the time domain we have

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t-\tau_n) \qquad (3.5)$$

Where τ_n is the delay for the n^{th} sensor and is given by

$$\tau_n = \frac{(n-1)d\ \cos\emptyset'}{c} \qquad (3.6)$$

Which is the time taken by the wave in the plane to travel between the reference microphone and n^{th} microphone. The process of finding the delays is known as time-delay estimation (TDE) and is closely related to the problem of source localization. Many TDE methods exist in the literature, and most are based on cross-correlation [1].

Filter-sum Beamforming

In filter-and-sum beamformers, both the amplitude and phase weights are frequency dependent. The filter-and-sum beamformer can be generalized to alter-and-sum beamformer where rather than a single weight, each microphone signal has an associated filter and the captured signals are filtered before they are combined. The filtered channels are then summed, according to

$$y(f) = \sum_{n=1}^{N} w_n(f)x_n(f) \qquad (3.7)$$

The multiplications in the frequency-domain signals are accordingly replaced by convolutions in the discrete-time domain. The discrete-time output signal is hence expressed as

$$y(t) = \sum_{n=0}^{N-1}\sum_{p=0}^{P-1} h_n(p)x_n(t-p-\tau_n) \qquad (3.8)$$

Where $h_n(p)$ is the p^{th} tap of the filter associated with n^{th} microphone. Clearly, delay-and-sum processing is simply filter-and-sum with a 1-tap filter for each microphone.

The above equation can be rewritten using matrix notation for simplified expression as.

$$y(f) = W(f)^T X(f) \qquad (3.9)$$

Where the weight vector $W(f)$ and data vector $X(f)$ are defined as

$$W(f) = [w_1(f) \cdots\ w_n(f) \cdots\ w_N(f)]^T$$

And

$$X(f) = [x_1(f) \cdots\ x_n(f) \cdots\ x_N(f)]^T$$

Where $(\cdot)^T$ denotes matrix transpose. A block diagram showing the structure of a general filter-sum beamformer is given in Fig. 3.2.

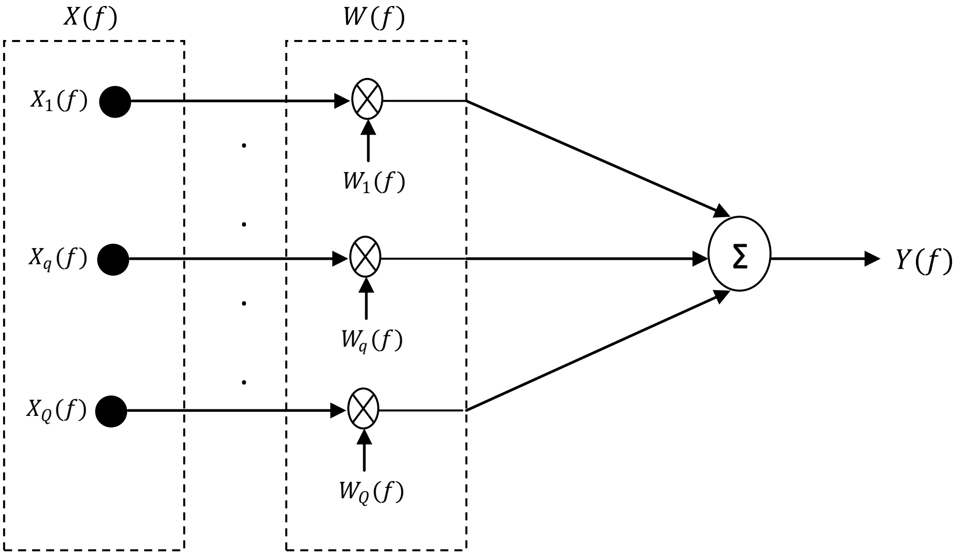

Fig. 3.2. Structure of Filter and sum beamforming.

Both the delay-and-sum and filter-and-sum methods are examples of fixed beamforming algorithms, as the array processing parameters do not change dynamically over time. If the source moves then the delay values will of course change, but these algorithms are still considered fixed parameter algorithms.

Post-Filtering

Post-filtering is a method to improve the performance of a filter-and-sum beamforming algorithm. A Wiener post-filter approach makes use of the information about the desired signal acquired by the spatial filtering, to achieve additional frequency filtering of the signal

[13]. It makes use of cross spectral density functions between channels, which improves the beamformer cancellation of noise.

3.2.2.2 Optimal beamforming

An optimal beamformer pre-computes an optimal set of filter weights based on a model of the array and sources, or alternatively it can be based on calibration information [14]. Examples of optimal beamforming are multi-channel Wiener filter, the eigenvector beamformer, the Linearly Constrained Minimum Variance (LCMV) beamformer, and the Minimum Variance Distortion-less Response (MVDR) beamformer.

3.2.2.3 Adaptive Beamforming

A beamforming which adaptively forms its directive patterns is called an adaptive beamforming. Adaptive beamforming is a powerful technique of enhancing a desired signal while suppressing the noise at the output of the array sensor. In adaptive beamforming, the array-processing parameters are dynamically adjusted according to some optimization criterion, either on a sample-by-sample or on a frame-by-frame basis. Adaptive beamforming alters the direction pattern in accordance with the changes in the acoustic environment, and thus provides a better performance than fixed beamforming and possesses high capability of noise reduction, particularly of prior unknown directional noise as compared to that of fixed beamforming.

Examples of adaptive beamformers are,

(i) *Frost algorithm:* This algorithm is a constrained LMS algorithm in which filter taps (weights) applied to each signal in the array are adaptively adjusted to minimize the output power of the array while maintaining a desired frequency response in the look direction.

(ii) *Generalized Side lobe Canceller (GSC):* The GSC consists of two structures, a fixed beamformer which produces a non-adaptive output and an adaptive structure for side lobe cancellation. The adaptive structure of the GSC is preceded by a blocking matrix which blocks signals coming from the look direction. The weights of the adaptive structure are then adjusted to cancel any signal common to both structures.

(iii) *Soft-constrained Recursive Least Squares (SC-RLS):* SC-RLS beamformer is a practical realization of the adaptive Wiener filter [15]. The SC-RLS structure is sensitive to movements amongst the calibrated desired sources, and an additional source tracking structure is required to track and to compensate for these movements.

(iv) *Sub-band beamforming:* Sub-band beamforming optimizes the array output by adjusting the weights of finite length digital filters so that the combined output contains minimal contribution from noise and interference [17]. This method is highly useful in speech extraction especially when it involves room reverberation suppression, reducing computational complexity and improving the overall performance of the filter.

Adaptive beamforming algorithms are very sensitive to steering errors and might suffer from signal leakage, degradation and significant signal cancellation still arises from the target signal reflections in reverberant environments. As a result, conventional adaptive filtering approaches have not gained wide spread acceptance for speech recognition applications.

3.3 Time domain beamforming with wiener filter

In this work, Filter-and-sum beamforming technique implemented with wiener filter in time domain and Subband beamforming is implemented in frequency domain which is discussed in the next chapter. Wiener filter is used for noise reduction [14, 15, 16]. All the unwanted disturbances or interferences and reverberation are considered to be noise here. So embedding wiener filter into beamforming technique will be one of the finest solutions to the process of speech enhancement. As a linear microphone array is used in this thesis, wiener filter in this case is referred to be a multi channel wiener filter.

For the input vector $x(t)$ at discrete-time instant t, containing mainly frequency components around the center frequency **Ω**, the spatial correlation matrix is given by

$$R_{xx}(t) = E[x(t)x^H(t)] \qquad (3.10)$$

Where, $x^H(t)$ is hermitian transpose of $x(t)$.

Considering that the speech signal, the interference and the ambient noise are uncorrelated, R can be written as

$$R_{xx}(t) = R_{ss}(t) + R_{ii}(t) + R_{nn}(t) \qquad (3.11)$$

Where $R_{ss}(t)$ is the source correlation matrix, $R_{ii}(t)$ is the interference correlation matrix and $R_{nn}(t)$ is the noise correlation matrix defined by the following equations.

$$R_{ss}(t) = E[x_s(t){x_s}^H(t)]$$

$$R_{ii}(t) = E[x_i(t){x_i}^H(t)]$$

$$R_{nn}(t) = E[x_n(t){x_n}^H(t)]$$

Wiener solution for the time domain beamforming

The optimal filter weight vector based on the Wiener solution [17] is given by

$$W_{opt} = [R_{xx}]^{-1} \quad r_{sx} \qquad (3.12)$$

Where the array weight vector, w_{opt} is arranged as

$$W_{opt} = \left[w_{1,} \;\; w_2 \;\; , \ldots . \;\; , \mathrm{w_N}\right] \qquad (3.13)$$

and r_{sx} is the cross-correlation vector defined as

$$r_{sx} = E[x_s(t)s^H(t)] \qquad (3.14)$$

The signal $s(t)$ is the desired source signal at time sample t. The output of the beamformer is given by

$$y(t) = {W_{opt}}^H \, X(t) \qquad (3.15)$$

CHAPTER 4 SUBBAND BEAMFORMING with WOLA FILTER BANK

4.1 Introduction

Sub-band beamforming with filter bank is an alternative solution for general adaptive beamforming to counter the drawbacks of signal leakage, degradation and significant signal cancellation in reverberant environment [17]. Fig. 4.1 illustrates the structure of sub-band beamforming for speech enhancement system using an array of microphones. Sub-band beamforming improves the performance of the filter by optimizing the array output by adjusting the weights of filter.

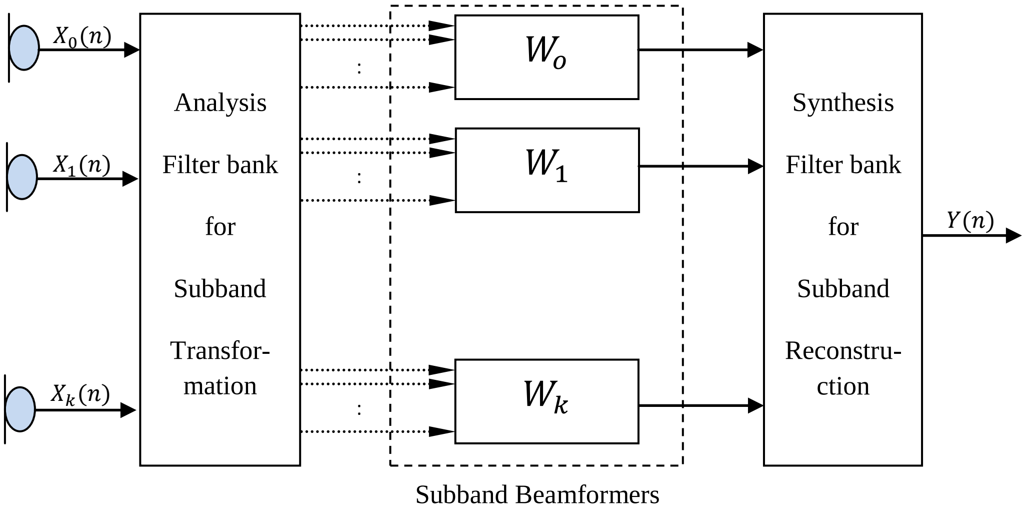
$X_0(n)$
$X_1(n)$
$X_k(n)$
Analysis Filter bank for Subband Transfor-mation
W_o
W_1
W_k
Subband Beamformers
Synthesis Filter bank for Subband Reconstru-ction
$Y(n)$

Fig. 4.1. Structure of Subband beamforming

A multichannel analysis filter bank is included in sub-band beamforming to decompose the received array signals into a set of sub-band signals, and a set of adaptive beamformers each adapting on the multichannel sub-band signals. The outputs of the beamformers are reconstructed by a synthesis filter bank in order to create a time-domain output signal [17].

Filter banks have been introduced in order to improve the time domain adaptive filters. The main improvements of these filter banks are faster convergence and the reduction of computational complexity due to the shorter adaptive filters in the sub-bands operating at a reduced sampling rate [18].Initially, the input signal is divided into sets of narrow band signals called sub-bands such that the bandwidth of these sub-bands should be approximately K times smaller in width than that of the input signal. Here, K represents the total number of sub-bands. This will therefore reduces considerably the complexity of the overall filtering structure. In order to reduce the aliasing effect between the sub-bands, over-sampled sub-band decomposition should be allowed by using a down-sampling factor or decimation factor D such that D is always less than K [19].

4.2 Filter Banks

Filter bank is a method that transforms a signal from the time domain to the time-frequency domain [19]. This time-frequency domain is required in most of the speech processing methods. The time-frequency domain means that a signal is represented in both time as well as a function of frequency which can be achieved by filtering the input time signal by a bank of bandpass filters, where the bandpass filters have very little mutual overlap in frequency. The transformed filter bank signals are denoted as sub-band signals since each of them describe a sub-band of the original signal. Through filter bank processing, larger problems are sub-divided into many smaller problems. Signal processing methods are generally more efficient when they are implemented using filter banks, since the processing load can be implemented in parallel for every subband. The basic structure of filter bank is shown in Fig. 4.2.

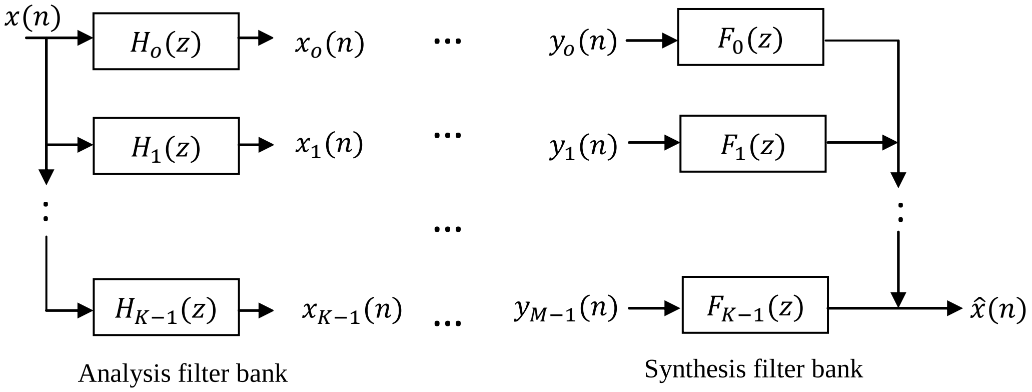

Fig. 4.2. Structure of a filter bank.

Filter bank analysis and synthesis strategies have many advantages in signal processing areas operating as a divide and conquer strategy tackling difficult problems into an equivalent series of much simpler problems. Many signal processing algorithms can be cast into a filtering (frequency-domain) framework. These include dynamic range compression, noise reduction, sub-band coding and directional processing, voice activity detection and echo cancellation. The frequency domain approach is an efficient method of meeting these constraints while delivering low power and flexibility.

The advantage of filter banks is that spatial characteristics of input signal are maintained if the same modulated filter bank is used for all microphone signals. Basically these modulated filter banks are defined by a low pass prototype filter to which all the filters in the bank are modulated by a relation

$$H_k(z) = H_0\left(zW_K^k\right) \qquad (4.1)$$

Where, $H_k(z)$ is the response of the filter used in filter bank, $H_0(z)$ is the prototype low pass filter and $W_k = e^{\frac{-j2\pi}{K}}$. For synthesis part, the filter bank consists of a set of frequency-shifted versions of the low-pass prototype filter.

There are many filter bank configurations and the major filter bank types are

- IIR Filter bank
- FIR Filter bank
- WOLA filter bank
- FFT Modulated Filter bank.

In this thesis, Sub-band beamforming is implemented with the help of WOLA filter bank.

4.2.1 WOLA filter bank

An oversampled DFT filter bank using WOLA (weighted overlap-add) processing provides an extremely efficient and elegant solution [19]. Fig. 4.3 shows a simplified block diagram of an oversampled Analysis of WOLA filter bank and Fig. 4.4 shows its synthesis part.

The input step size (R) is the FFT size (N) divided by the oversampling ratio (OS). The use of over sampling provides two benefits. One is that the gain of the filter bank bands can be adjusted over a wide range without aliasing and a group delay versus power consumption trade-off can be made. In operation, the input FIFO is shifted and R new samples are stored. The input FIFO is then windowed with a prototype low pass filter of length L. The resulting vector is added modulo N (i.e., “folded”) and the FFT of the resulting windowed time segment is computed. The outputs from the analysis filter bank provide both magnitude and phase information since FFT is used.

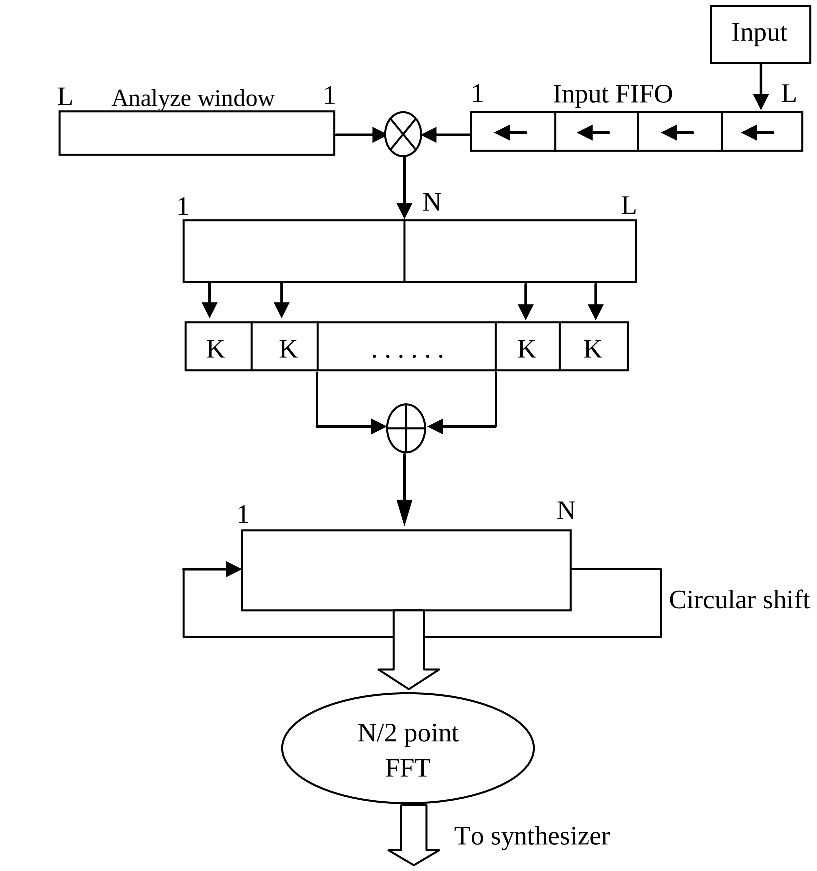

Fig. 4.3. Analysis stage of WOLA filter bank

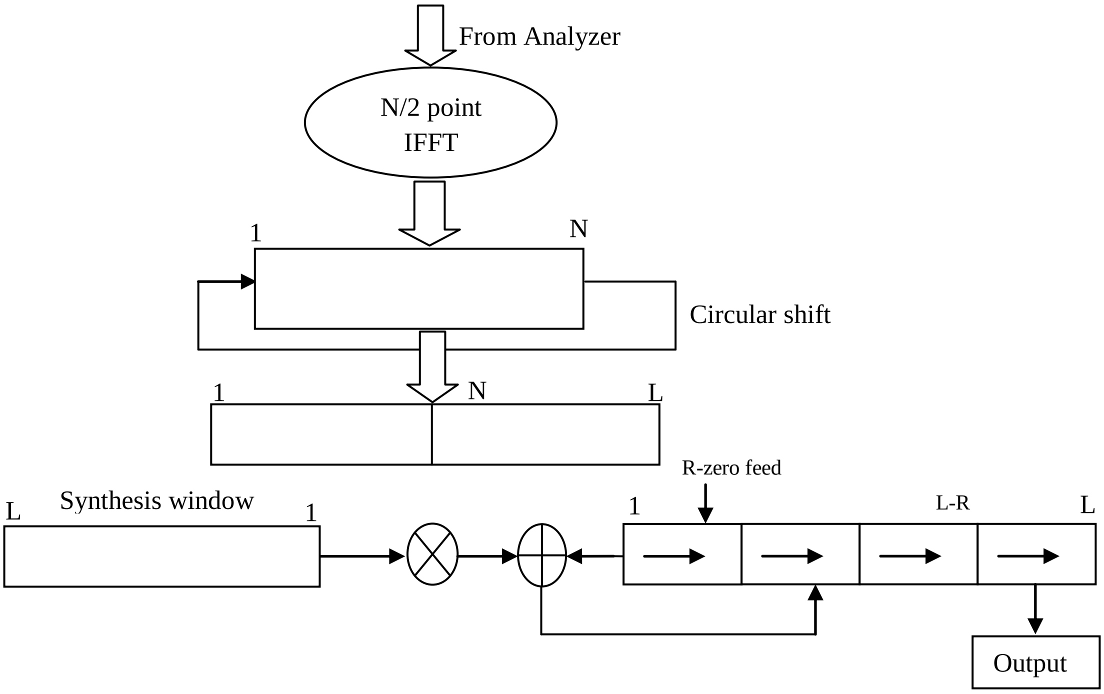

Fig. 4.4. Synthesis stage of WOLA filter bank

To generate a modified time-domain signal, the channel gains are applied to the N/2 FFT outputs (channel signals) and an inverse FFT is computed. The resulting time-domain “slice” is then windowed with a synthesis window and accumulated into the output FIFO. This generates R samples that are shifted out of the output FIFO. Finally, R zeros are shifted into the output FIFO and the entire process repeats for the next block of R input samples. Block-floating-point (BFP) computation units are used to increase the dynamic range and reduce the quantization error in order to improve the SNR of the WOLA filter bank. The BFP strategy decreases the quantization error without increasing the computation complexity. This is achieved by dividing data into non-overlapped groups (passes) and formatting the data at each node in data flow path with common exponent [20].

4.3 Adaptive subband beamforming.

As mentioned in earlier chapter, adaptive beamforming is one of the best solutions in the beamforming categories. For this purpose we need an adaptive algorithm to sub-band beamforming to get even better results. Recursive Least squares (RLS) algorithm has faster rate of convergence than Least Mean Square (LMS) algorithm due to the fact that it whitens the input data by using the inverse correlation matrix of data, assumed to be of zero mean [23]. This algorithm is developed using a relation in matrix algebra called matrix inversion lemma, in which it is used to obtain a recursive equation for computing least square solution $\widehat{w}(n)$ for tap weight vector.

Weighted Recursive Least Squares (WRLS) Algorithm

WRLS algorithm is often used for speech processing and there are many ways to derive it for subband beamforming but the one explained in [17] is most convenient for this thesis.

Assume a filter bank with K sub bands. Consider a subband signal, k from a range of $0\ to\ K-1$, where its corresponding normalized frequency is to be $f=\frac{2\pi k}{K}$. Let the observed microphone signals in sub band number k is denoted by $x_i^{(k)}(n), i=1,2,\dots,I$ at a sample instant n and N be total number of samples in the acquisition phase. The reference microphone is denoted by x_r .

Then the correlation matrices of speech and noise source are determined by the following equations.

When the speech signal is active,

$$X^{(k)}(n) = \left[x_1^{(k)}(n), x_2^{(k)}(n), \dots\ x_I^{(k)}(n)\right]^T \qquad (4.2)$$

$$\hat{r}_s^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} X^{(k)}(n)\ x_r^{(k)^*}(n) \qquad (4.3)$$

$$\hat{R}_{ss}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} X^{(k)}(n) X^{(k)^H}(n) \qquad (4.4)$$

When the noise signal is active,

$$\hat{R}_{nn}^{(k)}(N) = \frac{1}{N}\sum_{n=1}^{N} X^{(k)}(n) X^{(k)^H}(n) \qquad (4.5)$$

Then the above correlation matrices should be memorized in diagonal form using the equation 4.6

$$\left(\hat{R}_{nn}^{(k)}(N) + \hat{R}_{ss}^{(k)}(N)\right) = Q^{(k)^H}\, \Gamma^{(k)}\, Q^{(k)} \qquad (4.6)$$

Where, $Q^{(k)}$ is the set of eigenvectors represented as,

$$Q^{(k)} = \left[q_1^{(k)},\ q_2^{(k)}, \dots\ q_I^{(k)}\right] \qquad (4.7)$$

And $\Gamma^{(k)}$ is the set of eigenvalues denoted by,

$$\Gamma^{(k)} = diag\left(\left[\gamma_1^{(k)}, \gamma_2^{(k)}, \dots\ \gamma_I^{(k)}\right]\right) \qquad (4.8)$$

These eigenvectors and the eigenvalues, $q_1^{(k)}, \gamma_1^{(k)}, i = 1,2,\dots,I$ and the cross correlation vector, $\hat{r}_s^{(k)}(N)$ for each frequency $k = 0,1,\dots,K-1$, are stored in memory for subsequent use.

For operation phase, consider a sub-band weight variable $W_n^{(k)}$ for the k^{th} subband at a time instant n, such that

$$W_n^{(k)} = \left[w_1^{(k)}(n), w_2^{(k)}(n), \dots\ w_I^{(k)}(n)\right]^T \qquad (4.9)$$

Let $P_n^{(k)}$ be a variable used to represent the inverse of the total correlation matrix variable at time instant n, for the k^{th} subband which is to be initialized using equation 4.10

$$P_0^{(k)} = Q^{(k)^H}\,\Gamma^{(k)^{-1}} Q^{(k)} \tag{4.10}$$

Also assume λ and α be the forgetting factor for the WRLS and a smoothing factor for the weight update respectively and they should remain as constants for all frequencies.

With these assumptions, the final equations of WRLS algorithm will be

$$X^{(k)}(n) = \left[x_1^{(k)}(n), x_2^{(k)}(n), \dots\ x_I^{(k)}(n)\right]^T \tag{4.11}$$

$$P^{(k)} = \ \lambda^{-1}P_{n-1}^{(k)} - \frac{\lambda^{-2}P_{n-1}^{(k)}\,X_n^{(k)}\,X_n^{(k)^H}\,P_{n-1}^{(k)}}{1+\lambda^{-1}\,X_n^{(k)^H}\,P_{n-1}^{(k)}\,X_n^{(k)}} \tag{4.12}$$

$$P_n^{(k)} = P^{(k)} - \frac{\gamma_p(1-\lambda)P^{(k)}q_p^{(k)}\,q_p^{(k)^H}\,P^{(k)}}{1+\gamma_p(1-\lambda)\,q_p^{(k)^H}P^{(k)}q_p^{(k)}} \tag{4.13}$$

Where index $p = (n\,mod\,I) + 1,$

The weight vectors are updated according to the equation 4.14

$$W_n^{(k)} = \alpha W_{n-1}^{(k)} + \ (1-\alpha)P_n^{(k)}\hat{r}_s^{(k)} \tag{4.14}$$

And the final output from each subband is then

$$y^{(k)}(n) = W_n^{(k)^H}X_n^{(k)} \tag{4.15}$$

The operation phase consists of continuous decomposition of the microphone signals into discrete frequencies, by the analysis filter bank. The subband weights are updated by making use of both the memorized correlation estimates and the actual microphone observations. The output from each Subband signal is reconstructed with the reconstruction filter bank and the time domain output consists of the estimate of the speech signal. The algorithm is adapting continuously once the correlation estimates are placed into memory. The information gathered in the acquisition phase, will remain as a constant part of the correlation matrix while the contributions from the environmental noise will be subjected to the forgetting factor in the estimates.

Implementation of WRLS algorithm

The algorithm is implemented through the following steps [22]

- The filter output is calculated using the filter tap weights from the previous iteration and the current input vector

$$\bar{y}_{n-1}(n) = \bar{W}^T(n-1)x(n) \tag{4.16}$$

- The intermediate gain vector is calculated using the equation

$$u(n) = \breve{\psi}_\lambda^{-1}(n-1)x(n) \tag{4.17}$$

$$k(n) = \frac{1}{\lambda + x^T(n)u(n)}u(n) \tag{4.18}$$

- The estimation error value is calculated using equation

$$\bar{e}_{n-1}(n) = d(n) - \bar{y}_{n-1}(n) \tag{4.19}$$

- The filter tap weight vector is updated using the 4.19 and the gain vector is calculated using 4.17 and 4.18.

$$w(n) = \bar{W}^T(n-1) + \ k(n)\bar{e}_{n-1}(n) \tag{4.20}$$

- The inverse matrix is calculated using the equation

$$\breve{\psi}_\lambda^{-1}(n) = \lambda^{-1}\breve{\psi}_\lambda^{-1}(n-1) - + k(n) * \left[x^T(n)\breve{\psi}_\lambda^{-1}(n-1)\right] \tag{4.21}$$

CHAPTER 5 SOURCE LOCALIZATION ALGORITHMS

5.1 Introduction

Sound source localization is an important aspect of speech enhancement methods which depend on information of the speaker position. The challenge of identifying the speaker will be more complicated in multi speaker scenario or with a moving speaker. Recent experimental studies show that a steered response power algorithm with phase transform (SRP-PHAT) is a robust algorithm used for sound source localization in reverberant and multiple speaker environments [5].

This Chapter explains the concept and mathematical background behind the SRP-PHAT algorithm. Section 5.2 introduces the classification of existing microphone array based sound source localization techniques. Section 5.3 explains the concept of conventional GCC-PHAT localization algorithm and in section 5.4 SRP-PHAT model is explained.

5.2 Sound Source Localization Strategies

Sound source localization strategies using microphone arrays can be classified into three categories [5].

1 Steered beamformer based locators.
2 High resolution spectral estimation based locators.
3 TDOA based locators.

5.2.1. Steered beamformer based locators:

These locators use a focused beamformer, to steer the microphone array to various locations and searches for a peak in the resultant output power in order to estimate the maximum likelihood sound source location [5]. Delay and sum beamformers, the simplest of these locators time align each of the microphone channel responses and adds them up to get the resultant power. These locators are computationally expensive and the steered response of a conventional beamformer depends heavily on the spectral content of the sound source signal.

5.2.2. High resolution spectral estimation based locators:

These are based on beamforming techniques adapted from the field of high-resolution spectral analysis methods such as autoregressive modeling, minimum variance spectral estimation and Eigen analysis-based techniques [5]. They are used in a variety of array processing applications but they have the following limitations. These algorithms are less robust to source and sensor modeling errors and assume ideal source radiators, uniform sensor channel characteristics, exact knowledge of the sensor positions [5].

5.2.3. TDOA based locators:

The third category is TDOA based locators. These locators use the time delay data for each pair of microphones along with known microphone locations, to generate hyperbolic curves which are intersected in an optimal fashion to find the sound source location. The time delay estimation in these locators is complicated by the presence of background noise and room reverberations. In the noise only case with known noise statistics, the maximum likelihood time-delay estimate is obtained from a SNR-weighted version of the generalized cross correlation (GCC) function [5]. A more robust version of GCC locators known as GCC-PHAT uses phase transform (PHAT) to obtain a peak in the GCC-PHAT function corresponding to the dominant delay in the reverberated signal.

The TDOA based methods are computationally less expensive, but they have limitations as they assume a single source model. Multiple simultaneous sound sources, which is often a case in sound source localization applications, excessive ambient noise or moderate to high reverberation levels in the acoustic field typically results in unreliable sound source locations.

However the above mentioned limitations restrict the usage of these locators in reality. To overcome the limited use of these conventional source localization algorithms in realistic acoustic environment, an algorithm called SRP-PHAT is developed with a combination of Steered beamformer based locators and TDOA based methods which perform better in moderate ambient noise and reverberation levels compared to the previous locators [5].

5.3 GCC- PHAT

The main aim of Generalized Cross Correlation function is to determine the time difference of arrival (TDOA) between two microphones in a pair and has been a popular method [24, 25]. Then from multiple TDOA values, one can estimate the source location. Fig 5.1. is an example of linear microphone array with different arrival times of signal.

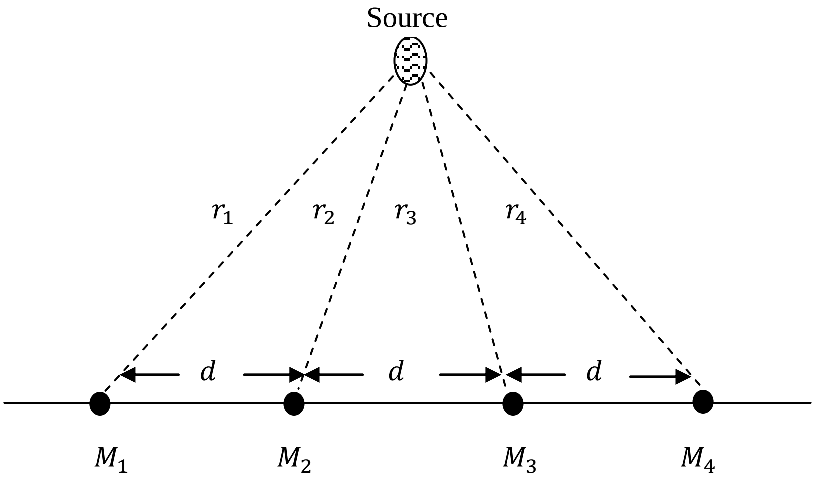

Fig. 5.1. TDOA between two microphones

Let, r_m and r_n be the distance from microphones m , n to the source. Then the travelling time of speech signal from the source to these microphones will be

$$\tau_m = \frac{r_m}{c}\ , \qquad \tau_n = \frac{r_n}{c} \tag{5.1}$$

And their TDOA is defined as

$$\tau_{mn} = \frac{r_m - r_n}{c} \tag{5.2}$$

5.3.1 Derivation of the GCC

Recall Equation 2.13 from Chapter 2 for a microphone signal at microphone m ,

$$x_m(t) = s(t) * h(d_s, t) + n_m(t) \tag{5.3}$$

Consider a signal at another microphone n,

$$x_n(t) = s(t) * h(d_s, t) + n_n(t) \tag{5.4}$$

Note that to be accurate; we would have to include the time delay τ_m into the source signal $s(t)$, i.e. $s(t - \tau_m)$ in equation 5.3 to show the signal received at microphone m is a delayed version of the source signal. Here the concern is all about the relative time-difference of arrival, τ_{mn} between these two microphones m and n.

The cross correlation of these two microphone signals will show a peak at the time-lag where these two shifted signals are aligned, corresponding to the TDOA τ_{mn}. The cross-correlation of $x_m(t)$ and $x_n(t)$ is defined as,

$$c_{mn}(\tau) = \int_{-\infty}^{\infty} x_m(t)\, x_n(t+\tau)\, dt \qquad (5.5)$$

Taking the Fourier Transform of the cross-correlation results in a *cross power spectrum,*

$$C_{mn}(\omega) = \int_{-\infty}^{\infty} c_{mn}(\tau) e^{j\omega\tau}\, d\tau \qquad (5.6)$$

Applying convolution properties of the Fourier Transform for 5.5 when substituting it into 5.6, we have,

$$C_{mn}(\omega) = X_m(\omega) X_n^*(\omega) \qquad (5.7)$$

Where $X_m(\omega)$ is the Fourier Transform of signal $x_m(t)$, and '*' denotes the complex conjugate.
The inverse Fourier Transform of 5.7 gives us the cross-correlation function in terms of the Fourier Transform of the microphone signals:

$$c_{mn}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_m(\omega) X_n^*(\omega) e^{j\omega\tau}\, d\omega \qquad (5.8)$$

The generalized cross-correlation (GCC) of $x_m(t)$ and $x_n(t)$ is the cross-correlation of their two filtered versions. Denoting the Fourier Transforms of these two filters as $W_m(\omega)$ and $W_n(\omega)$, we have the GCC, $R_{mn}(\tau)$ is defined as,

$$R_{mn}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \big(W_m(\omega) X_m(\omega)\big)\big(W_n^*(\omega) X_n^*(\omega)\big) e^{j\omega\tau}\, d\omega \qquad (5.9)$$

We define a combined weighting function, $\Psi_{mn}(\omega)$ as

$$\Psi_{mn}(\omega) = W_m(\omega) W_n^*(\omega) \qquad (5.10)$$

Substituting 5.10 into 5.9, the GCC becomes

$$R_{mn}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_{mn}(\omega) X_m(\omega) X_n^*(\omega) e^{j\omega\tau}\, d\omega \qquad (5.11)$$

The TDOA between two microphones m and n is the time lag τ that maximizes the GCC

$R_{mn}(\tau)$ in the real range limited by the distance between the microphones:

$$\hat{\tau}_{mn} = argmax\ [R_{mn}(\tau)] \qquad (5.12)$$

In reality, $R_{mn}(\tau)$ has many local maxima thus making it harder to detect the global maximum. The choice of the weighting functions, $\Psi_{mn}(\omega)$ would affect the performance of the GCC.

The Phase Transform (PHAT)

It has been shown that the phase transform (PHAT) weighting function is robust in realistic environments [26]. PHAT is defined as follows,

$$\Psi_{mn}(\omega) = \frac{1}{|X_n(\omega)X_m^*(\omega)|} \qquad (5.13)$$

Applying the weighting function PHAT from Equation 5.13 into the expression for GCC in Equation 5.11, the Generalized Cross-Correlation using the Phase Transform (GCC-PHAT) for two microphones m and n is defined,

$$R_{mn}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{|X_n(\omega)X_m^*(\omega)|} X_m(\omega)X_n^*(\omega)e^{j\omega\tau}\ d\omega \qquad (5.14)$$

5.4 SRP-PHAT

It has been shown that talker orientation strongly affects the performance of acoustic localization in smart rooms due to the combinative effects of talker directivity pattern and room reverberation [27]. However, techniques that join the estimated cross-correlations in a collaborative way, such as SRP-PHAT, have shown to be able to perform nearly independently on the talker orientation if the microphones are distributed appropriately in the room.

A filter and sum beamformer output in frequency domain can be defined as:

$$Y(\omega, q) = \sum_{n=1}^{M} G_n(\omega)X_n(\omega)e^{-j\omega\delta_n} \qquad (5.15)$$

Where $G_n(\omega)$ is the Fourier Transform of the adaptive filter, designed for n^{th} microphone input signal, and $X_n(\omega)$ is the Fourier Transform of the $x_n(t)$. Although adaptive filtering compensates the environmental noise and channel effect for some means in real time environment, yet it falls short on efficacy in practical scenarios.

5.4.1 Steered Response Power (SRP)

A conventional steered response power (SRP) is achieved by taking the power of the filter and sum beamformer, steering on the specific area for source localization. It can be expressed in frequency domain as:

$$P(q) = \int_{-\infty}^{\infty} Y(\omega, q) Y^{*}(\omega, q)\, d\omega \tag{5.16}$$

By substituting equation 5.15 in equation 5.16, we get

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{n=1}^{M} G_n(\omega) X_n(\omega) e^{-j\omega\delta_n} \right) \left(\sum_{m=1}^{M} G_m^{*}(\omega) X_m^{*}(\omega) e^{-j\omega\delta_m} \right) d\omega \tag{5.17}$$

Rearranging the expression, we get:

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{n=1}^{M} \sum_{m=1}^{M} \big(G_n(\omega) G_m^{*}(\omega)\big) \big(X_n(\omega) X_m^{*}(\omega)\big)\, e^{j\omega\delta_m - \delta_n} \right) d\omega \tag{5.18}$$

The steering delays δ_m and δ_n will be estimated using TDOA of each microphone pair, which can be written as:

$$\tau_{mn} = \delta_m - \delta_n \tag{5.19}$$

Substituting in equation 5.18 we get

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{n=1}^{M} \sum_{m=1}^{M} \big(G_n(\omega) G_m^{*}(\omega)\big) \big(X_n(\omega) X_m^{*}(\omega)\big)\, e^{j\omega\tau_{mn}} \right) d\omega \tag{5.20}$$

Weighting function can be defined for filter as:

$$\Psi_{nm}(\omega) = G_n(\omega) G_m^{*}(\omega) \tag{5.21}$$

Therefore equation 5.20 becomes

$$P(q) = \sum_{n=1}^{M} \sum_{m=1}^{M} \int_{-\infty}^{\infty} \Psi_{nm}(\omega) X_n(\omega) X_m^{*}(\omega)\, e^{j\omega\tau_{mn}}\, d\omega \tag{5.22}$$

A generalized SRP-PHAT for speaker localization is defined in equation 5.22 can be modified by changing the summation limits to minimize the computations. The modified equation is:

$$P(q) = \sum_{n=1}^{M} \sum_{m=n+1}^{M} \int_{-\infty}^{\infty} \Psi_{nm}(\omega) X_n(\omega) X_m^{*}(\omega)\, e^{j\omega\tau_{mn}}\, d\omega \tag{5.23}$$

The PHAT weighting functions can be defined as

$$\Psi_{nm}(\omega) = \frac{1}{|X_n(\omega)X_m^*(\omega)|} \tag{5.24}$$

Where$\Psi_{nm}(\omega)$ is the desired PHAT filter for the input signals of a microphone array and the relation of channel filter with weighting function can be expressed as,

$$G_n(\omega)G_m^*(\omega) = \frac{1}{|X_n(\omega)X_m^*(\omega)|} \tag{5.25}$$

Substituting equation 5.24 in equation 5.25, we get

$$P(q) = \sum_{n=1}^{M} \sum_{m=n+1}^{M} \int_{-\infty}^{\infty} \frac{1}{|X_n(\omega)X_m^*(\omega)|} X_n(\omega)X_m^*(\omega)\, e^{j\omega\tau_{mn}}\, d\omega \tag{5.26}$$

Where τ_{mn} is the time delay between microphones m and n

5.4.2 Angle of Arrival

Assume two microphones $x_1(t)$ and $x_2(t)$ in a linear array, separated by a distance d in far field zone and with a delay τ between the signals received by them. Let α be the angle where the sound source is located.

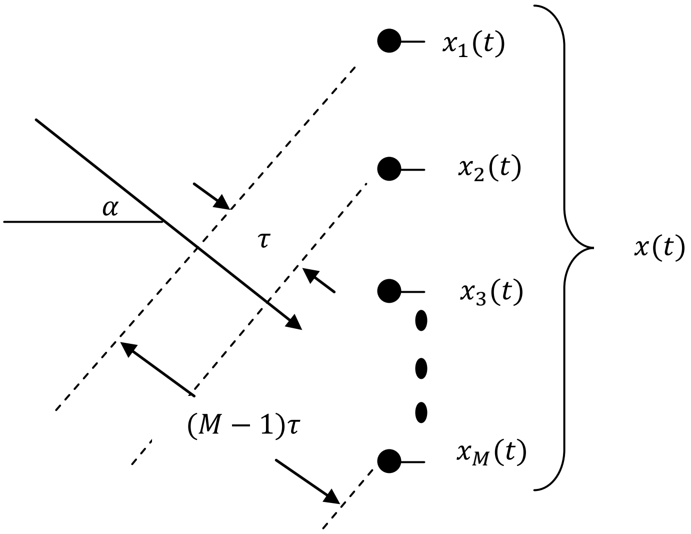

Fig. 5.2. DOA using two microphones in far field zone.

For speaker localization, one has to estimate the DOA of acoustic sound wave. From Fig. 5.2, we can calculate the DOA:

$$\sin(\alpha) = \frac{v * \tau}{d} \tag{5.27}$$

Where v is speed of sound i.e., 340 m/s. TDOA is also dependent of the sampling frequency f_s, as it will be calculated in seconds. So equation 5.27 becomes

$$\sin(\alpha) = \frac{v * \tau}{d * f_s} \tag{5.28}$$

And for estimating DOA, we get

$$\alpha = \sin^{-1}\left(\frac{v * \tau_s}{d * f_s}\right) \tag{5.29}$$

Thus SRP PHAT algorithm calculates DOA by estimating the TDOA to locate the speaker position in the conference room with the help of output parameter α.

Implementing SRP-PHAT algorithm can be summarized in the following steps.

1. Pre-compute theoretical delays from each possible exploration position to each microphone pair.
2. For each analysis frame compute the cross-correlations of each microphone pair.
3. For each position accumulate the contribution of cross-correlations (using delays pre-computed in 1).
4. Select the position with the maximum score.

CHAPTER 6 SPEECH QUALITY ASSESSMENT PARAMETERS

The classical objective measures for distortion assessment in speech signals can be implemented either on the time-domain or frequency domain and at the same time they can also be used for speech quality assessment. There are several objective speech quality measures. Here in this thesis, Signal-to-Noise Ratio (SNR) and Perpetual Evaluation of Speech Quality (PESQ) are the two parameters used to evaluate the results obtained.

6.1 Signal-to-noise ratio (SNR)

The Signal to Noise Ratio (SNR) is one of the most used measures in different conditions of signal processing, both for analog and digital systems. It compares the original and processed speech signals sample by sample. One of the main benefits of SNR is its mathematical simplicity, which makes it easy to be implemented. Over the years, many variations of the SNR have been developed, including the Classical SNR, Segmented SNR, and Segmented Averaged SNR over Frequency and many others [28].

The goal of SNR is to measure the distortion of processed speech signal from that of input speech signal. The general expression of SNR is,

$$SNR = 10\log_{10}\frac{\sum_{i=1}^{N} x^2(i)}{\sum_{i=1}^{N}(x(i)-y(i))^2} \qquad (6.1)$$

Where $x(i)$ and $y(i)$ are the original and processed speech samples indexed by *'i'* and N is the total number of samples.

For simplicity, the above equation can be written as,

$$SNR = 10\log_{10}\left(\frac{\sigma^2{}_{speech}}{\sigma^2{}_{noise}}\right) \qquad (6.2)$$

Where $\sigma^2{}_{speech}$ and $\sigma^2{}_{noise}$ are variance of speech signal and noise respectively. Hence, the improved SNR of the system can be obtained from,

$$SNR_{IMPROVED} = SNR_{OUTPUT} - SNR_{INPUT} \qquad (6.3)$$

$$SNR_{IMPROVED} = 10\log_{10}\left(\frac{\sigma^2{}_{speech_{output}}}{\sigma^2{}_{noise_{output}}}\right) - 10\log_{10}\left(\frac{\sigma^2{}_{speech_{input}}}{\sigma^2{}_{noise\ input}}\right) \qquad (6.4)$$

6.2 Perpetual Evaluation of Speech Quality (PESQ)

Determining the subjective speech quality has always been a laborious process and often expensive. PESQ is an objective measurement tool that predicts the results of subjective listening tests and provides a rapid and repeatable result in a few moments [29].

PESQ was initially proposed as an enhancement for Perpetual Speech Quality Measure (PSQM) which is a speech quality assessment for the telephonic bandwidth. Later it was accepted by International Telecom Union (ITU) and then it is referred as ITU-T P.862 standard [30].

PESQ uses a sensory model to compare the original, unprocessed signal with the processed signal. The resulting quality score is analogous to the subjective "Mean Opinion Score" (MOS). PESQ takes into account coding distortions and variable delay while calculating the MOS. The user interfaces have been designed to provide a simple access to this powerful algorithm, either direct speech signals or from recorded signals.

VAD (voice activity detection) which detects active speech is also a part of PESQ, making it eminent in expressing the voice quality in terms of MOS. The PESQ MOS as defined by the ITU recommendation P.862 ranges from 1.0 (worst) up to 4.5 (best). In practice, however its actual range is from 1 to 5 as shown in Table 6.1.

TABLE 6.1

CLASSIFICATION OF SPEECH QUALITY ACCORDING TO PESQ SCORE.

Quality of speech	Score
EXCELLENT	5
GOOD	4
FAIR	3
POOR	2
BAD	1

Other than speech quality assessment, PESQ can also be used for, providing rapid feedback on areas of signal processing, validation of design implementation, ranking alternative design solutions, providing a higher degree of confidence before submission to subjective testing.

CHAPTER 7 IMPLEMENTATION and EVALUATION of RESULTS

All the implementations are done using MATLAB in offline mode i.e., pre defined signals are used for both speech and noise. The order of implementation of various stages in this thesis is presented in the Fig. 7.1.

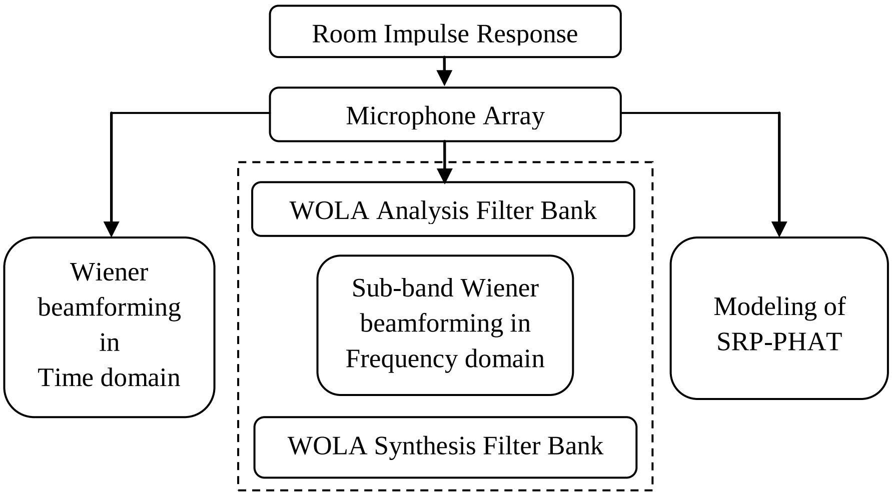
Room Impulse Response
Microphone Array
WOLA Analysis Filter Bank
Wiener beamforming in Time domain
Sub-band Wiener beamforming in Frequency domain
Modeling of SRP-PHAT
WOLA Synthesis Filter Bank

Fig. 7. 1. Various stages in the implementation of thesis

7.1 Simulation of RIR

The first step is to calculate Room Impulse Response (RIR). For this purpose, a virtual reverberant environment is simulated using image method which was discussed in chapter- 2. This RIR is used as a function to generate reverberation in the desired acoustic model. The room dimensions are in meters and chosen in *"length* x *breadth* x *height"* format. Since, Thiran's all pass filter is used, even fractional delays are accounted for while calculating RIR for each reverberant signal, keeping in mind the fact that MATLAB retains only integers discarding the fractional part of a reading. Sampling frequency f_s and reflection coefficient r (varies from 0 to 1) are the parameters used other than room dimensions in implementing reverberant environment. Reflection coefficient is assumed to be a factor which indicates the extent to which a signal gets reflected by the wall of the virtual room for the first time. Where $r = 0$ represents an ideal room without any reflections and $r = 1$ represents a room with maximum reflections.

For presenting the results of RIR, the sampling frequency is chosen at 16000Hz and dimensions of the room are 20m x19m x 21m. Room dimensions may look quite unusual but with these dimensions the results are adequate. By varying the reflection coefficient, we get different responses showing the corresponding energy decay.

The plot of the RIR with reflection coefficient $r = 0$ is shown in Fig. 7. 2. As mentioned earlier, zero reflection coefficient resembles an ideal room without any reflections as all the signals gets absorbed by the wall or simply it is a reverberation-free room.

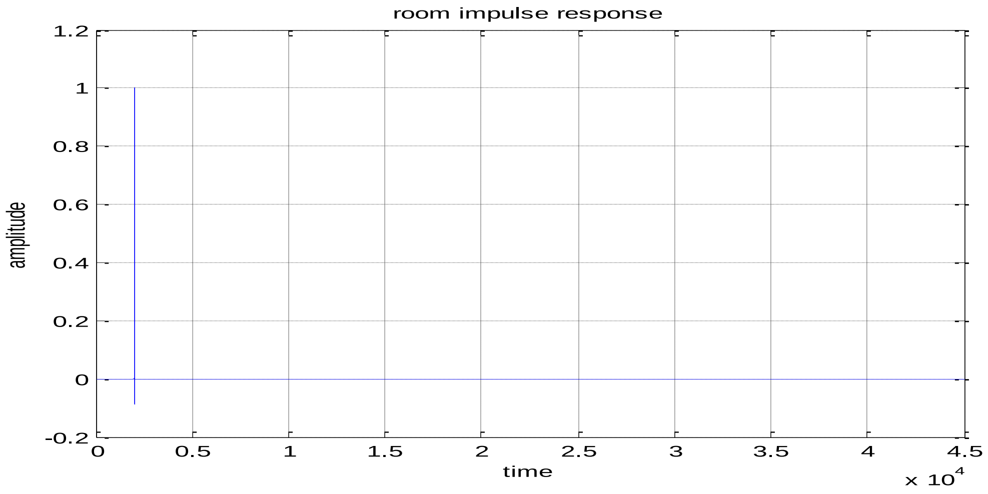

Fig. 7. 2. Energy decay for reflection coefficient r=0, f_s =16000Hz.

The plot of the RIR with reflection coefficient $r = 0.6$ is shown in Fig. 7. 3. In this case, 60% of speech signal is absorbed by the wall and remaining 40% of the signal is reflected back.

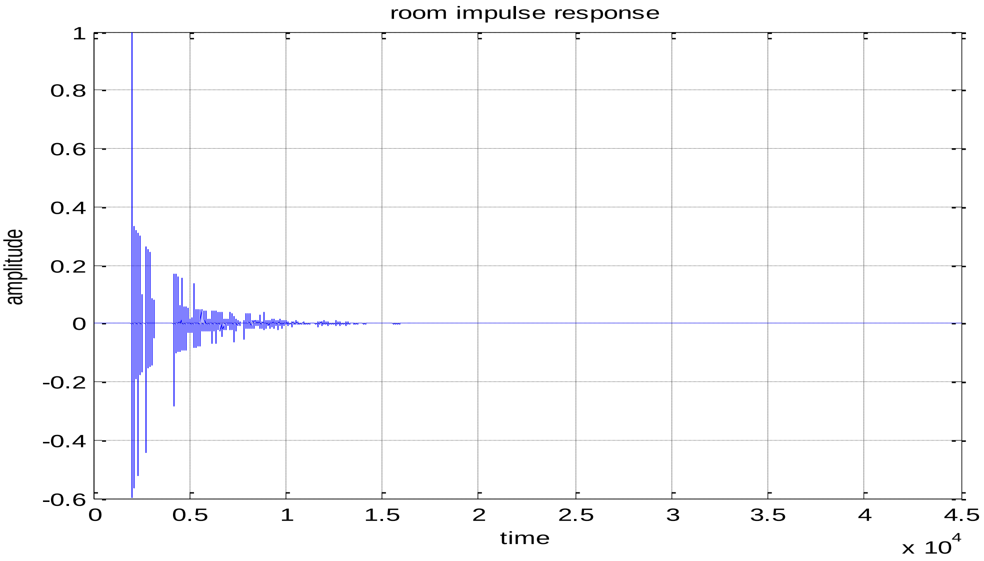

Fig. 7. 3. Energy decay for reflection coefficient r=0.6, f_s =16000Hz.

The plot of the RIR with reflection coefficient $r = 0.3$ is shown in Fig. 7. 4.

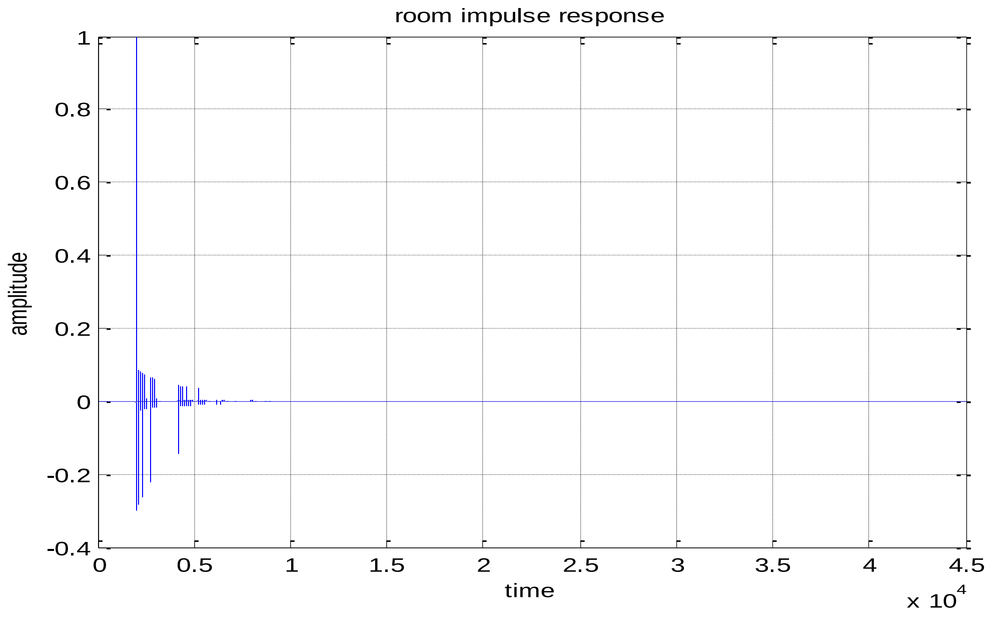

Fig. 7. 4. Energy decay for reflection coefficient r=0.3, f_s =16000Hz.

The plot of the RIR with reflection coefficient $r = 0.4$ is shown in Fig. 7. 5.

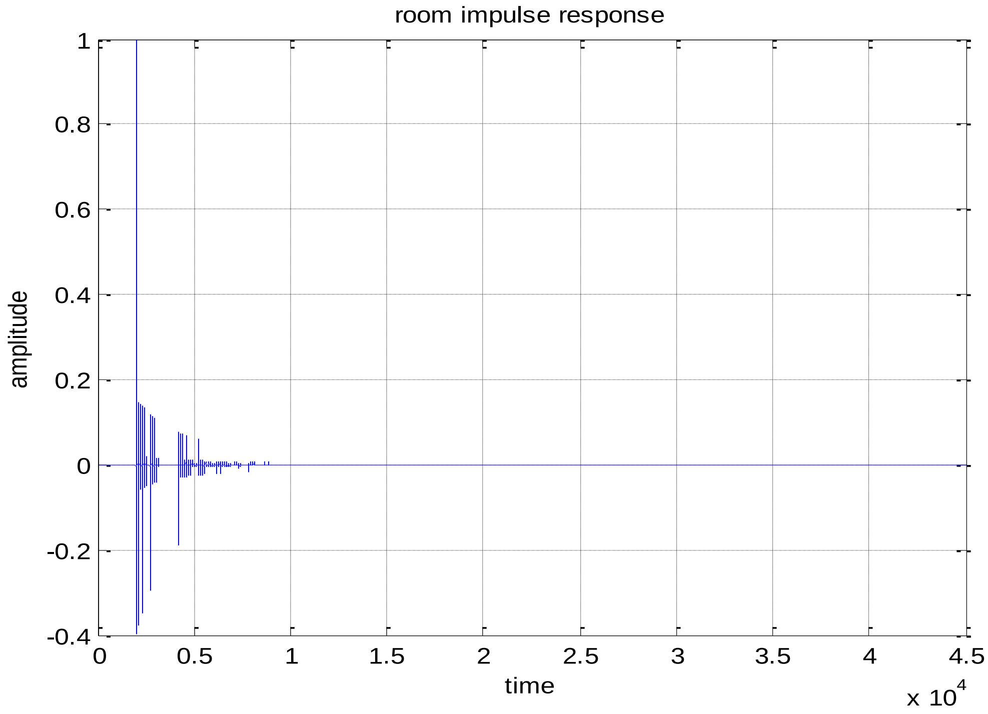

Fig. 7. 5. Energy decay for reflection coefficient r=0.4, f_s =16000Hz.

7.2 Simulation of Wiener beamforming in time domain.

The time domain Wiener beamforming is nothing but a filter-and-sum beamforming in which wiener solution. Initially the room dimensions are assumed to be 6m x 4m x2.8m and fixed locations of sound and noise are assumed. Then the output of RIR is convolved with both speech and noise signal separately. The final input to the liner microphone array will be of the form

$$x(t) = s(t) * h(d_s, t) + n(t) \qquad (7.1)$$

Where $s(t)$ is the source signal, $n(t)$ is the noise signal and $h(d_s, t)$ is the room impulse response. And in particular, for n^{th} microphone the input signal is

$$x_n(t) = s(t) * h(d_s, t) + n_n(t) \qquad (7.2)$$

The quality of processed or output speech signal is evaluated using the parameters explained in chapter 6 i.e., SNR and PESQ. The distance between the microphones is varied uniformly and corresponding SNR values are calculated. PESQ score is obtained by comparing recorded input speech and processed speech signals.

With the noise signal being wind noise and reflection coefficient $r = 0.3$ the improved SNR values are shown in Table 7.1

TABLE 7.1.
SNR IMPROVEMENT AND PESQ SCORE FOR WIND NOISE WITH ROOM DIMENSIONS $6m\,x\,4m\,x\,2.8m,\ r = 0.3$.

Distance between mics	SNR input(dB)	SNR output(dB)	SNR Improvement	PESQ score
0.02	13.4708	31.0235	17.5526	2.563
0.04	13.4708	31.4653	17.9944	2.490
0.06	13.4708	32.1397	18.6689	2.473
0.08	13.4708	30.8022	17.3314	2.571

With room dimensions changed to 8m x 8m x 8m, the corresponding SNR values are tabulated in Table 7.2

TABLE. 7. 2.
SNR IMPROVEMENT AND PESQ SCORE FOR WIND NOISE WITH ROOM DIMENSIONS $8m\,x\,8m\,x\,8m,\ r = 0.3$.

Distance between mics	SNR input	SNR output	SNR improvement	PESQ score
0.02	13.6721	31.8582	18.1861	2.581
0.04	13.6721	32.3435	18.6714	2.573
0.06	13.6721	32.5759	18.9038	2.535
0.08	13.6721	32.4912	18.8190	2.572

In both the cases, the improvement in SNR and PESQ are more or less alike. Therefore in the next step instead of wind noise, an Additive White Noise is used for evaluating the implemented system. Table 7.3 shows the corresponding SNR values.

TABLE. 7. 3
SNR IMPROVEMENT AND PESQ SCORE FOR WHITE NOISE WITH ROOM DIMENSIONS $8m\,x\,8m\,x\,8m,\ r = 0.3$

Distance between mics	SNR input	SNR output	SNR improvement	PESQ score
0.02	2.6898	26.8969	24.2071	2.812
0.04	2.6837	28.1777	25.4940	2.678
0.06	2.6915	23.3889	20.6974	2.976
0.08	2.6798	29.1262	26.4464	2.736

TABLE. 7. 4.
SNR IMPROVEMENT AND PESQ SCORE FOR WHITE NOISE WITH ROOM DIMENSIONS $8m\,x\,8m\,x\,8m,\ r = 0.6$.

Distance between mics	SNR input	SNR output	SNR improvement	PESQ score
0.02	-4.8411	19.0506	23.8918	3.198
0.04	-4.7921	19.9216	24.7137	3.147
0.06	-4.8152	17.3913	22.2066	3.317
0.08	-4.8156	19.7367	24.5523	3.244

From the Table 7.2 and Table 7.3, one can observe that high SNR improvement is achieved by changing the type of noise signal and by varying the reflection coefficient from 0.3 to 0.6, even better SNR improvement is achieved as shown in Table 7.4. Also the average PESQ score is at 3, which indicates that speech has fair quality. A graphic representation of SNR improvement and PESQ score for different noise signals is given in Fig. 7.6 and Fig. 7.7 respectively.

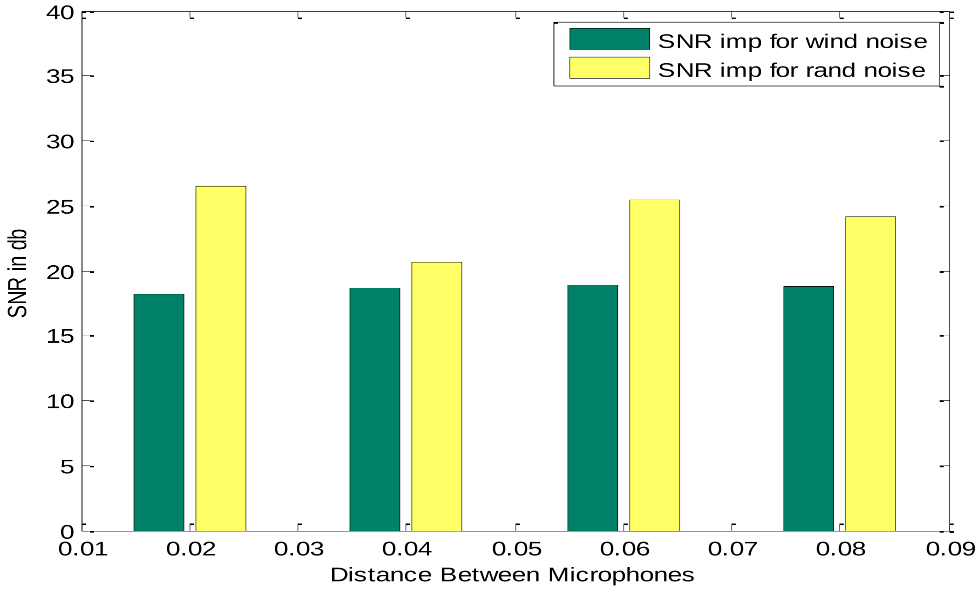

Fig. 7. 6. Representation of SNR improvement through blocks for both wind noise and random noise in time domain.

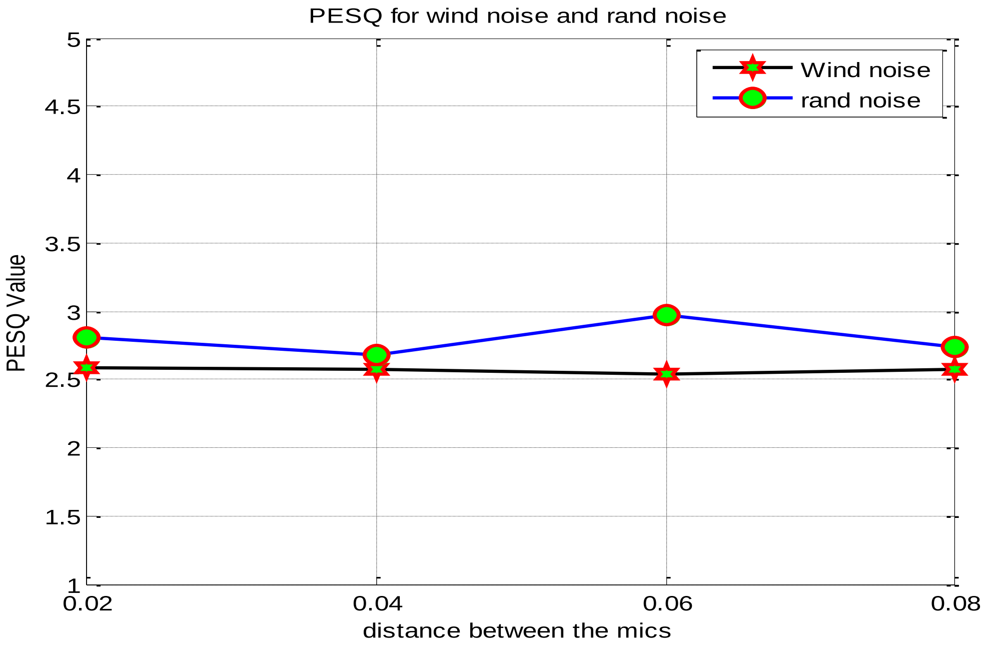

Fig. 7. 7. PESQ scores for wind noise and AWN in time domain.

When speech signal is added to wind noise, the output power spectrum of the beamformer for the best SNR improvement is shown in Fig. 7. 8.

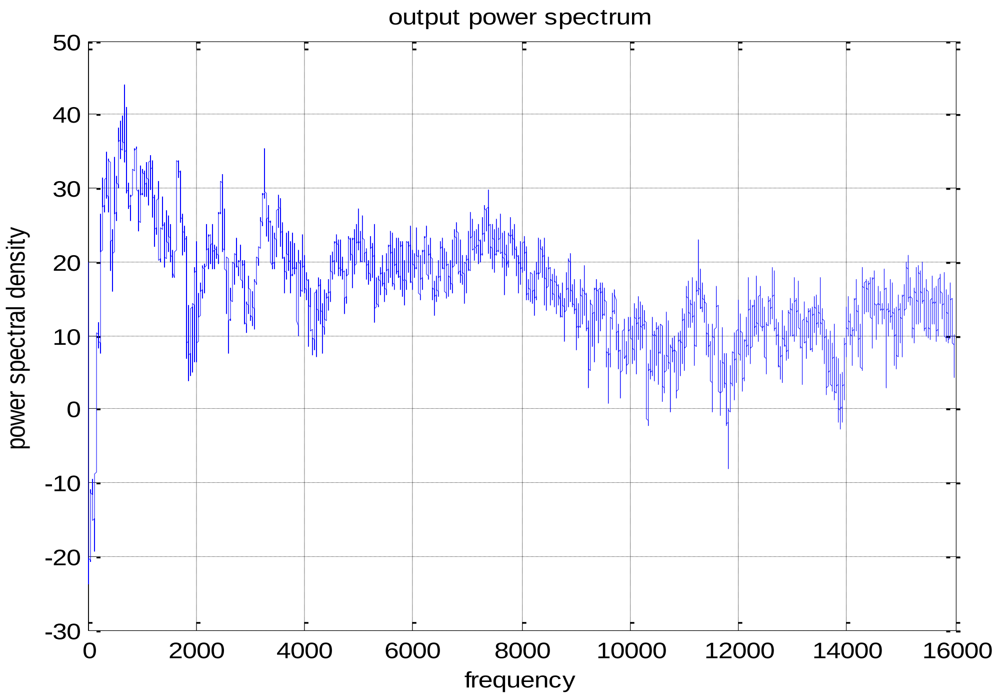

Fig. 7. 8. PSD of the output speech signal obtained from the beamformer with wind noise.

Similarly, when speech signal is added to white noise, the output power spectrum of the beamformer for the best SNR improvement is shown in Fig. 7. 9.

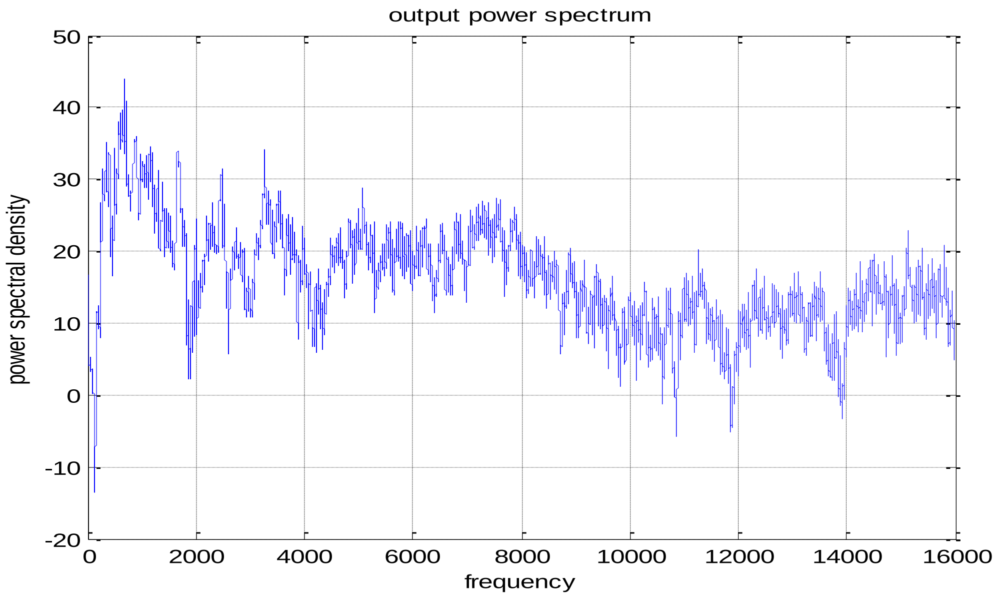

Fig. 7. 9. PSD of the output speech signal obtained from the beamformer AWN

7.3 Simulation of adaptive subband wiener beamforming in frequency domain

The implementation of subband wiener beamforming with WOLA filter bank is similar to that of time domain wiener beamforming. The only difference is that the former involves an extra stage named filter bank. With the aid of filter bank, one can transform a time domain signal into frequency domain and the signal is divided into a small number of frames called sub bands. The advantages of subband beamforming are discussed in chapter 4. As mentioned in chapter 4, WOLA filter bank is implemented in this thesis.

The following filter parameters are considered for testing the performance of filter bank.

Length of the filter L=128, number of sub bands k=64, over sampling ratio OS=2 and sampling frequency $f_s = 16000Hz$.

The power spectral density of the output speech signal is shown in Fig. 7.10.

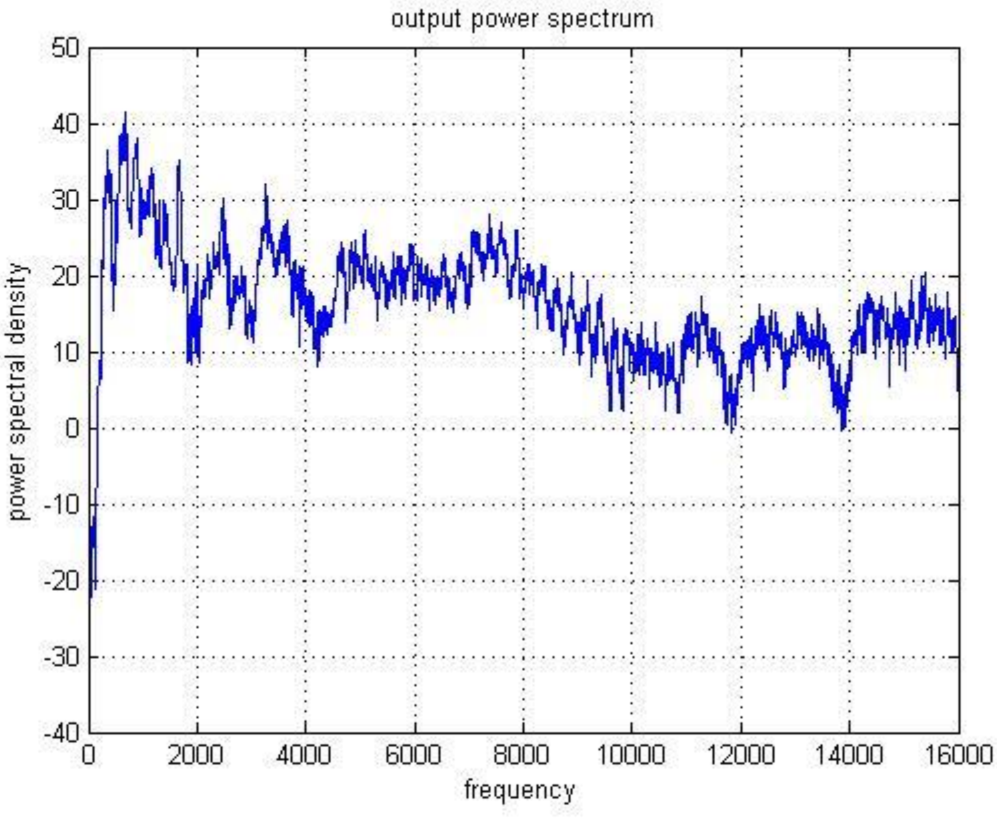

Fig. 7.10. PSD of the output speech signal obtained after processing from the filter bank.

The magnitude response and the impulse response of the WOLA filter bank are presented in Fig. 7.11 and Fig. 7.12.

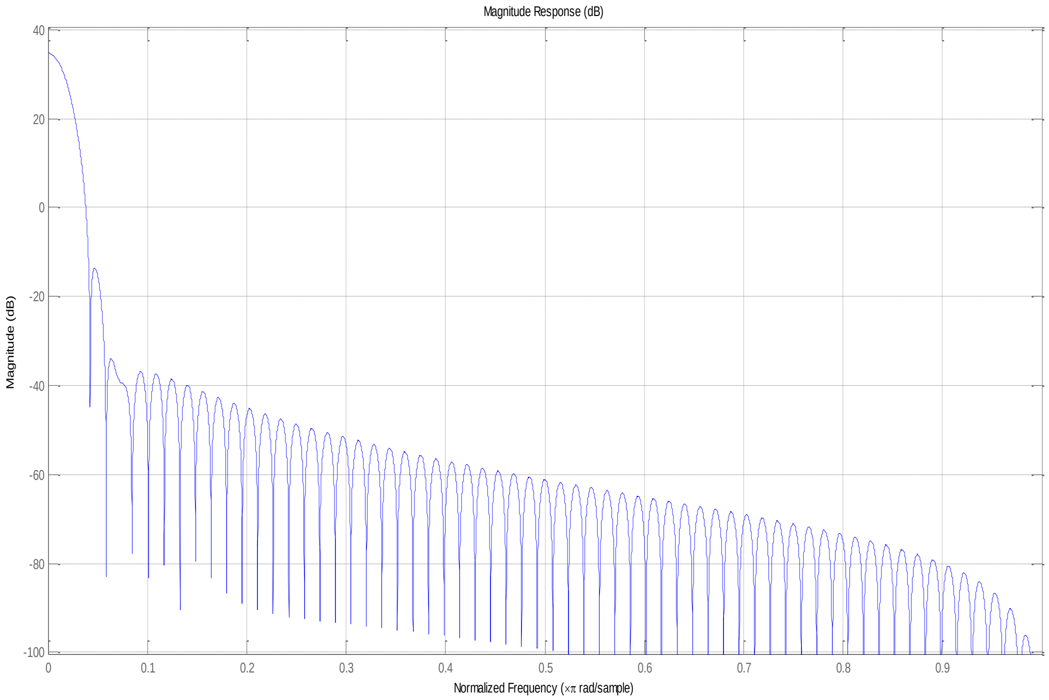

Fig. 7.11. Magnitude response of the WOLA filter bank.

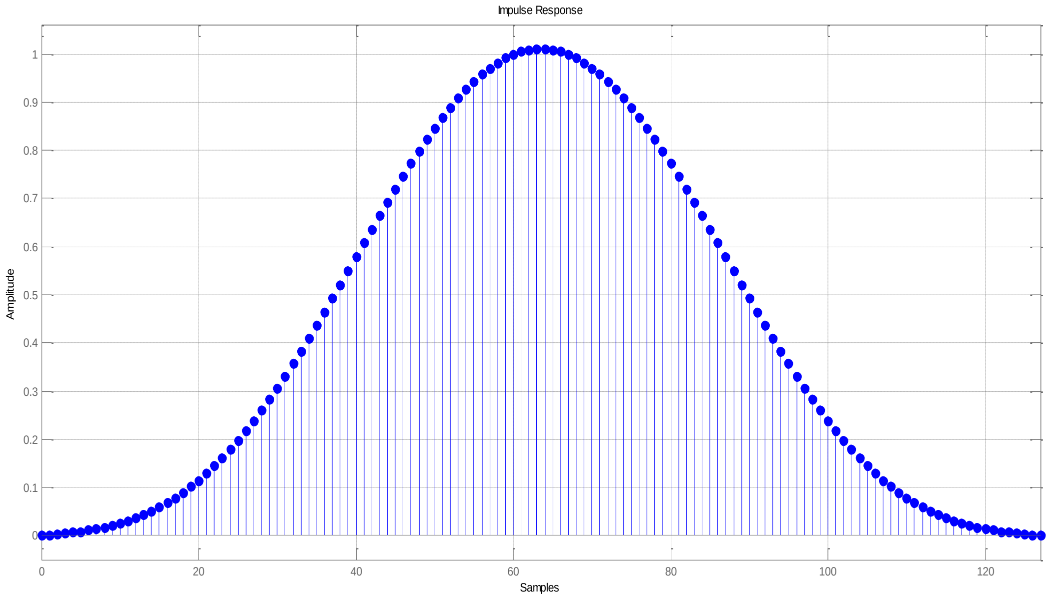

Fig. 7.12. Impulse response of the WOLA filter bank.

With these results, one can say that the filter bank is working fine and resultant speech has perfect reconstruction. So our next step is to embed this filter bank technique into wiener beamforming to take the advantage of sub band beamforming.

RLS algorithm is also used to make the beamformer adaptive. Outcome of RLS algorithm depends on a parameter called forgetting factor denoted by λ. It is a small positive number < 1.

Table 7.5 shows the results of frequency domain subband wiener beamforming with white noise as input along with speech signal with varying forgetting factor λ. Once again room dimensions are chosen as $8m\ x\ 8m\ x\ 8m$.

TABLE. 7. 5.
THE SNR IMPROVEMENT AND PESQ SCORE FOR WHITE NOISE WITH VARYING LAMBDA.

Value of λ	Number of Sub bands (K)	Over sampling ratio (OS)	SNR improvement	PESQ score
0.4	128	64	12.1904	2.984
0.5	128	64	14.2844	1.963
0.6	128	64	14.2677	1.134
0.7	128	64	16.3773	3.432
0.8	128	64	11.9247	1.106

From the above table, it is clear that at $\lambda = 0.7$ the system has high SNR improvement. Now keeping this as constant and varying the other parameters like number of sub bands and over sampling ratio, the following results are obtained.

TABLE. 7.6
THE SNR IMPROVEMENT AND PESQ SCORE FOR WHITE NOISE WITH CONSTANT LAMBDA.

Value of λ	Number of Sub bands (K)	Over sampling ratio (OS)	SNR improvement	PESQ score
0.7	64	32	8.5078	2.016
0.7	256	64	13.4519	3.628
0.7	256	32	14.9927	1.188
0.7	512	64	11.1921	2.786
0.7	256	128	14.3915	2.489

The results obtained when wind noise is chosen as input instead of white noise, are presented in Table 7.7

TABLE. 7. 7
THE SNR IMPROVEMENT AND PESQ SCORE FOR WIND NOISE WITH VARYING LAMBDA.

Value of λ	Number of Sub bands (K)	Over sampling ratio (OS)	SNR improvement	PESQ score
0.5	256	64	6.2159	1.369
0.6	256	64	7.1684	1.729
0.7	256	64	11.8998	1.818
0.8	256	64	12.8465	0.405
0.9	256	64	14.1897	0.946

The graphical representation of PESQ score and SNR improvement for different noise signals in frequency domain is given in Fig. 7.13 and Fig. 7.14 respectively. The waveforms in Fig.7.15 and Fig. 7.16 clearly show the noise reduction. By this it can be concluded that Sub band beamforming has been successfully implemented and desired results are achieved.

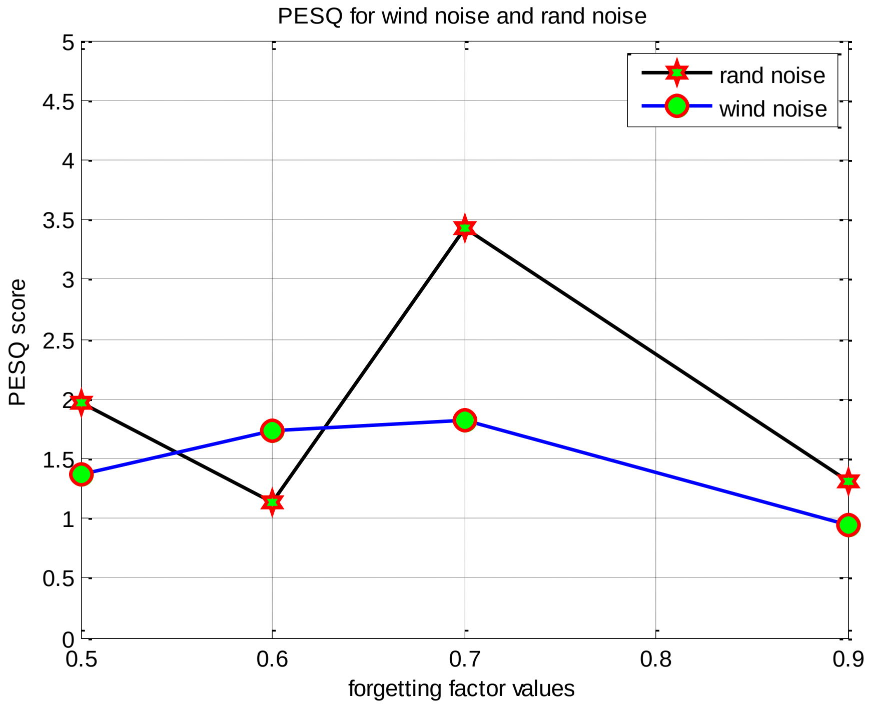

Fig. 7.13. PESQ score for wind noise and AWN in frequency domain.

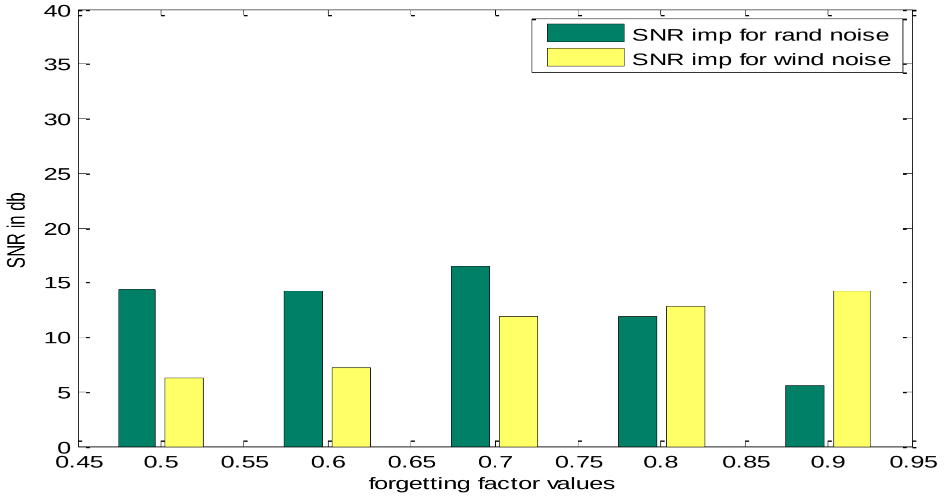

Fig. 7.14. Representation of SNR improvement through blocks for both wind noise and random noise in frequency domain

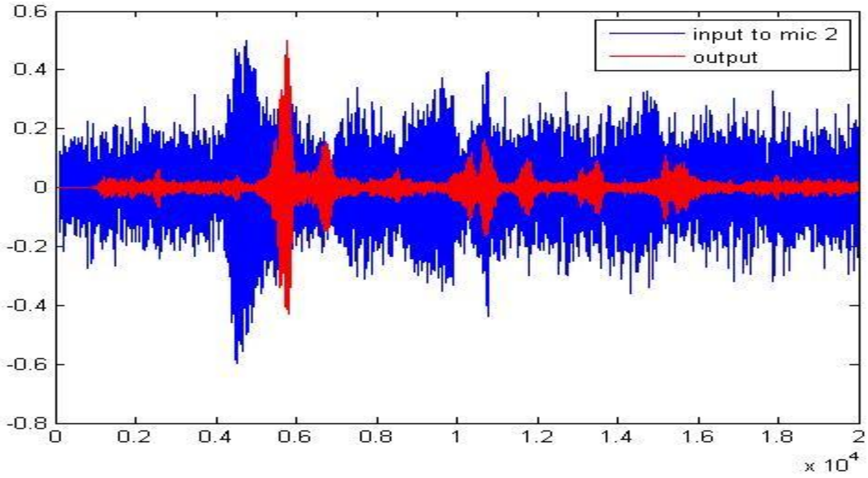

Fig. 7.15. Input speech and the output speech for WHITE noise.

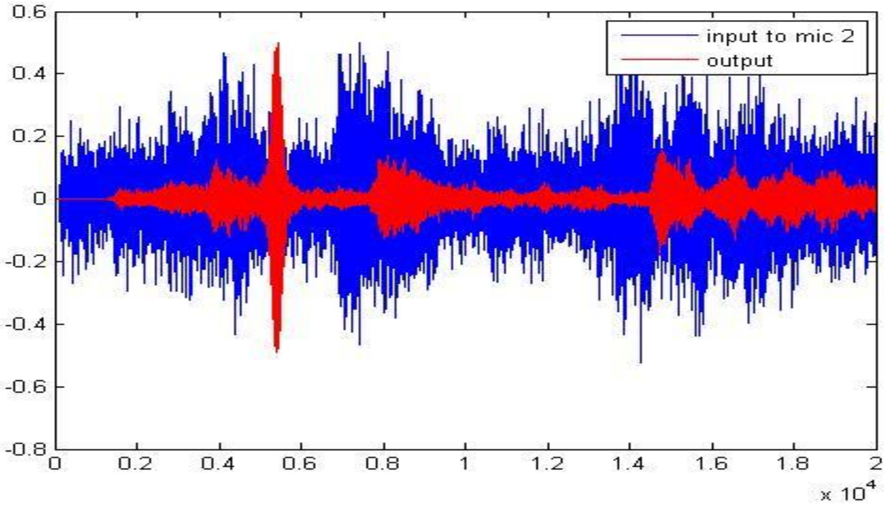

Fig. 7.16. Input speech and the output speech for WIND noise.

7.4 Simulation of SRP-PHAT

With the successful implementation of beamforming in both time domain and frequency domain, the next objective of this thesis moves to the problem of source localization.

The Steered response power (SRP) algorithm is a localization algorithm based on steered beamformers and TDOA methods which use filter and sum beamforming operation. The microphone signals received are time aligned by applying suitable time shifts and their correlation terms are summed together to obtain the steered response power. Recent studies show that SRP-PHAT is a robust algorithm and it shows a distinct performance benefit over the GCC-PHAT based method. SRP-PHAT is therefore chosen as a solution to source localization problem.

Recall steered response power derived in chapter 5 i.e., equation 5.26

$$P(q) = \sum_{n=1}^{M} \sum_{m=n+1}^{M} \int_{-\infty}^{\infty} \frac{1}{|X_n(\omega)X_m^*(\omega)|} X_n(\omega)X_m^*(\omega)\, e^{j\omega\tau_{mn}}\, d\omega \qquad (7.3)$$

Then the TDOA estimate can be calculated from

$$\tau_s = argmaxP(q) \qquad (7.4)$$

Then DOA of the signal is found using the equation below

$$\alpha = \sin^{-1}\left(\frac{v * \tau_s}{d * f_s}\right) \qquad (7.5)$$

With the help of α one can find the position of the speaker.

The evaluation of SRP PHAT is carried out by first calculating the reference position. Then if the input signal is delayed by one sample, the new position must be at variance with the reference position. Considering two mics are separated by a distance d, the test for SRP PHAT is carried out.

A random noise signal is taken as input and then delayed by one sample each time. Corresponding values are tabulated in Table 7.8.

TABLE. 7. 8.
TABLE REPRESENTING THE POWER AND POSITION VALUES IN SRP-PHAT.

Input delayed by n samples	Steered Response Power (SRP)	Position
0 (reference)	17.8792	11
1	17.5010	12
2	17.1255	13
3	16.4689	14
4	16.6669	15
5	15.0400	16

The graphical representations of the positions obtained are represented in Fig.7.17, Fig 7.18 and Fig 7.19.

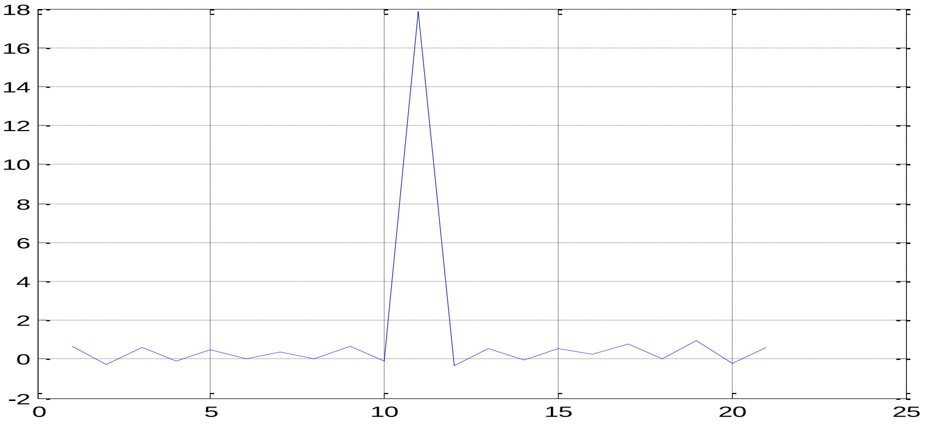

Fig. 7.17. Plot representing the position of the power for the given AWN signal

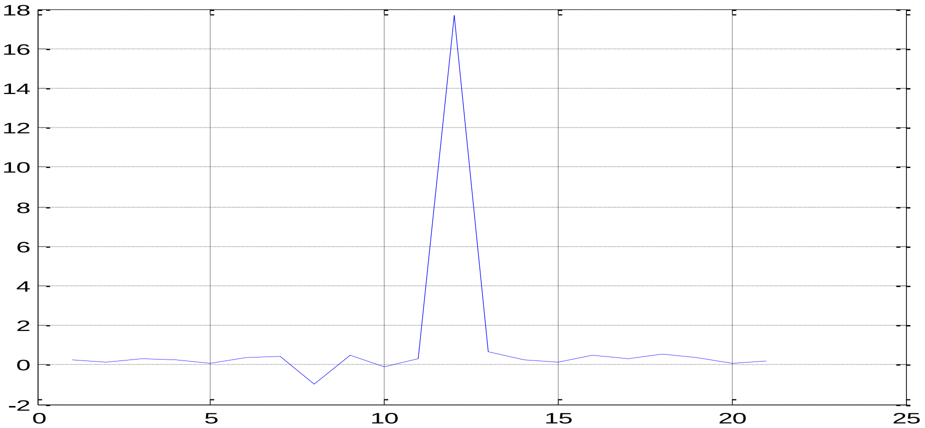

Fig. 7.18. Plot representing the position of the power for the given random noise signal delayed with one sample.

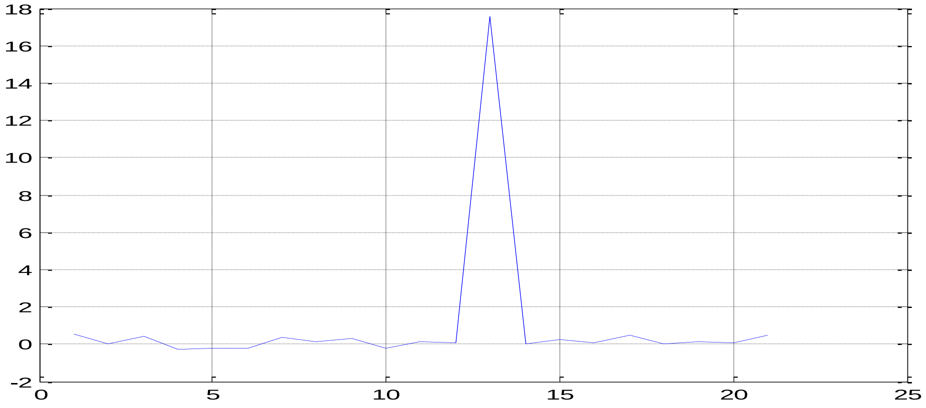

Fig. 7.19. Plot representing the position of the power for the given random noise signal delayed with two samples.

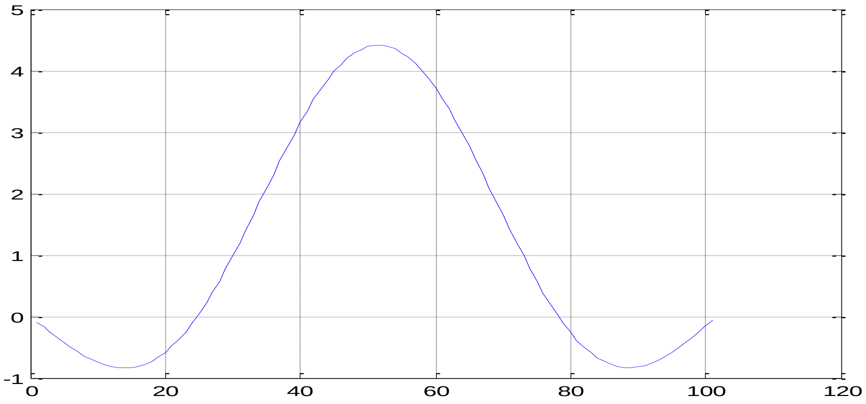

Fig. 7.20. Plot representing the position where the speech is identified for 2 mics.

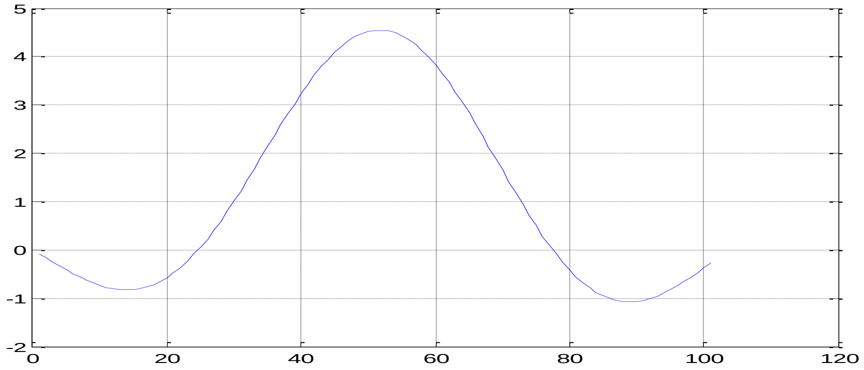

Fig. 7.21. Plot representing the position where the speech is identified for 4 mics.

The peak of the SRP indicates the location of the sound source. A strong reflection of the sound source may at times project a peak, which indicates wrong location of the sound source.

CHAPTER 8 CONCLUSION

The state of the art contributions of the thesis can be concluded as, speech enhancement with microphone linear arrays in a far field talking scenario by means of wiener beamforming in time domain and adaptive subband beamforming using WOLA filter in frequency domain. Also the same microphone geometry is used to locate the speaker position using SRP PHAT. The noise and reverberation are the problems faced in a far field talking scenario in which microphone array is placed at a considerable distance from the sound source. The use of multiple microphones, integrated with beamforming technique, as presented in this thesis, shows substantial improvement in the SNR and quality of speech, has been evaluated by means of PESQ score with satisfactory results. More precisely, two different but related research streams- speech enhancement and recognition with microphone arrays have been followed.

All the work is carried out in MATLAB with pre recorded signals. As mentioned earlier this study is conducted in four different stages. First, the Room impulse response of the virtual room is calculated. The corresponding results show that RIR effect is adequate to create reverberant environment. In the second stage, Wiener beamforming is implemented and tested with different noise signals. The results obtained under various scenarios were tabulated in Table 7.1, 7.2, 7.3 and 7.4 shows that the noise has been suppressed to minimum and processed speech has been enhanced. With these results in time domain, the test of adaptive subband beamforming in frequency domain is carried out with the same signals in third stage. Other parameters that affected the quality of speech in frequency domain are over sampling ratio, number of sub bands chosen by the filer bank and the forgetting factor λ. By varying these parameters, each time the corresponding results were tabulated in Table 7.5, 7.6, 7.7. The Inclusion of RLS algorithm in the subband beamforming resulted in SNR improvement to almost 8dB in comparison to the scenario with no adaptive algorithm.

Over all in time domain, the average SNR improvement is around 18dB and 25dB for wind and additive white noise respectively. In both the cases, on an average PESQ score has been at 2.5.

In frequency domain, the average SNR improvement is around 13dB and 15dB for wind and additive white noise respectively. But the PESQ score has been at 2.o for wind noise and 2.5 for white noise.

Speaker localization involving microphone array processing has been a challenging problem for experiments aimed at speech enhancement. Many studies came up with promising techniques to solve this problem. Out of them SRP PHAT is most convenient in terms of implementation. So far the performance evaluation of the microphone array beamforming system has been carried out with fixed speaker position. But with SRP PHAT one can get the same results with moving speaker as it steers the beam form towards the moving speaker by calculating the DOA of input signal. In the fourth stage, modeling of SRP PHAT is done to locate the speaker position. The test results show that SRP PHAT algorithm implemented in this thesis could locate the speaker position accurately but when it had to be integrated with Wiener beamforming, the results were not satisfactory. There is scope for future research in this area.

Future work

The whole scale of evaluation of the explained beamforming system is done in Matlab offline processing. There is scope for future work to evaluate the thesis in real time i.e., instead of recorded signals, real time signals are used. Then the system can be implemented on DSP processor for real time use. Another important aspect is the multiple speaker environments in which microphone array beam form is supposed to steer towards the current speaker at a very rapid pace. In particular, the application of microphone array processing methods to the multiple source scenarios using parallel beamformers to deliver the good results can be investigated. In which case the need to have a priori knowledge of the source directions would be eliminated with the integration of SRP PHAT.

In this thesis, a linear array of four microphones is used. Research needs to be done with more number of microphones and the results are to be systematically evaluated. Considering other microphone geometries such as circular array provide interesting avenues for further research.

REFERENCES

[1] M. Brandstein, D. Ward (Eds.), “Microphone Arrays Signal Processing Techniques and Applications,” Ed. Berlin, Germany: Springer-Verlag, 2001.

[2] W. Herbordt, W. Kellermann, “Adaptive Beamforming for Audio Signal Acquisition,” Jan. 2003.

[3] S. Y. Low, N. Grbic, and S. Nordholm, “Speech Enhancement using Multiple Soft Constrained Beamformers and Non-Coherent Technique,” in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Apr. 2003.

[4] J.-M. Valin, F. Michaud, J. Rouat and D. Letourneau, “Robust Sound Source Localization using a Microphone Array on a Mobile Robot,” in *Proc. Int. Conf. on Intelligent Robots and Syst.*, Oct. 2003, pp. 1228-1233 vol. 2.

[5] J. H. DiBaise, H. F Silverman and M. S. Brandstein, “Microphone Arrays Signal Processing Techniques and Applications: Robust Localization in Reverberant Rooms,” Ed. Berlin, Germany: Springer, 2001, pp. 157-180.

[6] A. Ramamurthy, H. Unnikrishnan and K. D. Donohue, “Experimental Performance Analysis of Sound Source Detection with SRP PHAT-β,” in *IEEE Southeastcon*, March 2009, pp. 422-427.

[7] S. Bharitkar, C. Kyriakakis, “The Influence of Reverberation on Multichannel Equalization: An Experimental Comparison between Methods,” in Conf. Rec. of the Thirty Seventh Asilomar Conf. on Signals, Syst. and Comput., Los Angeles, CA, USA, Nov. 2003, pp. 546-549, Vol. 1.

[8] J. B. Allen, D. A. Berkley,” Image method for efficiently simulating small-room acoustics.” Acoustics Research Department, Bell Laboratories. [Online].

Available: http://www.umiacs.umd.edu/~ramani/cmsc828d_audio/AllenBerkley79.pdf

[9] V. Välimäki, “Simple Design of Fractional Delay All pass Filter,” in Laboratory of Acoust. and Audio Signal Process., Helsinki University of Technology, Finland.

[10] J. P. Thiran, “Recursive Digital Filters with Maximally Flat Group Delay,” in *IEEE Trans. Circ. Theory,* Nov. 1971, pp. 659–664 vol. 18.

[11] V. Valimaki, T. I. Laakso, "Principles of fractional delay filters," in IEEE Int. Conf. on *Acoust., Speech, and Signal Process., Proc., ICASSP,* 2000, pp. 3870-3873 vol.6.

[12] D. Johnson, D. Dudgeon, “Array Signal Processing - Concepts and Techniques,” Ed. Prentice Hall, 1993.

[13] R. Zelinski, “A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms,” in *IEEE Int. Conf. on Acoust., Speech and Signal Process*., Apr. 1988, pp. 2578 – 2581 vol. 5.

[14] Zohra Yermeche, “Subband Beamforming for Speech Enhancement in Hands-Free Communication,” Ph.D. dissertation, Dept. Elect. Eng., Blekinge Institute of Technology, Ronneby, Sweden, 2004.

[15] N. Grbic, S. Nordholm, “Soft Constrained Subband Beamforming for Hands-Free Speech Enhancement,” in *IEEE Int. Conf. on Acoust., Speech and Signal Process.,* May 2002, pp. I-885–I-888.

[16] J. Chen, J. Benesty, Y. Huang and S. Doclo, “New Insights Into the Noise Reduction Wiener Filter,” in IEEE Trans. on *Audio, Speech, and Language Process.,* July 2006, pp. 1218-1234.

[17] N. Grbic, “Optimal and Adaptive sub-band beamforming, Principles and applications” Ph.D. dissertation, Dept. Elect. Eng., Blekinge Institute of Technology, Ronneby, Sweden, 2001.

[18] Jan Mark de Haan, “Filter Bank Design for Subband Adaptive Filtering Methods and Applications,” Ph.D. dissertation, Dept. Elect. Eng., Blekinge Institute of Technology, Ronneby, Sweden, May. 2001.

[19] P. P. Vaidyanathan, “Multirate Systems and Filter Banks,” Ed. Prentice-Hall, 1993.

[20] R. E. Crochiere, L. R. Rabiner, “Multirate Digital Signal Processing,” Ed. Prentice-Hall, 1983.

[21] R. Brennan, T. Schneider, “An Ultra-Low-Power WOLA Filter bank Implementation in Deep Submicron Technology,” in DSP factory Ltd, Waterloo, Canada.

[22] A. Munjal, V. Aggarwal and G. Singh, “RLS Algorithm for Acoustic Echo Cancellation,” in *Nat. Conf. Challenges and Inform. Technology COIT,* Mar 2008.

[23] S. Haykin, “Adaptive Filter bank Theory,” 2nd Ed. New Jersey: Prentice-Hall Inc.

[24] Hoang Tran Huy Do,” Real –Time SRP-PHAT Source Location Implementations on a Large-Aperture Microphone Array,” M. S. Thesis, Dept. Elect. Eng., Brown Univ., Rhode Island, USA, 2009.

[25] H. F. Silverman, Y. Yu, J. M. Sachar and W. R. Patterson, “Performance of Real-Time Source-Location Estimators for a Large-Aperture Microphone Array,” in *IEEE Trans. Speech, Audio Process.*, July 2005, pp. 593-606.

[26] K. D. Donohue, J. Hannemann and H. G. Dietz, “Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments,” in Signal Process., Jan. 2007, pp. 1677-1691, Vol. 87.

[27] Y. M. Malik, ”Speaker Localization, tracking and remote speech pickup in a conference room,” M. S. Thesis, Dept. Elect. Eng., Blekinge Institute of Technology, Karlskrona, Sweden, 2009.

[28] J. R. Deller, J. G. Proakis, J. G. Hansen and H. L. John, “Discrete-Time Processing of Speech Signal,” Ed. USA: Macmillan Publishing Company, 1993.

[29] ITU-T P.862 "Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", ITU-T publications [Online].

Available: http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm

[30] A.W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, “Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs,” in *IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, 2001, pp. 749-752, vol. 2.