Microphone Array Wiener Beamformer and Speaker Localization

With emphasis on

WOLA Filter Bank

Hemanth Yerramsetty

Master's Thesis

This thesis is presented as part of Degree of Master of Science in Electrical Engineering

Blekinge Institute of Technology

January 2012

Supervisor: Dr. Nedelko Grbic
Examiner: Dr. Benny Sallberg
Department of Signal Processing
School of Engineering (ING)
Blekinge Institute of Technology

Table of Contents

Acknowledgements

I would like to thanks my supervisor, Dr. Nedelko Grbic, for his encouragement, patient guidance and support he provided me throughout this thesis work.

I would also like to thanks to my friends, who supported me a lot during the software part for this thesis. Otherwise it would be difficult for me to complete that part in short time period.

I must express my gratitude to my parents, for their continuous support and encouragement.

Abstract

This thesis describes the design and implementation of a speech enhancement system that uses 4-channel microphone array beam forming and speech enhancement algorithms applied to a speech signal in a multiple source environment. To locate the accurate Direction of Arrival (DOA) from the source, it is necessary to design a suitable microphone array system with more efficient localization algorithm. The goal of the system is to improve the quality of the primary speech signal.

A filter bank is a signal processing tool that can facilitate manipulation of signals in the frequency domain. The WOLA (Weighted Overlap and Add) filter is an efficient method used to implement a uniformly distributed multi-channel filter bank. The WOLA is generally used in applications that demand high quality filters in term of stop band rejection and filter shape.

Beamformers work by means of steering an array of microphones towards a desired look direction through utilizing signal information rather than physically moving the array. In this research, Wiener beam former is examined the input signals are first split into frequency bands so that Wiener beam forming techniques can be used.

There are many algorithms developed for estimating the number of sources and locating the DOA, such as Bayesian algorithm, kalman filtering, Generalized Cross Correlation (GCC) and Steered Response Power (SRP) algorithm. But SRP algorithm with its steered beam forming technique for speaker localization is more robust using microphone array. The Phase Alignment Transform (PHAT) has gained a lot of attention in the recent research for its quite robust response in low noise, but reverberant environment. So combining SRP-PHAT will become the robust localizer in reverberant environment.

Experiments were done on recorded data of human talkers. The algorithm gives accurate DOA from the dominant speaker. In addition to these, listener opinion testing is performed.

1. Introduction

Thesis Statement

A microphone array is the promising solution for realizing hands-free speech recognition in real environments. Accurate talker localization is very important for speech recognition using the microphone array. However localization of a moving talker is difficult in noisy reverberant environments. The talker localization errors degrade the performance of speech recognition. A speech recognition algorithm is implemented to solve the problem, which considers multiple talker direction hypotheses.

This research addresses the problem of primary source enhancement in a multiple source environment. To improve the quality and recognition of the speech signal of interest, a microphone array along with beam forming and speech enhancement algorithms can be used to separate the primary speech signal from the interfering speech signals. Thus, it is the goal of this research to enhance the quality of the primary speech signal of interest through the development and implementation of beam forming and enhancement algorithm.

For accurate and robust speaker localization there is a need to pay attention on two major concerns. Firstly how to design and implement a suitable microphone array and secondly which speech processing algorithm will be used for robust and accurate speaker localization in a conference room.

Thesis Overview

The thesis is divided into 10 chapters, after defining the thesis statement in introduction. Background of microphone array fundamentals, Beamformer, Speaker localization has been discussed in chapter 2. Chapter 3 discusses about Acoustic signal modeling, Room impulse response with Fractional Delay. Filter bank (WOLA) mathematical equations and diagram will elaborate in detail in Chapter 4. Beamforming techniques are explained in chapter 5. And in

chapter 6 we can discuss about SRP-PHAT. Results are discussed in chapter 7. Conclusion and Future work are discussed in chapter 8 & chapter 9. Finally references are showed in chapter 10.

2. Background

Person-machine communication has been one of the most important research milestones of the speech community for decades. Particularly, there is a clear belief that spoken dialogue would be the most natural and powerful user interface to computers. With recent improvements in computer technology, speech and language processing, and also in other fields such as image processing, new goals for computer communication and assistance are starting to appear feasible. Currently, it becomes evident that computers have to adapt to human requirements, being involved in human communication activities, and requiring the minimal possible awareness from the users. Consequently, there is a need of perceptual user interfaces which are multimodal and robust, and which use unobtrusive sensors.

An interesting example of these new challenging multimodal research efforts can be found in the development of smart-rooms. A smart-room is a closed space provided of multiple microphones and cameras, which is designed to assist and complement human activities. For instance, some of the innovative services offered are lecture summarization, identification of people attending a conference, or composition of a draft about what was said in a meeting.

Such highly sophisticated multimodal services are based on the information provided by many basic technological components of the various modalities. In the case of the audio processing, some of the technologies that are involved in these services are Voice Activity Detection, Automatic Speech Recognition, Speaker Identification and Verification or Acoustic Event Classification.

While most of these technologies perform reasonably well in controlled scenarios, in the context of the development of hands-free speech applications where close-talking microphones are not allowed, they present common problems. In this case, the situation is similar to the one described by the cocktail party phenomenon. Audio signals recorded with microphones that can be located several meters away from the source of interest are severely degraded by noise, reverberation, and also position, orientation and dynamics of multiple concurrent speakers. As a

consequence of these sources of degradation, audio technologies show a dramatic lost of performance.

One field of growing interest to reduce problems introduced by distant microphone recordings consists in taking advantage of the multi-microphone availability. More concretely, microphone array processing has been broadly investigated as a pre-processing stage in order to enhance the recorded signal that might be used for any speech application. The basic idea of beamforming is to generate a directive beam that targets the desired direction based on the combination of the signals arriving to an array of microphones.

Unfortunately, most of the speech enhancement techniques based on multi-microphone processing relay on one fundamental cue that is mostly unknown: the source position. The need for reliable target position estimation in the beamforming applications is one of the reasons for the increasing interest in the acoustic source localization and tracking topic. Furthermore, accurate knowledge of the position of the events or the speakers present in a room is also useful for other multimodal services like analyzing group dynamics or behaviors, deciding which the active speaker among all the presents is, or providing information for an automatic steering camera system. Furthermore, knowledge about speaker head orientation can be useful information in such applications.

Hence, in real smart-room environments, a multi-microphone approach to speech processing would permit to enhance multiple captured speech signals on the basis of an estimated speaker position, and eventually these enhanced signals might be used for improving the performance of an automatic speech recognition application, or any other speech based application [1].


BLEKINGE TEKNISKA HÖGSKOLA
BTH

3. A Model for Room Acoustics

3.1 Introduction

Many people who are working in the field of acoustic signal processing reach a point where they want to simulate room acoustics. This report gives a short overview of image methods that can be used for simulating room acoustics. The image method [2], which was proposed by Allen and Berkley in 1979, is probably one of the methods most commonly used in the acoustic signal processing community. A RIR-function, which can be used in MATLAB, has been created to generate multichannel Room Impulse Responses (RIR) using the image method. This function enables the user to control the reflection order, room dimension and microphone directivity.

The image model can be used to simulate the reverberation in a room for a given source and microphone location, and is discussed in Section 3.1. Using the image method Allen and Berkley [2] developed an efficient method to compute a Finite Impulse Response (FIR) that models the acoustic channel between a source and a receiver in rectangular rooms. The image model and image method some additional refinements will discussed in Section 3.2 and in section 3.3. In section 3.4, Fractional Delay has been discussed.

3.2 Image Model

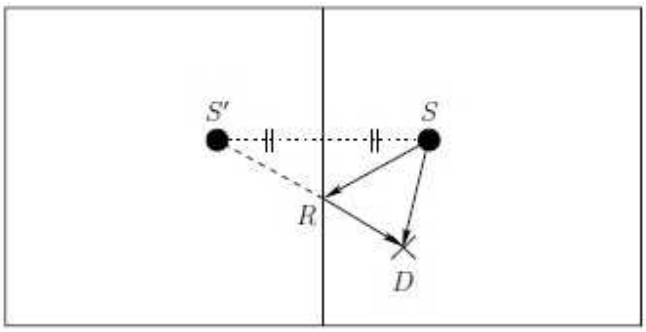


Figure.1: Path involving one reflection obtained using one image.

Figure.1 shows a sound source *S* located near a rigid reflecting wall. At destination *D* two signals arrive, one from the direct path and a second one from the reflection. The path length of the direct path can be directly calculated from the known locations of the source and the destination. Also shown is an image of the source, *S′*, located behind the wall at a distance equal to the distance of the source from the wall. Because of symmetry, the triangle *SRS′* is isosceles and therefore the path length *SR* + *RD* is the same as *S′D*. Hence, to compute the path length of the reflected path, an image of the source is constructed and computes the distance between destination and image. Also, the fact that the computing distance using one image means that there was one reflection in the path.

Figure.2 shows a path involving two reflections. The length of this path can be obtained from the length of *S″D*. In Figure 3 the length of a path involving three reflections is obtained from the length of *S‴D*. These figures can also be extended to three dimensions to take into account reflections from the ceiling and the floor.

In general the path lengths (and thus the delays) of reflections can be obtained by computing the distance between the source images and the destination. The strength of the reflection can be obtained from the path length and the number of reflections involved in the path. The number of reflections involved in the path is equal to the level of images that was used to compute the path [2].

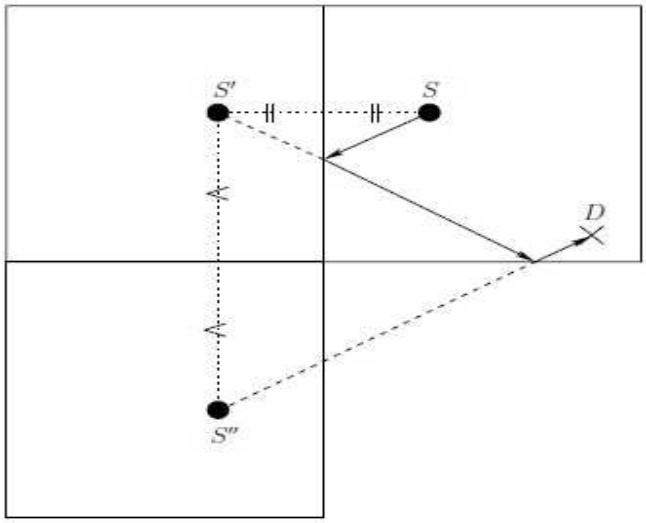


Figure.2: Path involving two reflections obtained using two images.

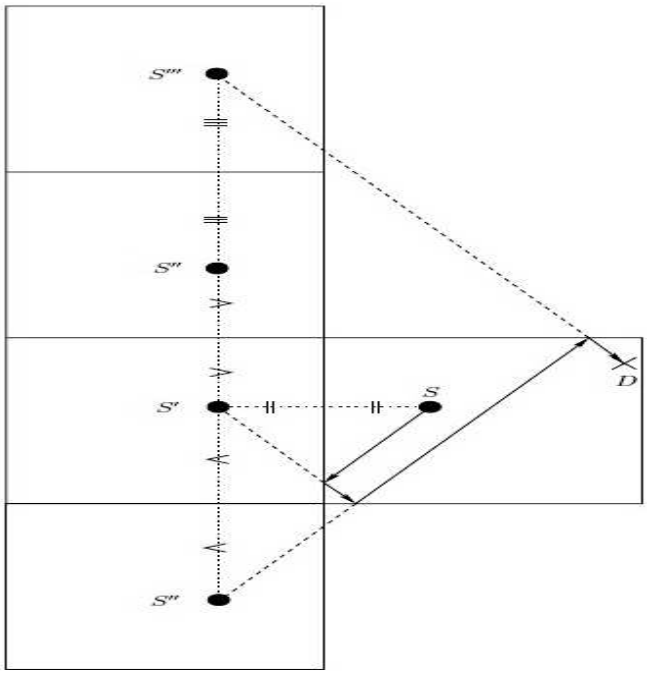


Figure.3: Path involving three reflections obtained using three images.

3.3 Image Method

Consider a rectangular room with length, width and height. x_s is the x-coordinate of the sound source and x_r is the length of the room in the x-dimension and let the microphone be at a location represented by the vector x_m, both vectors are with respect to the origin, which is located at one of the corners of the room. Let's just have a look at the model in just one dimension. This is depicted in the figure below.

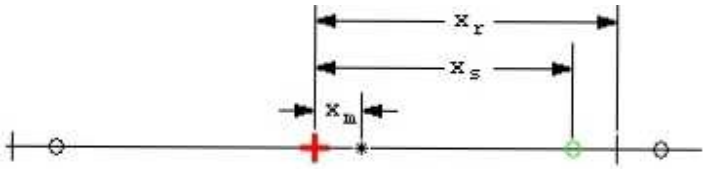


Figure.4: Model of Image method in one dimension

The red part in the above Figure.4 is the origin. The x-coordinate of the virtual sources can be expressed using the sequence below.

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

The location of the i^{th} virtual source is determined by plugging in an integer for i. if i is negative then the virtual source will be located on the negative x-axis. If $i=0$ then the virtual source is actually the real source. To find the distance between the i^{th} virtual sound source and microphone by subtracting the microphone's x-coordinate x_m, fromx_i. This is shown below.

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

In the similar manner the relative positions of the virtual sources along the y and z axes are calculated

To find the distance to each virtual sourcex_i, y_j and z_k has been used in the Pythagoras theorem and is calculated as follows

$$d_{ijk} = \sqrt{x_i^2 + y_j^2 + z_k^2} \qquad (3)$$

This equation represents the distance in the form of a three dimensional matrix.

To finding the unit impulse response function of each virtual source

Let's assume the equation

$$u_{ijk}(t) = t - \frac{d_{ijk}}{c} \qquad (4)$$

From the above equation *t* is the time d_{ijk} is the distance given by the equation 3, and *c* is the speed of the sound. In the above equation $\frac{d_{ijk}}{c}$ is the effective time delay of each echo. From the above equations, impulse response function has a magnitude one when $u_{ijk}(u)$ is equal to zero. Detail about this, there is two things that will affect the magnitude of the echoes. The first thing is the distance it travels from the source to the microphone. This is represented by the following equation.

$$b_{ijk} \propto \frac{1}{d_{ijk}} \qquad (5)$$

The second thing is the number of reflections the sound wave makes while it is transmitted. If all the wall reflection coefficients are the same, Let assume the wall reflection coefficient to be r_w and raise this reflection coefficient to the exponent *n* where $n=|i|+|j|+|k|$. *n* represents the total number of reflections the sound has made. This is shown in equation (6) for the virtual source with the indices *i*, *j*, and *k*.

$$r_{ijk} = r_w^{|i|+|j|+|k|} \qquad (6)$$

Previously the reflection coefficient is same for all the walls but, if each wall has a different reflection coefficient then the situation becomes more complex. If $r_{x=0\ is}$ the reflection coefficient for the wall perpendicular to the x-axis which is close to the origin and $r_{x=xr}$ is the reflection coefficient for the wall opposite to that. Then the combined reflection coefficient for all the reflections made by the i^{th} virtual source along the x-axis can be found using the following equation.

$$r_{x_i} = r_{x=0}^{\left|\frac{1}{2}i-\frac{1}{4}+\frac{1}{4}(-1)^i\right|} r_{x=x_r}^{\left|\frac{1}{2}i+\frac{1}{4}-\frac{1}{4}(-1)^i\right|} \quad (6.1)$$

In the similar manner the combined reflection coefficients for the j^{th} and k^{th} virtual sources can be calculated using the equations below.

$$r_{y_i} = r_{y=0}^{\left|\frac{1}{2}j-\frac{1}{4}+\frac{1}{4}(-1)^j\right|} r_{y=y_r}^{\left|\frac{1}{2}j+\frac{1}{4}-\frac{1}{4}(-1)^j\right|} \quad (6.2)$$

$$r_{z_k} = r_{z=0}^{\left|\frac{1}{2}k-\frac{1}{4}+\frac{1}{4}(-1)^k\right|} r_{z=z_r}^{\left|\frac{1}{2}k+\frac{1}{4}-\frac{1}{4}(-1)^k\right|} \quad (6.3)$$

In order to find the total reflection coefficient of the virtual sources with indices *i*, *j*, and *k* simply multiply the reflection coefficients of the *i*, *j*, and k^{th} virtual sources along the *x*, *y* and *z*-axes. The equation which represents this is as follows

$$r_{ijk} = r_{x_i} r_{y_j} r_{z_k} \quad (6.4)$$

Now the total magnitude of each echo is calculated by multiplying the equations 5 and 6 together as shown in equation 7.

$$e_{ijk} = b_{ijk} r_{ijk} \quad (7)$$

Construction of the Impulse Response

In the final stage the impulse response will obtained by multiply unit impulse response function, total magnitude of each echo together and sum over all the three indices. This can be thought of as the summation of the all the sounds as they stream from all of the virtual sources. This is shown in the equation below [3,4].

$$h(t) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} \sum_{k=-n}^{n} a_{ijk} e_{ijk} \quad (8)$$

3.4 Fractional delay

Fractional delay filters are useful in numerous digital signal processing applications where accurate time delays are needed or the locations of sampling instants must be changed, such as in telecommunications, music synthesis, and speech coding . Many design methods have been proposed for fractional delay filters of FIR and IIR type. The transfer function of a digital all-pass filter is given by

$$H(z) = \frac{z^{-N}D(z^{-1})}{D(z)} \qquad (9)$$

Where N is the order of the filter and $D(z) = 1 + a_1Z^{-1} + a_2Z^{-2} + \ldots + a_NZ^{-N}$ is the denominator polynomial with real-valued coefficients a_k, and the numerator polynomial is a reversed version of the denominator.

The design of fractional delay all-pass filters is usually based on solving a set of linear equations, such as the least squares method proposed by Lang and Laakso. These methods produce optimal or very nearly optimal designs, but their usefulness is limited when high-order filters are needed or when coefficient values should be calculated online in a real-time application

Only one FD all-pass filter design method is known that can be implemented using closed-form formulas: the maximally flat group delay method that is based on Thiran's allpole filter design. A drawback of this method is that the fractional delay approximation is excellent only on a narrow band at low frequencies and a dramatic widening of the bandwidth of good approximation requires the filter order to be increased excessively.

In 1971, Thiran published a closed-form design method for all-pole filters that have a prescribed maximally flat group delay. Fettweis showed that the design formulas can be used for obtaining allpass filters that have the same property. When the desired group delay of an allpass filter is d, it is only necessary to make the substitution $d' = d/2$ in Thiran's formula, since the group delay of an allpass filter is twice that of its denominator. The Thiran design formula for a fractional delay allpass filter can be written as

$$H(z) = \frac{a_N z^N + a_{N-1} z^{N-1} + \cdots + a_1}{a_0 z^N + a_1 z^{N-1} + \cdots + a_N} \qquad (10)$$

The coefficients $a_0, \ldots, a_N$ are given by:

$$a_k = (-1)^K \binom{M}{k} \prod_{n=0}^{M} \frac{d+n}{d+k+n} \qquad (11)$$

Where d is the real-valued delay parameter and k = 1, 2, 3... N. Closed-form formulas that are at most Nth-order rational polynomials of delay d can be obtained from (11). For example, when N = 2, the filter coefficients are $a_1 = -2(D-2)/(D+1)$ and $a_2 = (D-1)(D-2)/(D+1)(D+2)$. Here a notation has been introduced $D = N + d$, where D denotes the group delay (in samples) that the allpass filter produces at low frequencies. In it was shown that the numerator polynomial $D(z)$ of the original Thiran all-pole filter has all its zeros inside the unit circle for $d > -0.5$. This implies that the allpass filter designed using Eq. (11) is stable for $d > -1$, because the group delay of the allpass filter is twice that of the numerator.

However, the error increases with frequency, and particularly in the case of low-order filters, the deviation soon becomes large. Also note that when the order of the filter is increased, the bandwidth of good approximation (error smaller than, e.g., 0.1 samples) is not becoming much wider. It is also of interest to examine the frequency response error (FRE) of the allpass filter as a measure of approximation quality. The following definition is used for the FRE:

$$E(e_{j\omega}) = e^{-j\omega(N+d)} - H(e_{j\omega}) \qquad (12)$$

where the first part on the right-hand side represents the frequency response of an ideal fractional delay filter producing a delay of N + d sampling intervals, and the second term is the frequency response of the allpass filter obtained from (9) using coefficients (11), which approximates a constant delay of N + d samples[5].

4. Filter Bank Design for Wiener Beamformer

4.1 Introduction

Filterbank analysis and synthesis strategies prove advantageous in many signal processing areas operating as a divide and conquer strategy tackling difficult problems into an equivalent series of much simpler problems. For example, large convolution systems encountered in applications such as echo cancellation and feedback cancellation may require a large number of filter taps. Using the filterbank technique, it may equivalently be implemented as a parallel combination of much shorter subband filters. When properly designed, the filterbank subband signals are minimally overlapping in frequency yielding signals that are approximately orthogonal to each other. Lately, digital filterbank techniques, with their great precision, have enabled many strategies to be implemented that were difficult or impractical with analog structures. Accordingly, much theory has been developed including the so-called perfect reconstruction filterbank.

An oversampled DFT filterbank using WOLA (weighted overlap-add) processing provides an extremely efficient and elegant solution. The WOLA filterbank structure is highly configurable and best performance is of course only achieved with an understanding of the optimizations and tradeoffs that can be made within its structure. The Weighted Overlap-Add (WOLA) filterbank discussed here is an important component that meets these difficult requirements.

4.2 WOLA filter bank

Over the last two decades, multi-rate digital signal processing techniques have been considerably developed and widely practiced in various engineering disciplines. The conditions to obtain perfect reconstruction maximally decimated (or critically sampled) filter banks have been extensively investigated and well-documented. Perfect reconstruction systems impose severe constraints that are not suitable in some applications. For applications requiring significant adjustment in the frequency bands, other structures are preferable. The WOLA structure can meet these design constraints.

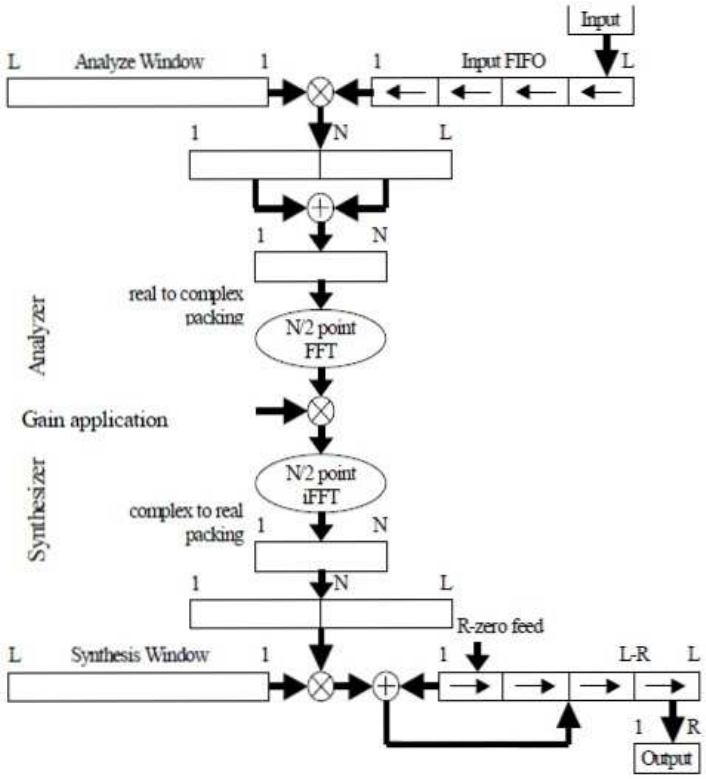


Figure.5: Simplified block diagram of a WOLA filter bank

WOLA structure Figure.5 shows a simplified block diagram of an oversampled WOLA filter bank [4], [6]. The input step size (*R*) is the FFT size (*N*) divided by the oversampling ratio (*OS*). The use of oversampling provides two benefits (*i*) the gain of the filter bank bands can be

adjusted over a wide range without the introduction of audible aliasing and (ii) a group delay versus power consumption trade-off can be made.

The Weighted Over-Lap Add (WOLA) filter bank is an efficient filter bank technique frequently used in DSPs. Four variables, together with an analysis window function $w[n]$, define the WOLA filter bank, namely; L the length of the analysis window, D the decimation rate (block rate), K the number of subbands, and D_F the synthesis window decimation rate.

The analysis stage, see figure.6, accepts a block of D new input data samples. Each new block is fed into an input FIFO buffer $u[n]$, of length L samples. The data in the input FIFO is element-wise weighted by the analysis window function and stored into a temporary buffer $t_1[n] = u[n].w[n]$, of length L samples. The temporary buffer is time-folded into another

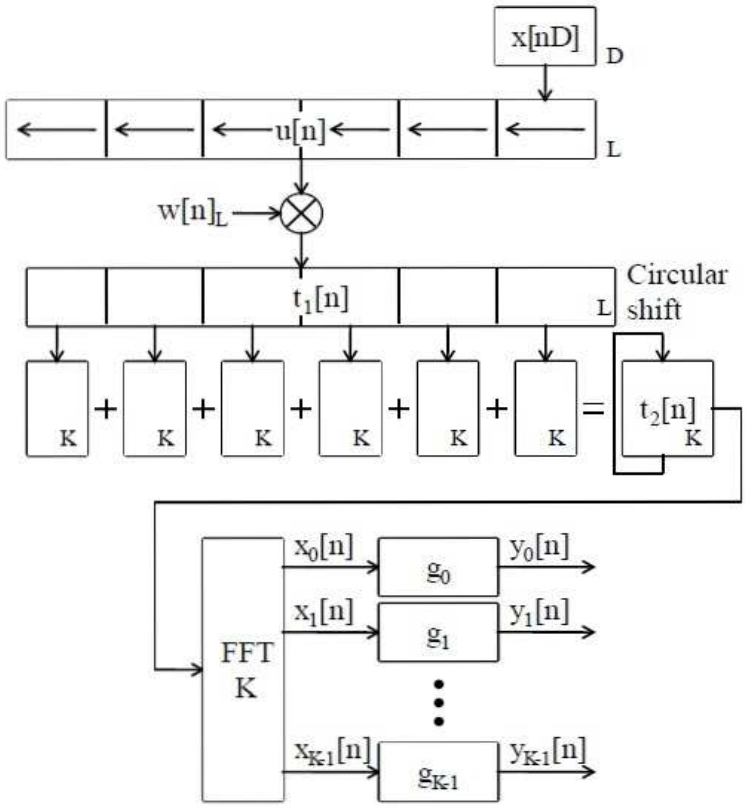


Figure.6: Analysis stage of a WOLA filter bank structure

temporary vector $t_2[n]$, of length K samples. The time-folding means that the elements of $t_1[n]$ are modulo-K-added to $t_2[n]$, according to

$$t_2[n] = \sum_{m=0}^{\frac{L}{K-1}} t_1[n + mK] \qquad (13)$$

The temporary buffer $t_2[n]$ is circularly shifted by $K/2$ samples in order to produce a zero-phase signal for the FFT. This means that the upper half of $t_2[n]$ is swapped place with the lower half of $t_2[n]$. The circularly shifted buffer $t_2[n]$ is then fed into a K-sized FFT to compute the subband signals $x_k[n]$.

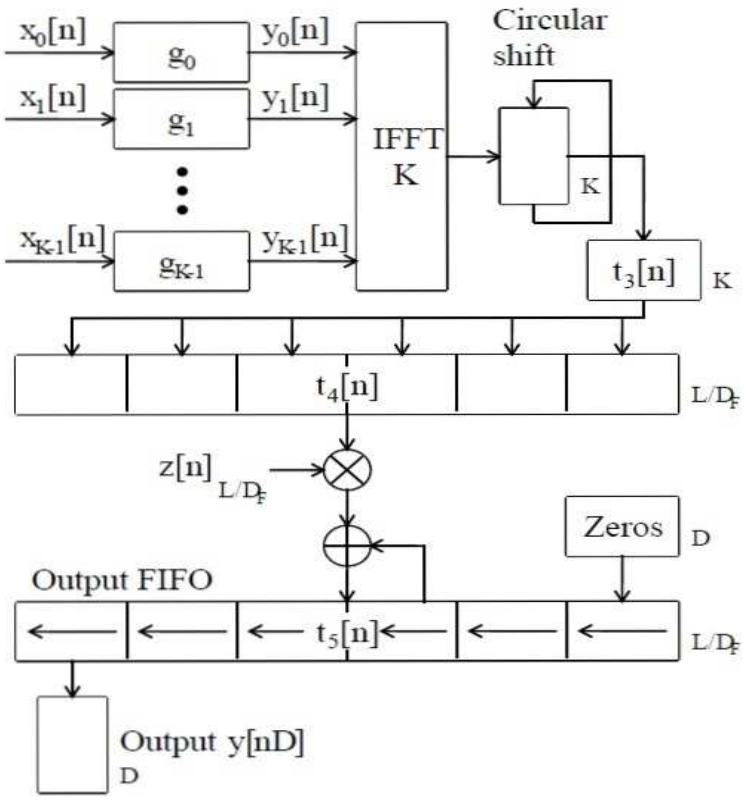


Figure.7: Synthesis stage of a WOLA filter bank structure

The synthesis stage of the WOLA filter bank implements the actual weighted overlap-add procedure, i.e., the WOLA procedure. The synthesis stage, see figure.7, starts by applying a size-*K* IFFT to the processed subband signals $y_k[n]$.

The IFFT output is circularly shifted *K*/2 samples, to counter-act the circular shift used in the analysis stage, and the circularly shifted data is stored in a temporary buffer $t_3[n]$, of size *K* samples. The buffer $t_3[n]$ is then stacked, by repetition, in the buffer $t_4[n]$ of length L/D_F, where L is the analysis window length, and D_F is the synthesis window decimation factor. The buffer $t_4[n]$ is weighted by a synthesis window function $z[n]$ of size L/D_F, defined as $z[n] = w[nD_F]$, i.e., a factor D_F decimated analysis window function. The weighted data is summed with the data in the output FIFO, $t_5[n]$ of length L/DF , and the output FIFO data is over-written with the summation result, i.e.,$t_5[n] \leftarrow t_5[n] + z[n].t_4[n]$. the output FIFO is then shifted left by D samples, i.e., D zeros are filled from the FIFO's rear, and the out-shifted data is the actual output data block, $y[nD]$ [6].

5. Microphone Array Beamforming techniques

5.1 Introduction

Microphone arrays spatially sample the sound pressure field. When combined with spatio-temporal filtering techniques known as *beamforming*, they can extract the information from (spatially constrained) signals, of which only a mixture is observed.

In this section an introduction to the principle of beamforming is first given, followed by a description of the classical beamforming techniques: the Delay-and-Sum beamformer and the Filter-and-Sum beamformer. Frequency and time domain bemaforming are discussed.

5.2 Beamforming

Beamforming is a signal processing technique used in sensor arrays for directional signal transmission or reception. This is achieved by combining elements in the array in a way where signals at particular angles experience constructive interference and while others experience destructive interference. Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity. The improvement compared with an omnidirectional reception/transmission is known as the receive/transmit gain (or loss) [8].

A beamformer combines sampled data from each transmitted element the same way an FIR filter would combine temporally sampled information. Beamformers are of two general types, a narrowband beamformer and a wideband beamformer. A narrowband beamformer is shown in the Figure 8. The output of the beamformer is given by,

$$z(k) = \sum_{i=1}^{M} w_i^* y_i(k) \qquad (14)$$

Where the complex weights *w* are used to steer the beam towards the desired user and steer nulls towards interferers. The above equation can be written in vector form as

$$z(k) = W^H y(k) \qquad (15)$$

A wideband beamformer is often used when signals of wide band are of interest and is more complex than a narrow band beamformer.[7 new].

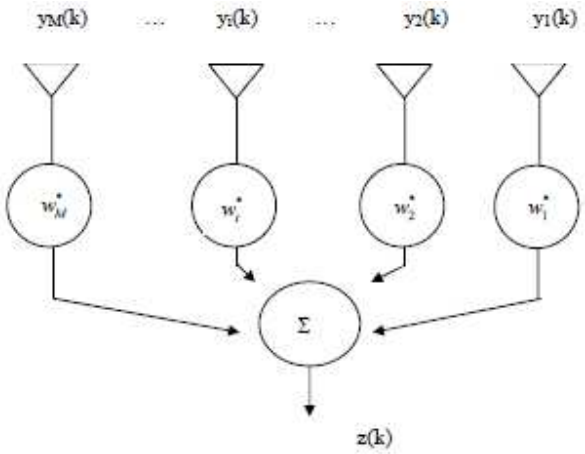


Figure.8: Narrowband beamformer

5.3 Classical Beamformers

The complex weighting element w_m in the far-field horizontal directivity pattern of a linear sensor array can be expressed in terms of its magnitude and phase components as

$$W_m = a_m e^{j\varphi m} \qquad (16)$$

The directivity pattern $B(\Omega, \theta)$ is

$$B(\Omega, \theta) = \sum_{m-1}^{M} w_m e^{j\Omega(\mathrm{m}-1)\mathrm{d}\cos(\theta)/\mathrm{c}} \qquad (17)$$

the above equation is reformulated as

$$B(\Omega, \theta) = \sum_{m-1}^{M} w_m e^{j[\Omega(\mathrm{m}-1)\mathrm{d}\cos(\theta)/\mathrm{c}+\varphi_\mathrm{m}} \qquad (18)$$

While the amplitude weights a_m control the shape of the directivity pattern, the phase weights ϕ_m control the angular location of the response's main lobe. Beamforming techniques are algorithms for determining the complex sensor weights w_m in order to implement a desired shaping and steering of the array directivity pattern.

5.3.1 Delay-and-sum beamformer

The time-domain implementation as shown in Fig.9, the delay and sum beamformer (Time domain beamformer) basically consists on the alignment of the different microphone signals to compensate for the different path lengths from the source to the various microphones, and the combination of these aligned signals together. It can be expressed mathematically as follows:

$$y(n) = \sum_{q=1}^{Q} \propto_q x_q(n - \tau_q) \qquad (19)$$

where $\propto_q$ is the weight given to each different microphone and τ_q is the delay that compensates the different propagation delays. Usually, the weight $\propto_q$ is equal to 1/Q resulting in the average of the aligned signals, however it is possible to select other criteria for microphone weight for instance depending on the propagation model, compensation of different sensor gain or even different signal to noise ratio. Obtaining τ_q is a problem of time delay estimation or more generally of speaker localization. The simplicity of the delay-and-sum beamformer is the most important strength, resulting in many cases a convenient and practical choice for many microphone array applications. Thus, delay-and-sum is widely used despite its frequency depending response, the impossibility of reducing highly directive noise sources.

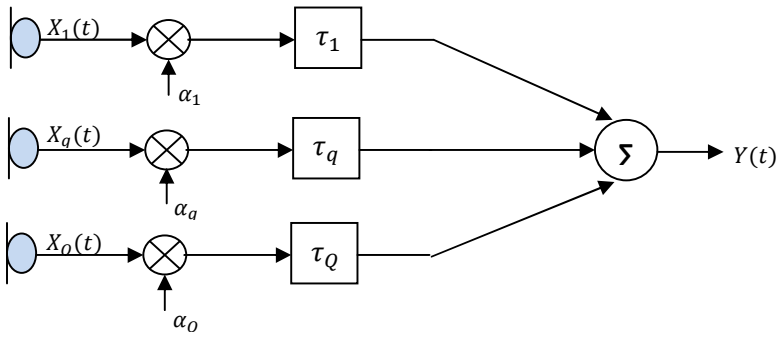


Figure.9. Delay and sum beamformer in a time domain implementation.

5.3.2 Filter-and-Sum Beamformer (FSB)

In the filter-and-sum beamformer (FSB), both the amplitude and the phase of the complex weights are frequency dependent, resulting in a filtering operation of each array element input signal. The filtered channels are then summed. Using the following vector notations

$$W(\Omega) = [\mathrm{W}_1(\Omega), \mathrm{W}_2(\Omega) \ldots\ldots \mathrm{W}_\mathrm{M}(\Omega)],$$
$$X(\Omega) = [\mathrm{X}_1(\Omega), \mathrm{X}_2(\Omega), \ldots. \mathrm{X}_\mathrm{M}(\Omega)],$$

the array output is given by

$$Y(\Omega) = W(\Omega)^{\mathrm{H}} X(\Omega).$$

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

$$y(n) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} w_m(l) x_m(n-l) \qquad (20)$$

where $X_m(n)$ are sampled observations from sensor m, $W_m(l)$ ($l = 0, 1, \ldots, L-1$) are the filter weights for channel m and L is the filter length.

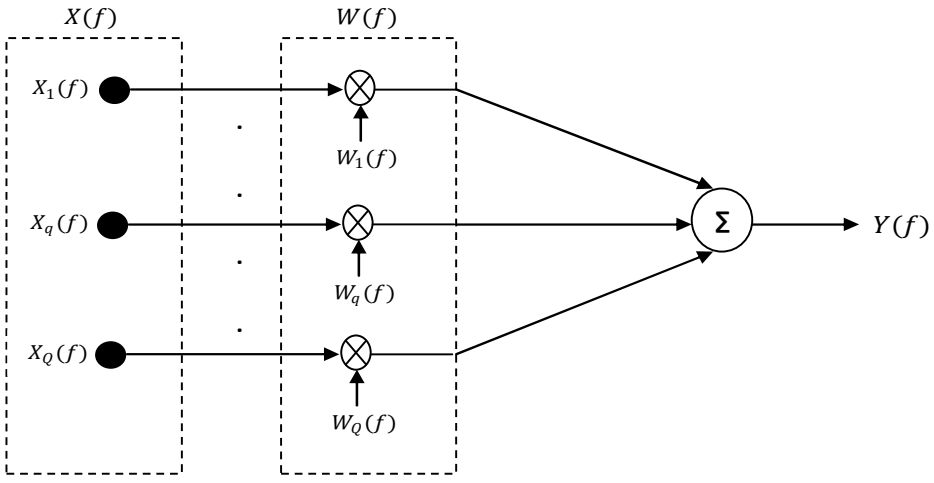


Figure. 10. Filter and sum beamforming in the frequency domain of frequency bin *f*.

Beamforming techniques can be broadly divided into two categories:

- conventional beamformers
- adaptive beamformers

5.4 Conventional beamformers

Conventional beamformers use a fixed set of weightings and time-delays (or phasings) to combine the signals from the sensors in the array, primarily using only information about the location of the sensors in space and the wave directions of interest. In contrast, adaptive beamforming techniques generally combine this information with properties of the signals actually received by the array, typically to improve rejection of unwanted signals from other directions. This process may be carried out in either the time or the frequency domain.

5.5 Adaptive beamformers

As the name indicates, an adaptive beamformers is able to automatically adapt its response to different situations. Some criterion has to be set up to allow the adaption to proceed such as minimising the total noise output. Because of the variation of noise with frequency, in wide band systems it may be desirable to carry out the process in the frequency domain.

Different types of adaptive beamformers or phased arrays are as follows

- Time domain beamformers
- Frequency domain beamformers

A time domain beamformer works by doing time based operations. The basic operation in this time domain beamformer is “delay and sum”. It delays the incoming signal from each array element by a certain period of time and adds them together. Sometimes multiplication is done with a window across the array to increase the main lobe/side lobe ratio, and also to insert zeroes in the characteristic.

The frequency domain beamforming is once again classified into two types.

The first type separates the different frequency components that are present in the received signal into different frequency bins or bands (using either an FFT or filter bank). When different delay and sum beamformers are applied to each frequency bin or band, it is possible to

point the main lobe to different directions for different frequencies making this approach more flexible.

The other type of frequency domain beamformers makes use of spatial frequency. It means that an FFT is taken across different array elements, but not in time. Hence the output of the N point FFT is N channels evenly divided in space. This approach is not flexible as different directions are fixed [13].

Beamforming finds solution in aperture synthesis, ISA, phased array antennas, microphone arrays, synthetic aperture radar, synthetic aperture sonar etc. [14].

A microphone array is any number of microphones operating in tandem. There are many applications:

1. Systems for extracting voice input from ambient noise (notably telephones, speech recognition systems, hearing aids).
2. Surround sound and related technologies.
3. Locating objects by sound: acoustic source localization, e.g. military use to locate the source(s) of artillery fire. Aircraft location and tracking.
4. High fidelity original recordings.

Typically, an array is made up of omnidirectional microphones distributed about the perimeter of a space, linked to a computer that records and interprets the results into a coherent form. Arrays may also be formed using numbers of very closely spaced microphones. Given a fixed physical relationship in space between the different individual microphone transducer array elements, simultaneous DSP processing of the signals from each of the individual microphone array elements can create one or more "virtual" microphones. Different algorithms permit the creation of virtual microphones with extremely complex virtual polar patterns and even the possibility to steer the individual lobes of the virtual microphones patterns so as to home-in-on, or to reject, particular sources of sound [15].

5.6 Optimal Beamformers

Optimal beamformer consists of set of microphones placed at different locations in order to sample the sound pressure field. These beamformers focus on minimizing the mean-square error between a reference signal which is highly correlated with the desired signal i.e. speech signal and the output signal. These beamformers mainly concentrate on obtaining the reference speech signal with a good correlation to the desired speech signal but doesn't keep constraint on distortion of the signal [18]. This degraded signal in high noise field environments can be enhanced by considering a limitation on the filter weights based on the knowledge of the source position in order to restrict the path of the desired speech signal. These filter weights are calculated and the optimization is measured based on the powers of signal and noise signals i.e. Signal-to-noise ratio (SNR).

The optimal beamformer is designed based on the power criteria by taking the observed microphone signals which consists of speech signal and noise signal. These beamformers optimizes the array output by adjusting the weights at output which contains minimum contributions from noise and interference. The optimum weights and Signal-to-Noise Ratio (SNR) are generated and calculated by considering the numerical methods to solve the generalized eigenvector problem. In the reverberant noise field environments assumptions, based on the type of environments the optimization procedure simplifies and closed form solution is obtained based on the existence of matrix inversion. The knowledge of signal source location is the basic aspect in finding the closed form solution for optimum beamformer.

In this context, we are dealing with study of optimal beamformers such as Wiener Beamformer and Maximum Signal-to-Noise Ratio (SNR) in time-domain. The optimum beamformers results a closed form solution if the reference signal is continuously accessible which is considered as serious constraint. In this case, the performance of each beamformer is mainly based on SNR, Speech Distortion (SD), Noise Distortion (ND) and PESQ. In this thesis, the output of the beamformer is obtained based on various parameters such as number of microphones, distance between the microphones (D), the speech source and noise positions i.e. angles at which they are placed from microphone.

Generally, received microphone speech signal in case of wide-band signal we need a beamformer that can delay a range of frequencies. This delay can be achieved by using digital linear filters at each microphone signal. The process of making the speech signal to pass from location of speaker and attenuating the signals arriving from other locations is known as broadband Beamforming. In this context, we are using broadband Beamforming in case of wiener beamformer and maximum SNR beamformer.

5.7 Wiener filtering for microphone arrays

In 1988 Zelinski proposed an adaptive Wiener post-filter with delay-and-sum beamformer as the one shown in Figure. It is shown that incorporating a post-filter with the beamformer allows use of knowledge obtained with spatial filtering and also allows effective frequency filtering of the signal, resulting in a both spatial and frequency enhancement.

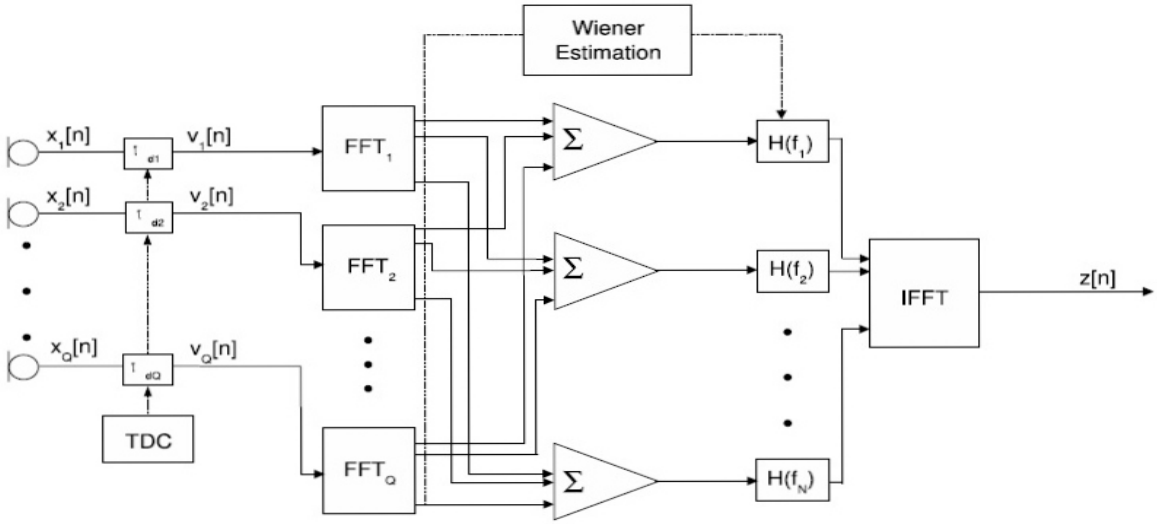


Figure. 11. Multichannel Wiener post-filter of a delay and sum beamformer. The received speech signals are previously time-aligned by a TDC module.

The general wiener post filter is formulated in terms of the cross spectral densities of noise at the beamformer output and the desired source as:

$$H(f) = \frac{\emptyset_{ss}(f)}{\emptyset_{ss}(f)+\emptyset_{nn}(f)} \qquad (21)$$

Where $\emptyset_{ss}(f)$ and $\emptyset_{nn}(f)$ are the auto spectral densities of the signal and the noise.

The wiener filter can be estimated as there is availability of multiple inputs that permits computing the power spectral density of the target signal and the one of the noise combining the cross power spectral densities and the power spectral density of the different microphones of the array. Assume that the received signal is an additive mixture of the desired signal and noise. They are uncorrelated and that noise is uncorrelated also between microphones and have an equal power spectral density, then

$$\emptyset_{v_i v_j}(f) = \emptyset_{ss}(f) \tag{22}$$

$$\emptyset_{v_i v_j}(f) = \emptyset_{ss}(f) + \emptyset_{nn}(f) \tag{23}$$

In this case wiener filter equation can be estimated by averaging as

$$H(f) = \frac{\frac{2}{Q(Q-1)} R\left\{\sum_{i=1}^{Q-1}\sum_{j=i+}^{Q} \widehat{\emptyset}_{v_i v_j}(f)\right\}}{\frac{1}{Q}\sum_{i=1}^{Q} \widehat{\emptyset}_{v_i v_i}(f)} \tag{24}$$

The real operator $R\{.\}$ is used because the term being estimated in the numerator, $\emptyset_{ss}(f)$ is necessarily real. It should be noted that the denominator in fact provides an over-estimate of the noise power at the beamformer output, as it is calculated using the input signals.

It is clear from the above assumptions the post-filter is particularly convenient in presence of spatially white noise, however it is also useful in diffuse noise fields which reasonably approximate these conditions. In fact, the post-filter was deeply studied in general case of a filter-and-sum beamformer and was also studied for different non-ideal conditions in terms of beamformer characteristics, such as noise reduction and array gain. In this work, it was shown that the post-filter is effectively able to cancel any incoherent noise, that the rejection to coherent noise correlated and uncorrelated with the desired signal is improved if they are not arriving from the same direction and that it is robust to minor steering errors. The general expression of the Wiener post-filter for any beamformer is:

$$H(f) = \frac{\sum_{i=1}^{Q} |w_i(f)|^2}{\sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} w_i(f) w_j{}^*(f)} \frac{R\left\{\sum_{i=1}^{Q-1} \sum_{j=i+}^{Q} \widehat{\emptyset}_{v_i v_j}(f)\right\}}{\frac{1}{Q} \sum_{i=1}^{Q} \widehat{\emptyset}_{v_i v_i}(f)} \qquad (26)$$

The Zelinski post-filter has been extensively used in microphone array works, for instance as part of a GSC-like beamformer or in combination with a speech dereverberation technique based in the separate processing of the minimum-phase and all-pass components of the input speech signal.

6. The Steered Response Power (SRP)

6.1 Introduction

Array signal processing techniques rely on the ability to *focus* on signals originating from a particular location or direction in space. Most of these techniques employ some type of *beamforming*, which generally includes any algorithm that exploits an array's sound-capture ability. Beamforming, in the conventional sense, can be defined by a *filter-and-sum* process, which applies some temporal filters to the microphone signals before summing them to produce a single, focused signal. These filters are often adapted during the beamforming process to enhance the desired source signal while attenuating others. The simplest filters execute time shifts that have been matched to the source signal's propagation delays. This method is referred to as *delay-and-sum* beamforming; it delays the microphone signals so that all versions of the source signal are time-aligned before they are summed. The filters of more sophisticated filter-and-sum techniques usually apply this time alignment as well as other signal-enhancing processes.

Beamforming techniques have been applied to both source-signal capture and source localization. If the location of the source is known (and perhaps something about the nature of the source signal is known as well), then a beamformer can be focused on the source, and its output becomes an enhanced version (in some sense) of the inputs from the microphones. If the location of the source is not known, then a beamformer can be used to scan, or *steer*, over a predefined spatial region by adjusting its steering delays (and possibly its filters). The output of a beamformer, when used in this way, is known as the *steered response*. The steered response power (SRP) may peak under a variety of circumstances, but with favorable conditions, it is maximized when the steering delays match the propagation delays.

Beamforming has been used extensively in speech-array applications for voice capture. For this application, the filters applied by the filter-and-sum technique must not only suppress the background noise and contributions from unwanted sources, they must also do this in way that does not significantly distort the desired signal. However, when beamforming techniques are applied to source-localization, these filters need only boost the power of the desired source signal

in the beamformer's output when the array is focused on it. This important distinction is exploited in this chapter where a new type of filter is proposed for localization. These filters are derived from the phase transform (PHAT), which applies a magnitude-normalizing weighting function to the cross-spectrum of two microphone signals. This procedure produces a function that is useful for TDOA estimation but is obviously a distortion of the input (and source) signals. In the same way, beamformer filters can be designed to produce a steered response that is useful for source localization but not for voice-capture.

This chapter discussed the application of filters that makes the steered response power (SRP) equivalent to the sum of all possible pair wise phase transforms. The new technique, which has been dubbed "SRP-PHAT", exploits microphone redundancy by combining the microphone signals, rather than combining a multitude of TDOA estimates, to enhance the accuracy of location estimation.

6.2 Conventional Beamformers

Conventional beamformers use a fixed set of weightings and time-delays (or phasings) to combine the signals from the sensors in the array, primarily using only information about the location of the sensors in space and the wave directions of interest. In contrast, adaptive beamforming techniques generally combine this information with properties of the signals actually received by the array, typically to improve rejection of unwanted signals from other directions. This process may be carried out in either the time or the frequency domain.

A microphone array having M received signals, defined as

$$x_n(t) = s(t) * h(d_s, t) + n_n(t) \qquad (27)$$

Where $x_n(t)$ consist of delayed, filtered and noise signal $s(t)$. A delay-and-sum beamformer will align all the microphone‟s input by giving an appropriate steering delay $\delta_n$to each microphone input $x_n(t)$and summing all the inputs to get an unmodified signal from a spatial locationq_s.

The conventional delay-and-sum beamformer is defined as:

$$y(t, q_s) = \sum_{n=1}^{M} x_n(t - \delta_n) \qquad (28)$$

The delay-and-sum beamformer will give an output $y(t, q_s)$, which is overall delayed signal from all microphones. The delay δ_n is estimated and computed individually between all microphone pairs, which make the operation causal for this practical system.

In ideal environment delay-and-sum beamformer gives scaled and summed version of desired signal. However, in a real time environment channel characteristics are not even, which degrades the efficiency of delay-and-sum beamformers. One reason of degradation could be additive noise. Adaptive filters are used to minimize the noise in the input signal of each microphone. Microphone signals are first filtered and then computed in the delay-and-sum beamformer to get the desired output.

The conventional filter-and-sum beamformer output in frequency domain can be defined as:

$$Y(\omega, q) = \sum_{n=1}^{M} G_n(\omega) X_n(\omega) e^{-j\omega\delta_n} \qquad (29)$$

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 compensate the environmental noise and channel effect for some means in real time environment, but yet it is not too much robust for practical scenarios.

6.3 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. Power of filter-and-sum beamformer can be expressed in frequency domain as:

$$P(q) = \int_{-\infty}^{\infty} Y(\omega, q)\omega Y^*(\omega, q) d\omega \qquad (30)$$

Inserting filter-and-sum beamformer output in frequency domain equation in equation (30)

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{l=1}^{M} G_l(\omega) X_l(\omega) e^{-j\omega\delta_l} \right) \left(\sum_{l=1}^{M} G_K^*(\omega) X_K^*(\omega) e^{-j\omega\delta_k} \right) d\omega \qquad (31)$$

Rearranging the expression, get:

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{l=1}^{M} \sum_{k=1}^{M} \left(G_l(\omega) G_K^*(\omega) (X_l(\omega) X_K^*(\omega)) \right) e^{-j\omega(\delta_k - \delta_l)} \right) d\omega \qquad (32)$$

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

$$\tau_{kl} = \delta_k - \delta_l \qquad (33)$$

$$P(q) = \int_{-\infty}^{\infty} \left(\sum_{l=1}^{M} \sum_{k=1}^{M} \left(G_l(\omega) G_K^*(\omega) (X_l(\omega) X_k^*(\omega)) \right) e^{-j\omega\delta_{kl}} \right) d\omega \qquad (34)$$

Weighting function can be defined for filter as:

$$\Psi_{lk}(\omega) = G_l(\omega) G_K^*(\omega) \qquad (35)$$

The integral is on the filter and the microphone input signals for a finite length, rearranging the equation 34 get:

$$P(q) = \sum_{l=1}^{M} \sum_{k=1}^{M} \int_{-\infty}^{\infty} \Psi_{lk}(\omega) X_l(\omega) X_k^*(\omega) \, e^{-j\omega\delta_{kl}} \, d\omega \qquad (36)$$

The peak of the SRP indicates the location of the sound source. The strong reflection of the sound source sometime also gives peak, which indicates the wrong location of the sound source. The complication for the search of global maxima also increases by these strong reflections of the sound source.

6.4 SRP-PHAT

The strong reflections of the sound source can result the wrong DOA. To minimize this error, a weighting function phase alignment transform (PHAT), which is robust in real time environment [10] is applied on SRP to estimate the correct DOA of sound source. PHAT is not robust under high reverberant environment, but it is still effective under low and moderate reverberant conditions.

A generalized SRP-PHAT for speaker localization is defined in equation (39) can be modified by changing the summation limits to minimize the computations. The modified equation is

$$P(q) = \sum_{l=1}^{M} \sum_{k=l+1}^{M} \int_{-\infty}^{\infty} \Psi_{lk}(\omega) X_l(\omega) X_k^*(\omega)\, e^{-j\omega\delta_{kl}}\, d\omega \qquad (37)$$

The PHAT weighting functions can be defined as

$$\Psi_{lk}(\omega) = \frac{1}{|X_l(\omega) X_k^*(\omega)|} \qquad (38)$$

Where $\Psi_{lk}(\omega)$ 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:

$$G_l(\omega) G_K^*(\omega) = \frac{1}{|X_l(\omega) X_k^*(\omega)|} \qquad (39)$$

Inserting equation 38 in equation 37, get

$$P(q) = \sum_{l=1}^{M} \sum_{k=l+1}^{M} \int_{-\infty}^{\infty} \frac{1}{|X_l(\omega) X_k^*(\omega)|} X_l(\omega) X_k^*(\omega)\, e^{-j\omega\delta_{kl}}\, d\omega \qquad (40)$$

τ_{kl} is the time delay between microphone k and microphone l. The far field assumption is used for the calculation of theτ_{kl}. Planner sound waves will arrive from the speaker at microphone array as shown in the figure below.

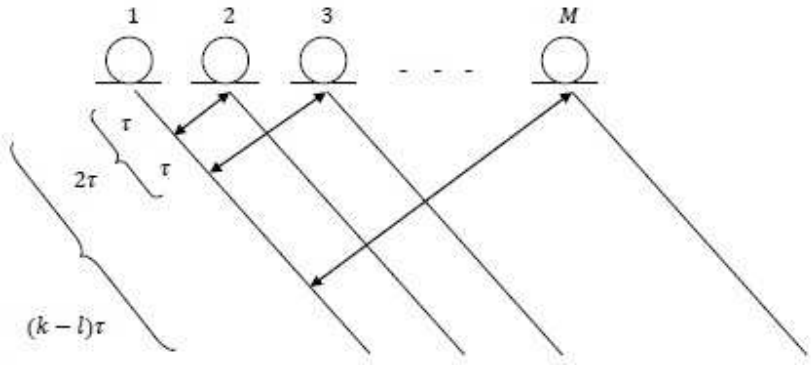


Figure.12: Planner Waves arrival from a Far Field Sound Source.

6.5 TDOA Estimation using SRP-PHAT

To estimate the speaker location, TDOA τ_s should be first estimated. The GCC-PHAT algorithm used in [11] is defined as

$$\tau_s = arg\max_{\tau_{kl}} P_{kl}(\tau_{kl}) = arg\max_{\tau_{kl}} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{|X_l(\omega) X_k^*(\omega)|} X_l(\omega) X_k^*(\omega) e^{j\omega\delta_{kl}} \, d\omega \right) \quad (43)$$

By inserting equation 42 in equation 43 get

$$\tau_s = arg\max_{\tau} P(\tau)$$

$$= arg\max_{\tau}\left(\sum_{l=1}^{M}\sum_{k=l+1}^{M}\int_{-\infty}^{\infty}\frac{1}{|X_l(\omega)X_k^*(\omega)|}X_l(\omega)X_k^*(\omega)\,e^{j\omega\tau(k-l)}\,d\omega\right) \quad (44)$$

The TDOA τs will be the value which will give the maximum output power of SRP-PHAT. This SRP-PHAT algorithm has only one parameter output $\propto$, which indicates the DOA of sound source as expressed below [12]

$$\propto = \sin^{-1}\left(\frac{\mathrm{v} * \tau_s}{\mathrm{d} * f_s}\right) \quad (45)$$

7. Results

This chapter will discuss about the results of this research. Primitively it discuss about the room impulse response with fractional delay filter based on Thiran approximation.

Consider the room coordinates as 5x2x1 with sample frequency (Fs) 8000 KHz and with discrete reflection coefficient(γ) can observe different types of different plots of the room impulse response showing the energy decay is shown.

Initially the RIR with reflection coefficient value of zero has been plotted. If the reflection coefficient is zero it means that the room absorbs each and everything so there will be no RIR plotted.

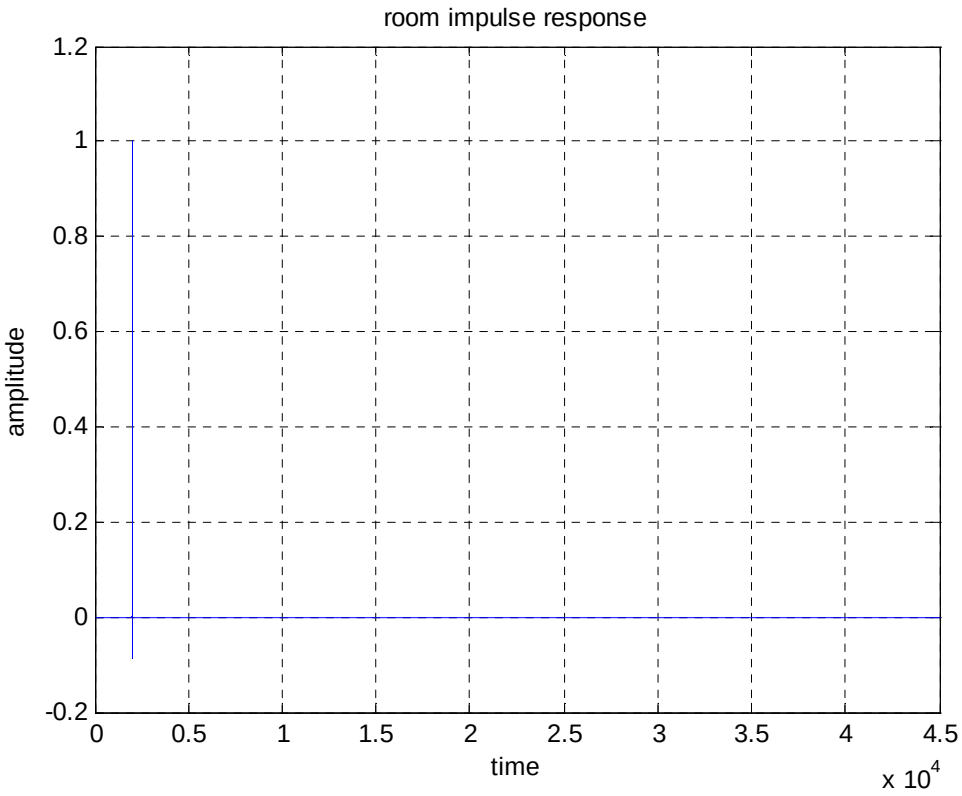


Figure.13: Room Impulse Response for reflection coefficient $\gamma = 0$

Similarly, by varying the reflection coefficient i.e., ($\gamma = 0.6, 0.8, 0.95$) room impulse response plots are shown below

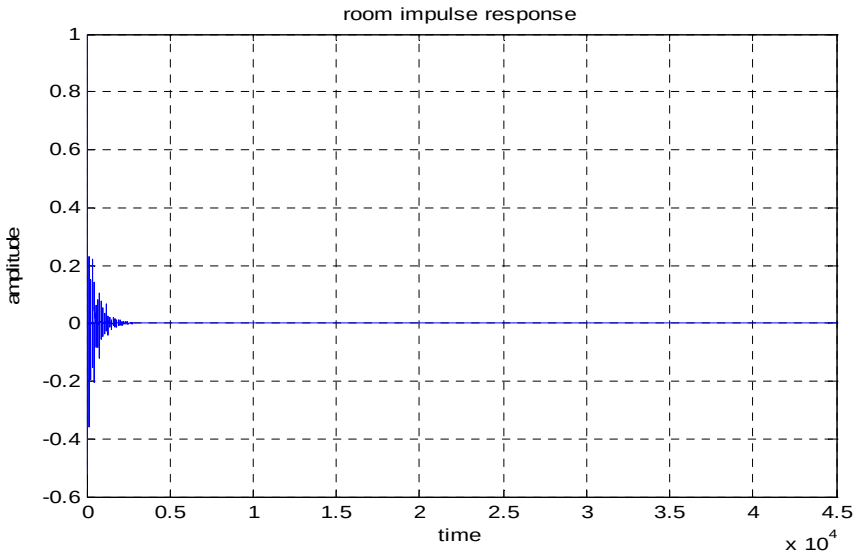


Figure.14: Room Impulse Response for reflection coefficient $\gamma = 0.8$

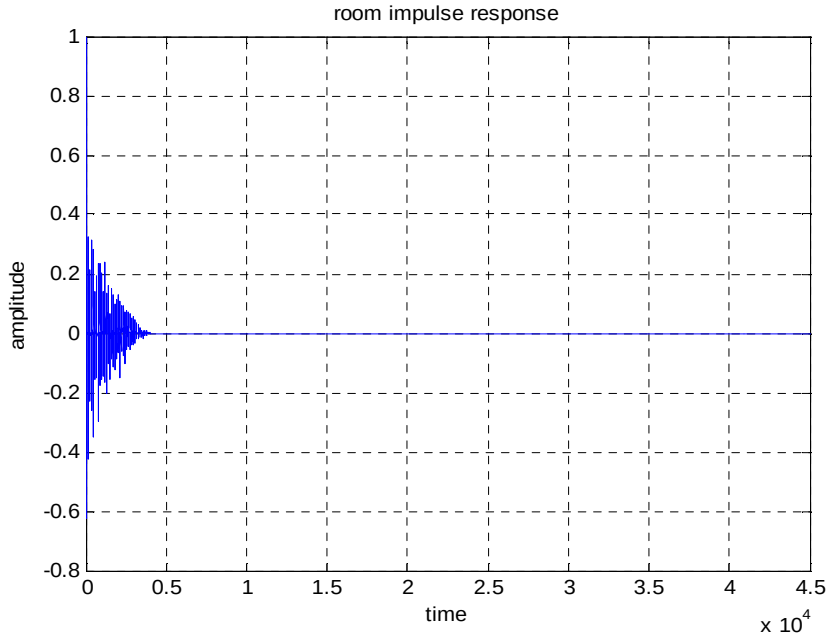


Figure.15: Room Impulse Response for reflection coefficient $\gamma = 0.95$

After the implementation of RIR a delay and sum beamformer and filter and sum beamformer are designed. By means of beamforming it is possible to spatially filter signals arriving to a microphone array by enhancing fixed desired directions while others are rejected thus reducing the noise. In the proposed beamforming design the noise is reduced and was able to produce a clean desired speech.

PESQ (Perceptual Evaluation of Speech Quality) is another objective measurement tool that predicts the results of subjective listening tests on telephony systems. PESQ uses a sensory model to compare the original, unprocessed signal with the degraded signal from the network or network element. The resulting quality score is analogous to the subjective MOS measured using panel tests according to ITU-T P.800. The PESQ scores are calibrated using a large database of subjective tests. The most eminent result of PESQ is the MOS. It directly expresses the voice quality. The PESQ MOS as defined by the ITU recommendation P.862 ranges from 1.0 (worst) up to 4.5 (best) [16].

PESQ takes into account coding distortions, errors, packet loss, delay and variable delay, and filtering in analogue network components. The user interfaces have been designed to provide a simple access to this powerful algorithm, either directly from the analogue connection or from speech files recorded elsewhere.

For Delay and Sum beamformer a room is designed with dimensions 6x4x2.8. Four microphones are arranged linearly. The distance between the each microphone will be same. The distance between each microphone is varied for each time and the SNR improvement and the PESQ values are determined for AWN and wind noise. The noise in first case is the wind noise.

The reflection coefficient value is 0.3 and the sampling frequency is 16000Hz. The position of the first mic is 4x2x1 and the source is located at 5x2x1. The position of the noise source is 3x1x0.5.

TABLE. 7. 1. THE SNR IMPROVEMENT AND PESQ SCORE FOR WIND WITH R=0. 3 AND ROOM DIMENSIONS 8X8X8.

Distance between mics	SNR input(dB)	SNR output(dB)	SNR Improvement	Input PESQ	Output PESQ
0.01	13.6721	32.0933	18.4211	1.983	3.691
0.015	13.6721	32.7745	19.1023	1.983	3.816
0.012	13.6721	32.3697	18.6976	1.983	3.763
0.02	13.6721	32.2902	18.6181	1.983	3.758

Next, by changing the room dimension to 6x4x2.8, the SNR improvement values and the PESQ scores are noted. The values obtained are as follows.

TABLE. 7. 2. THE SNR IMPROVEMENT AND PESQ SCORE FOR WIND NOISE WITH R=0. 3 AND ROOM DIMENSIONS 6X4X2.8.

Distance between mics	SNR input	SNR output	SNR improvement	Input PESQ	Output PESQ
0.01	13.4488	27.2124	13.7636	1.983	3.324
0.015	13.4488	29.0652	15.6165	1.983	3.528
0.012	13.4488	28.3333	14.8845	1.983	3.456
0.02	13.4488	29.5738	16.1250	1.983	3.592

After this, the wind noise is replaced with AWN but with the same reflection coefficient. Then the values obtained are as follows. The room dimensions are 6x4x2.8 and the reflection coefficient value is 0.3.

TABLE. 7. 3. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH R=0. 3 AND ROOM DIMENSIONS 8X8X8.

Distance between mics	SNR input	SNR output	SNR improvement	Input PESQ	Output PESQ
0.01	1.1017	24.8423	23.7406	2.140	3.182
0.015	1.0689	26.0943	25.0254	2.140	3.461
0.012	1.0890	26.3819	25.2929	2.140	3.537
0.02	1.0740	24.9300	23.8560	2.140	3.237

In the similar manner the reflection coefficient is varied from 0.3 to 0.6 and observed whether there is an improvement in the SNR. The values obtained are as follows.

TABLE. 7. 4. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH R=0. 6 AND ROOM DIMENSIONS 8X8X8.

Distance between mics	SNR input	SNR output	SNR improvement	Input PESQ	Output PESQ
0.015	-8.8049	15.0448	23.8497	2.140	2.379
0.012	-8.7857	15.4066	24.1923	2.140	2.356
0.02	-8.7687	14.9060	23.6747	2.140	2.337

TABLE. 7. 5. THE SNR IMPROVEMENT AND PESQ SCORE FOR WIND NOISE WITH ROOM DIMENSIONS 6X4X2.8 AND VARYING REFLECTION COEFFICIENTS WITH DISTANCE BETWEEN THE MICS 0.02M.

Reflection coefficient R	SNR improvement	Input PESQ	Output PESQ
0.2	18.9284	1.983	3.832
0.3	16.1250	1.983	3.592
0.4	10.6512	1.983	2.935
0.5	8.0175	1.983	2.623
0.7	5.1616	1.983	2.018

TABLE. 7. 6. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH ROOM DIMENSIONS 6X4X2.8 AND VARYING REFLECTION COEFFICIENTS WITH DISTANCE BETWEEN THE MICS 0.02M.

Reflection coefficient R	SNR improvement	Input PESQ	Output PESQ
0.2	25.1984	2.140	3.594
0.3	23.8560	2.140	3.237
0.4	22.9259	2.140	2.894
0.5	22.6913	2.140	2.588
0.7	25.4246	2.140	2.076

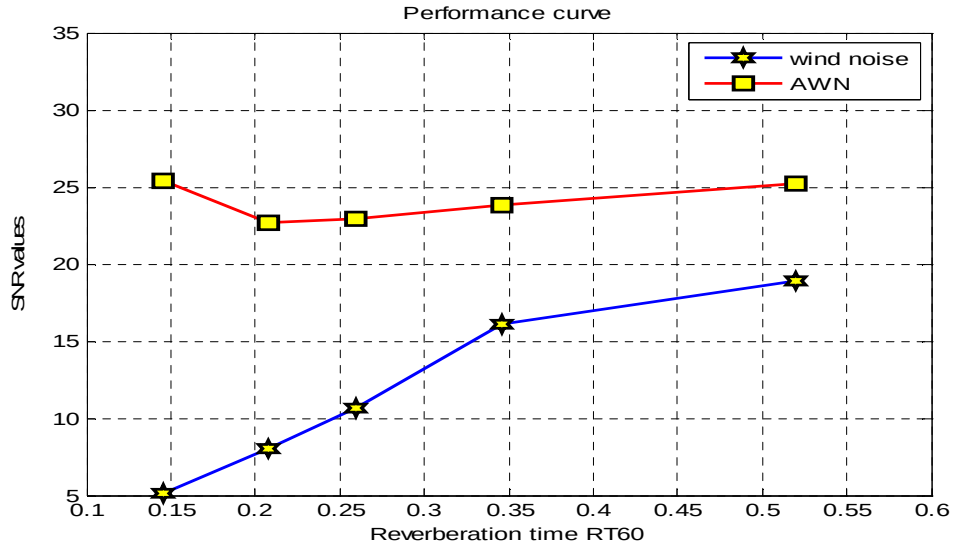


Figure. 7. 8. Performance curve showing performance with respect to reverberation time RT60.

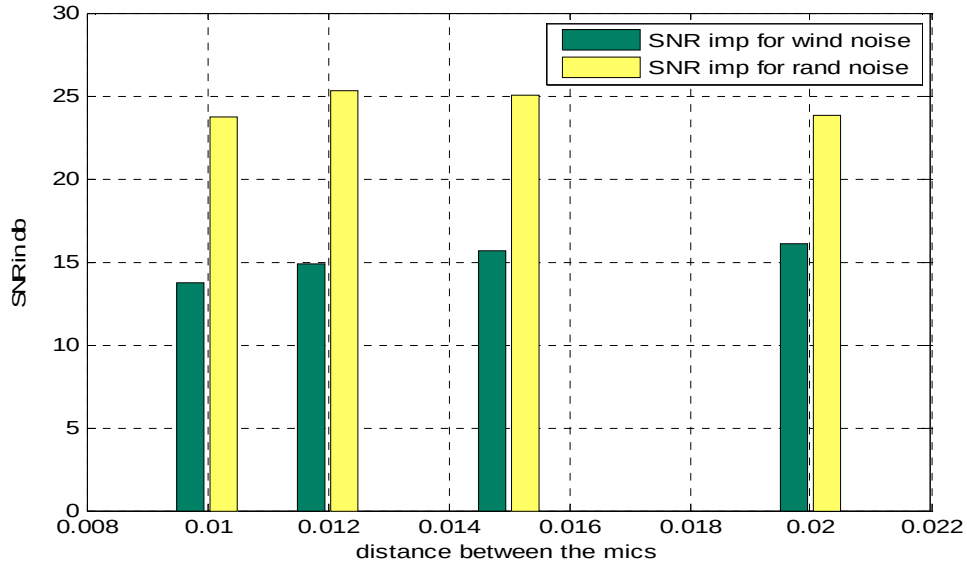


Figure. 7. 9. Representation of SNR improvement through blocks for both wind noise and random noise in time domain.

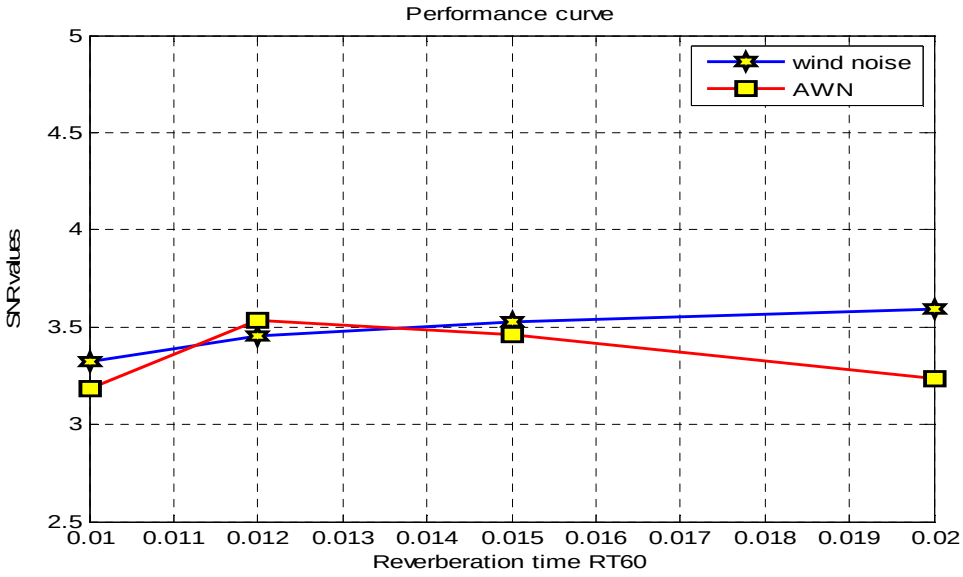


Figure. 7. 10. PESQ score for wind noise and AWN in time domain.

The input and output power spectrums of the beamformer when wind noise is given with speech signal are in Figure. 7. 11 and Figure. 7. 12.

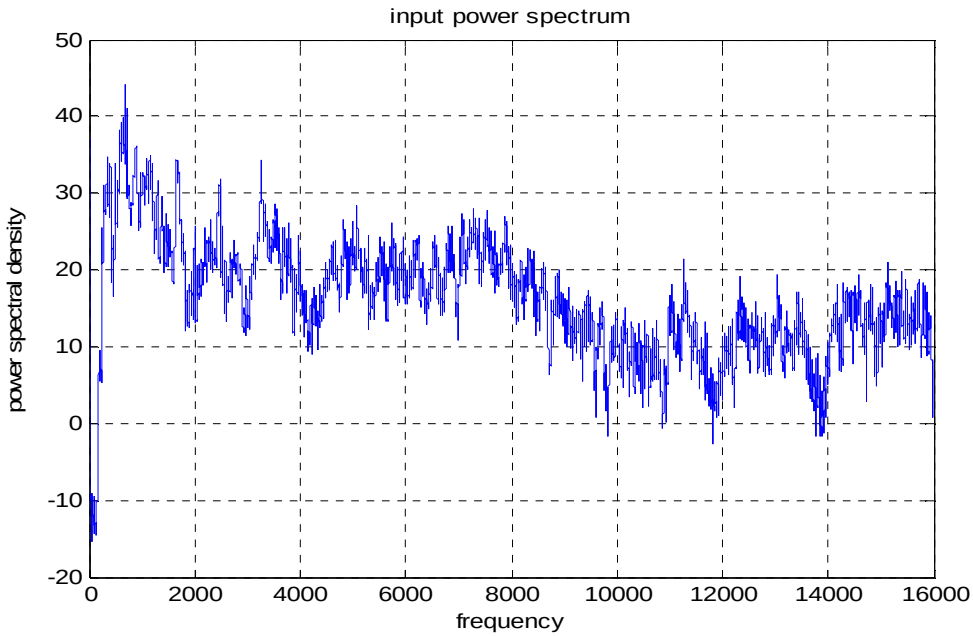


Fig.. 7. 11. PSD of the input speech signal obtained from the time domain beamformer with wind noise

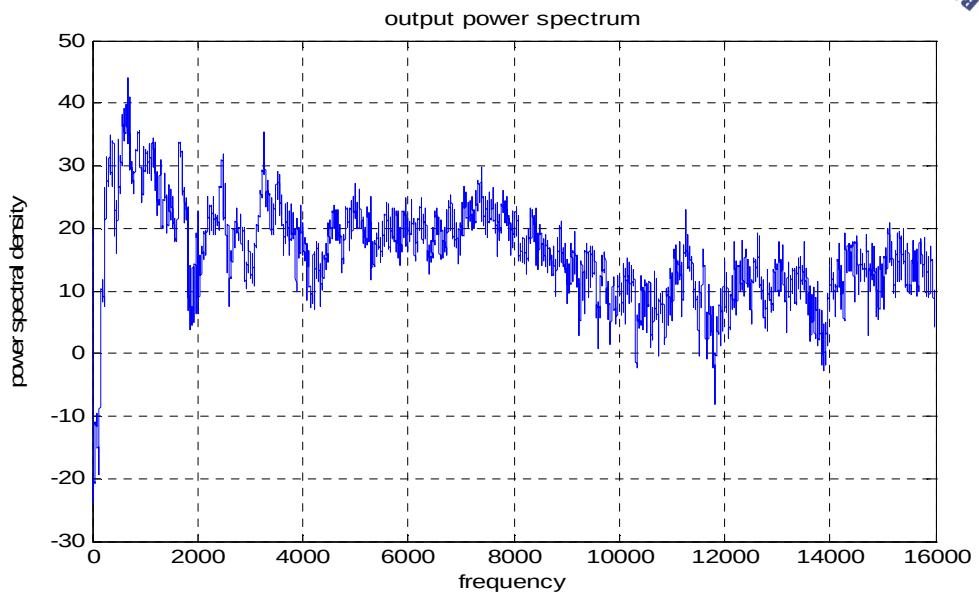


Figure. 7. 12. PSD of the output speech signal obtained from the time domain beamformer with wind noise.

The input and output power spectrums of the beamformer when additive white noise is given with speech signal are in Figure. 7. 13 and Figure. 7. 14.

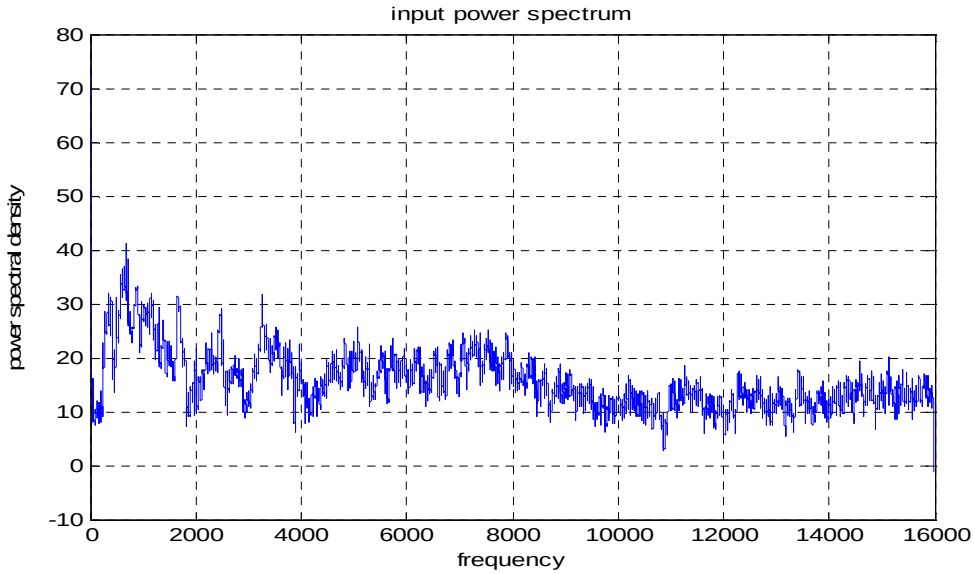


Fig. 7. 13. PSD of the input speech signal obtained from the time domain beamformer with additive white noise

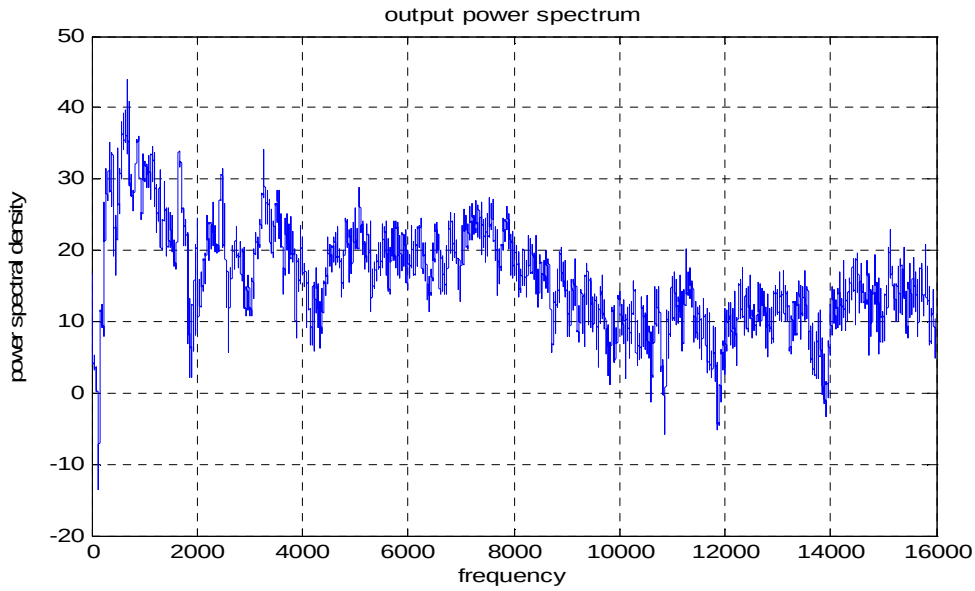


Figure. 7. 14. PSD of the output speech signal obtained from the time domain beamformer with additive white noise.

After the implementation of the time domain beamformer, frequency domain beamformer has then implemented by using WOLA filter bank. Before the implementation frequency domain beamformer WOLA filter bank has been designed and tested it by taking the sampling frequency (Fs) as 16000 KHz. The results of these filter banks are shown below.

The power spectral density of the output is Fig. 7. 15.

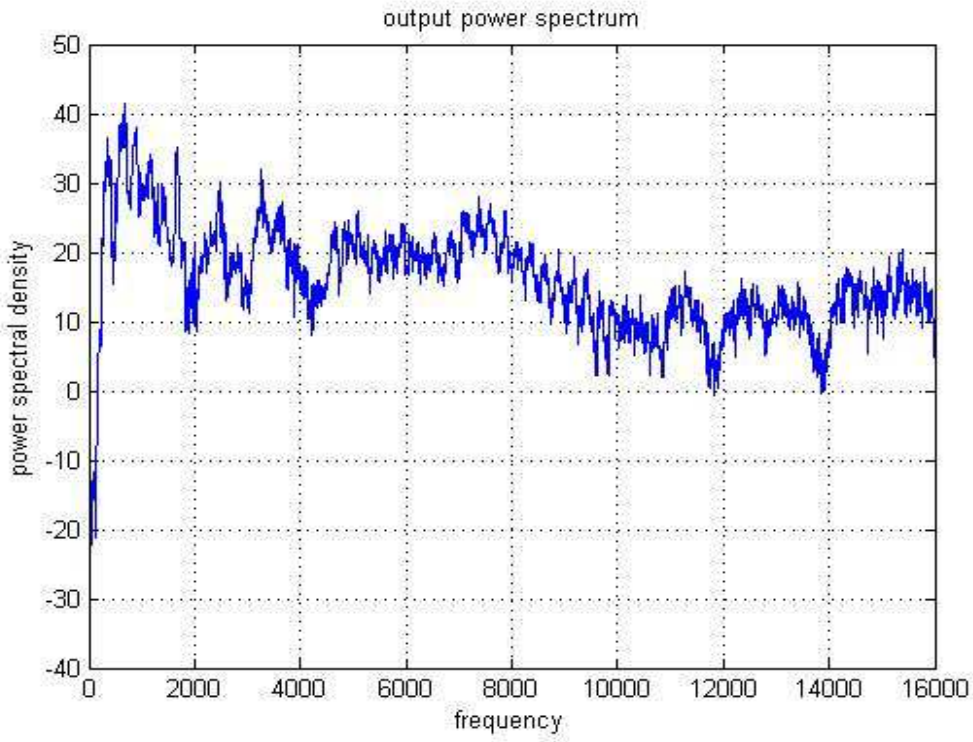


Figure. 7. 15. PSD of the output speech signal obtained after processing from the filter bank.

The magnitude response and the impulse response of the WOLA filter bank are represented in Figure. 7. 16 and Figure. 7. 17.

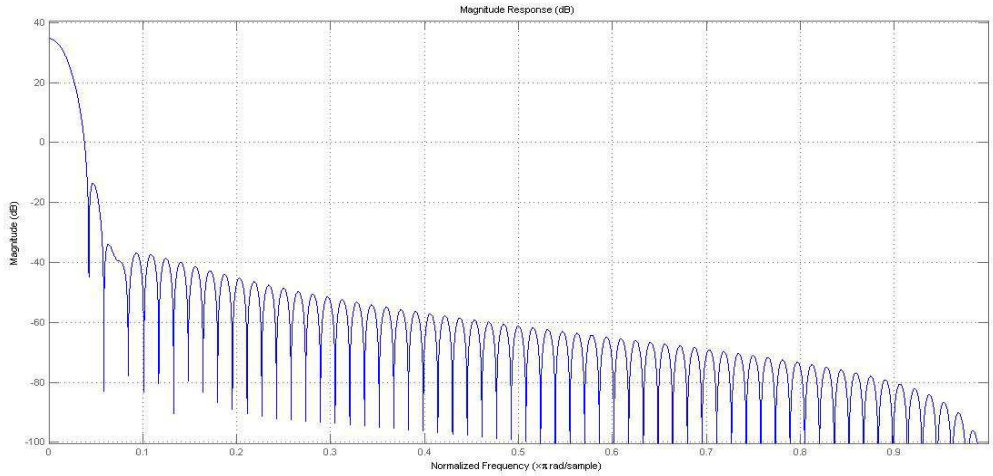


Fig. 7. 16. Magnitude response of the WOLA filter bank for 128 sub-bands.

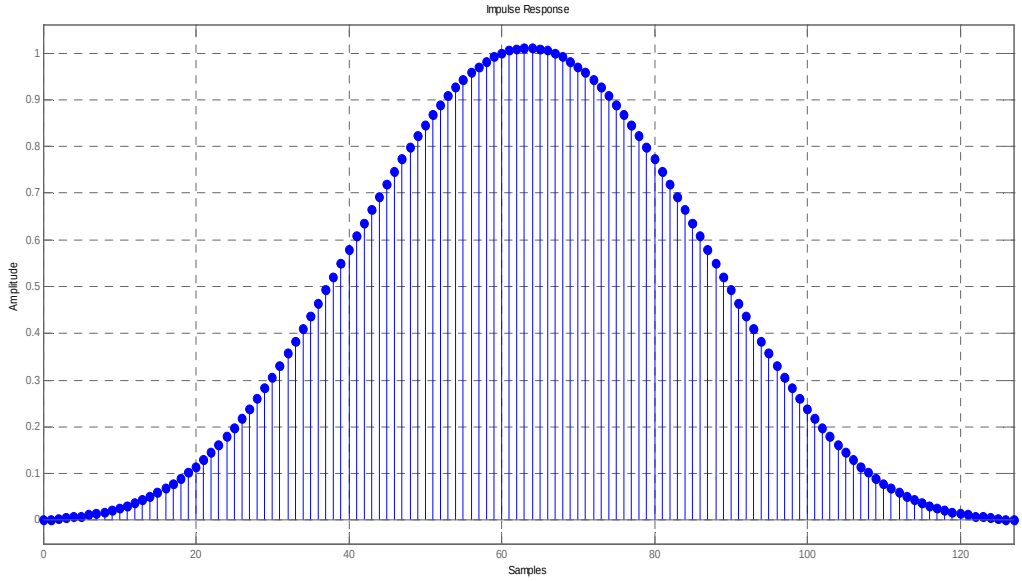


Figure. 7. 17. Impulse response of the WOLA filter bank for 128 sub-bands..

In the Filter and Sum Beamformer both the AWN and wind noise are tested and analysed by means of SNR improvement and the PESQ score.

An RLS algorithm is used for reducing the unwanted echoes and to improve the speech quality. Here in this RLS algorithm a parameter λ is used. λ is called the forgetting factor. This is a small positive constant very close to, but smaller than 1. With values of λ<1 more importance is given to the most recent error estimates and thus the more recent input samples, this results in a scheme that places more emphasis on recent samples of observed data and tends to forget the past. r_{dx} is the desired signal.

The RLS algorithm is implemented through the following steps [19].

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

$$\bar{y}_{n-1}(n) = \bar{W}^T(n-1)x(n) \qquad (46)$$

The intermediate gain vector is calculated using the equation

$$u(n) = \check{\psi}_{\lambda}^{-1}(n-1)x(n) \quad (47)$$

$$k(n) = \frac{1}{\lambda + x^T(n)u(n)}u(n) \quad (48)$$

The estimation error values is calculated using equation

$$\bar{e}_{n-1}(n) = d(n) - \bar{y}_{n-1}(n) \quad (49)$$

The filter tap weight vector is updated using the 49 and the gain vector are calculated using 47 and 48

$$w(n) = \bar{W}^T(n-1) + k(n)\bar{e}_{n-1}(n) \quad (50)$$

The inverse matrix is calculated using the equation

$$\check{\psi}_{\lambda}^{-1}(n) = \lambda^{-1}\check{\psi}_{\lambda}^{-1}(n-1) - + k(n) * \left[x^T(n)\check{\psi}_{\lambda}^{-1}(n-1)\right] \quad (51)$$

Each iteration of the RLS algorithm requires $4N^2$multiplication operations and $3N^2$ additions.

To begin with the beamformer, AWN is given and the λ is varied. SNR improvement and PESQ score are noted. The number of sub-bands used is 128 and the OS is 64. The room dimensions are 6x4x2.8. Mic is located at the point 4x2x1, source is located at the point 5.8x2x1.5 and the noise source position is 2x1x0.5. Here the sampling frequency is 8000Hz.

TABLE. 7. 7. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH VARYING LAMBDA AND THE ROOM DIMENSIONS 8X8X8.

λ value	No of Sub-bands	OS	SNR improvement	PESQ score before processing	PESQ score after processing
0.4	128	64	12.1904	1.884	2.984
0.5	128	64	14.2844	1.884	1.963
0.6	128	64	14.2677	1.884	1.134
0.7	128	64	16.3773	1.884	3.432
0.8	128	64	11.9247	1.884	1.106

As better output is obtained for λ value of 0.7 the other parameters are modified in order to improve the quality of the speech still further.

TABLE. 7. 8. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH CONSTANT LAMBDA AND VARYING OS AND NO OF SUB BANDS ROOM DIMENSIONS 8X8X8.

λ value	No of Sub-bands	OS	SNR improvement	PESQ score before processing	PESQ score after processing
0.7	64	32	8.5078	1.884	2.016
0.7	256	64	13.4519	1.884	3.628
0.7	256	32	14.9927	1.884	1.188
0.7	512	64	11.1921	1.884	2.786
0.7	256	128	14.3915	1.884	2.489

In the above manner, AWN is replaced by wind noise and the SNR improvement and PESQ score are noted down in order to find out at which values a desired signal will be obtained with more improved quality.

TABLE. 7. 9. THE SNR IMPROVEMENT AND PESQ SCORE FOR AWN WITH VARYING LAMBDA AND ROOM DIMENSIONS 8X8X8.

λ value	No of Sub-bands	OS	SNR improvement	PESQ score before processing	PESQ score after processing
0.5	256	64	6.2159	1.041	1.369
0.6	256	64	7.1684	1.041	1.729
0.7	256	64	11.8998	1.041	1.818
0.8	256	64	12.8465	1.041	1.405
0.9	256	64	14.1897	1.041	1.946

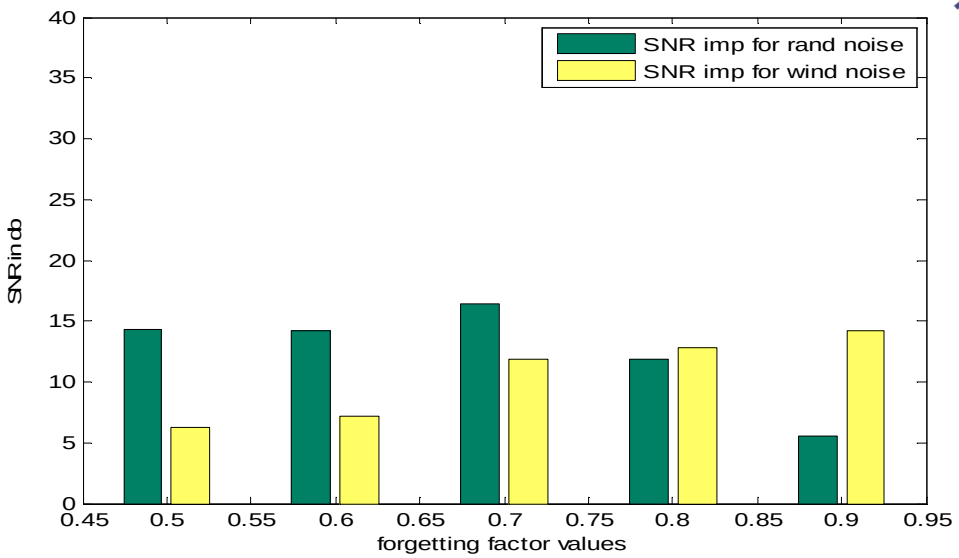


Figure. 7. 18. Representation of SNR improvement through blocks for both wind noise and random noise in frequency domain.

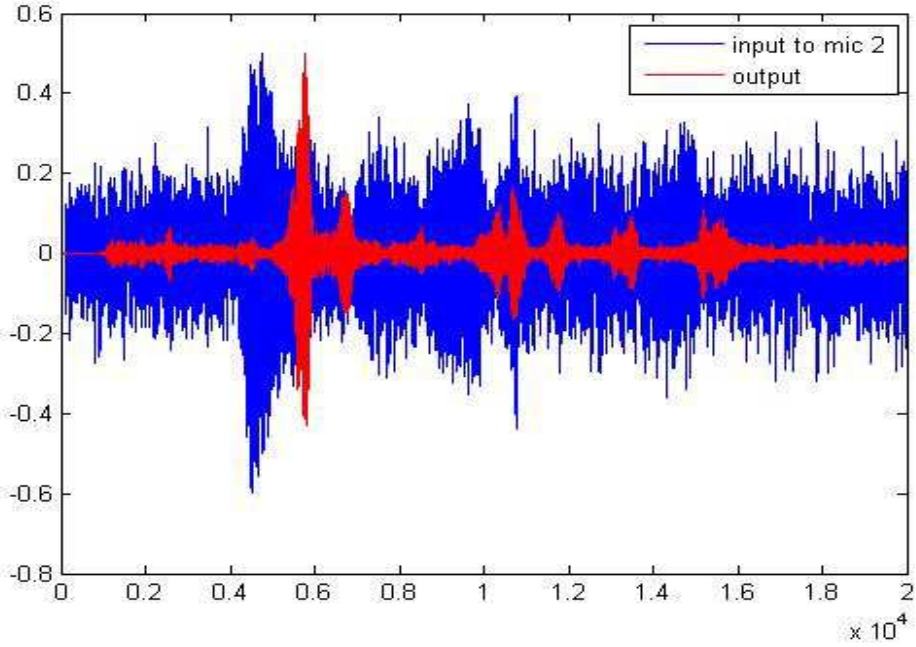


Figure. 7. 19. Input speech and the output speech for AWN noise for frequency domain beamformer.

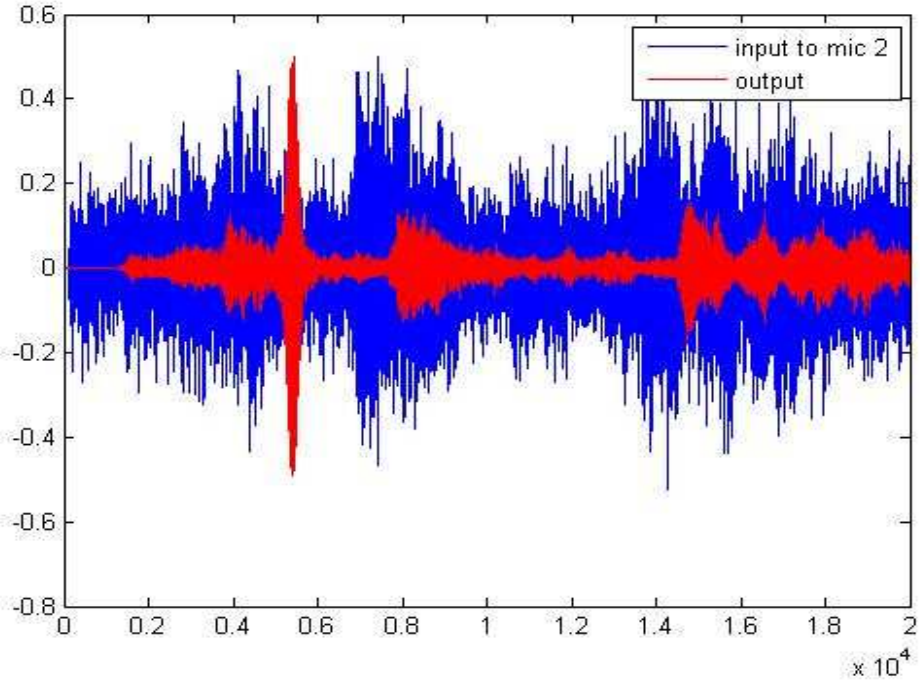


Figure. 7. 20. Input speech and the output speech for wind noise for frequency domain beamformer.

SRP-PHAT is a robust algorithm used for sound source localization this is shown by recent experimental studies. In the Filter and sum beamformer SRP-PHAT is used for speaker localization. The speaker position is identified using SRP-PHAT but unfortunately could not recover the speech signal properly. Initially SRP-PHAT is tested by taking two mics. Mics are arranged in linearly. τ value is calculated and the position at which τ has been obtained is determined. Here the power of the speech signal τ is. The signal is delayed by one sample after finding τ and τ is again calculated and the position is identified.

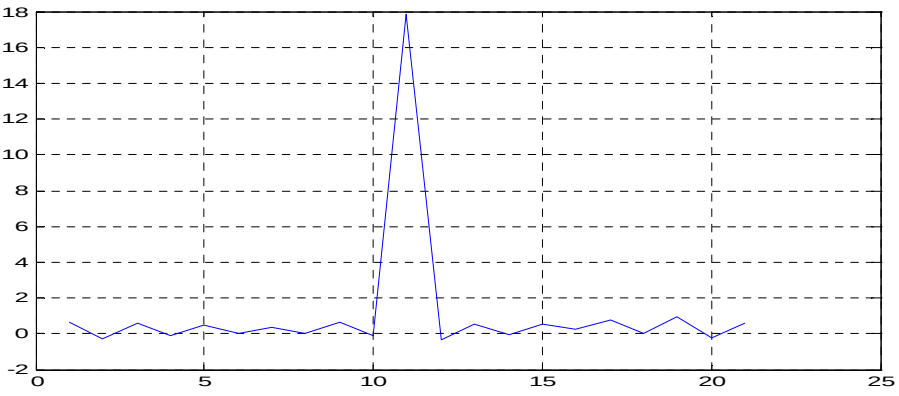
Before finding the SRP for two mics initially a random noise signal is taken and delayed by one sample every time respectively and the τ's position and value are calculated. The values obtained are shown in the table below.

TABLE. 7. 10. TABLE REPRESENTING THE POWER AND POSITION VALUES IN SRP-PHAT.

	Power(τ)	position
Initial input signal	17.8792	11
Signal delayed by one sample	17.5010	12
Signal delayed by two samples	17.1255	13
Signal delayed by three samples	16.4689	14
Signal delayed by four samples	16.6669	15
Signal delayed by five samples	15.0400	16

The graphical representations of the positions obtained are represented in Figure.7. 21 , Figure 7. 22 and Figure 7. 23.

Figure. 7. 21. Plot representing the position of the power for the given AWN



signal.

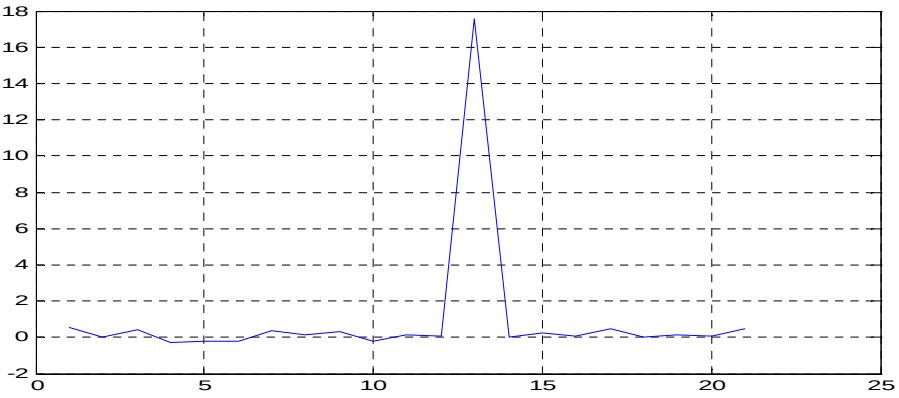


Figure. 7. 22. Plot representing the position of the power for the given AWN signal delayed with one sample.

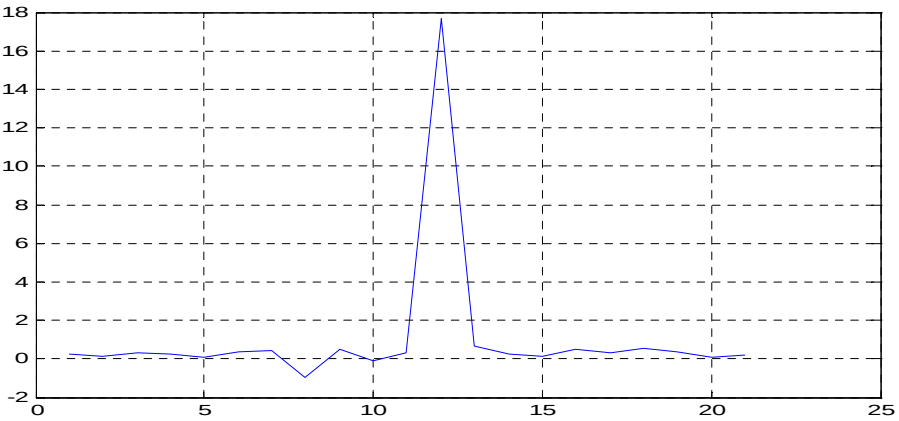


Figure. 7. 23. Plot representing the position of the power for the given AWN signal delayed with two samples.

After testing this, SRP-PHAT with two mics is implemented and the position where power is obtained is identified.

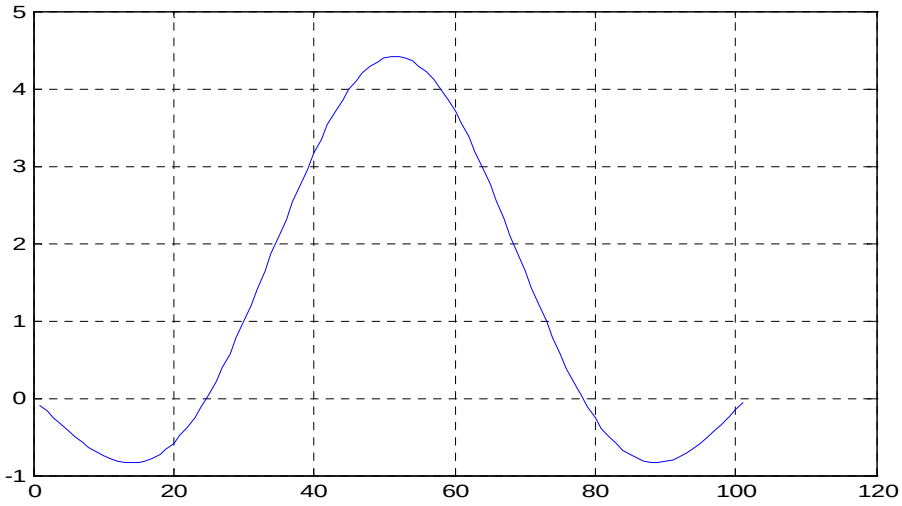


Figure. 7. 24. Plot representing the position where the speech is identified for 2 mics.

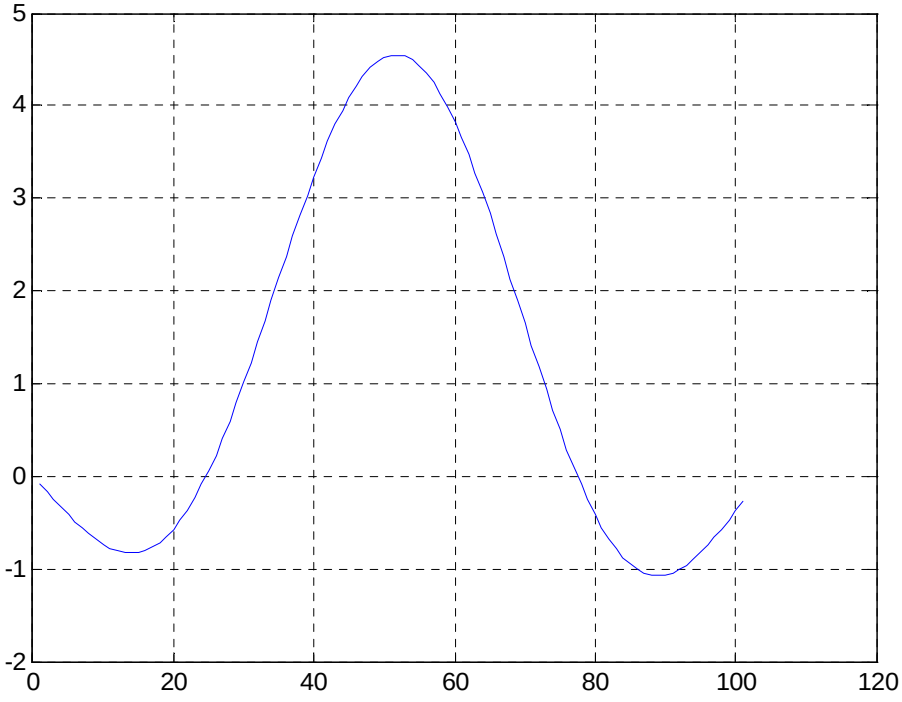


Figure. 7. 25. Plot representing the position where the speech is identified for 4 mics.

8. Conclusions

The use of Wiener filter Beamformer and SRP-PHAT algorithm, as presented in this research, shows substantial improvement in the SNR and segmental SNR for a range consistent with a multiple interfering speaker environment. In the multiple speaker environments, interfering talkers will have enough power to generate an acceptable beamformed signal estimate for the enhancement algorithms. As a result, the speech enhancement methods presented in this paper are able to contend with non-stationary, broadband noise that occurs in a multiple speaker environment.

The algorithm performed robustly in the low noisy and less reverberant environment. The modification made in SRP-PHAT algorithm makes it much faster by reduced computations. SRP-PHAT algorithm is also suitable for real time processing of speaker localization in a conference room.


BLEKINGE TEKNISKA HÖGSKOLA
BTH

9. Future Work

Humans naturally communicate using speech. It is therefore evident that the future of man-machine communications will focus on speech a main interface. Also, as device become smaller and more personal, a speech interface become essential. It is well know that present speech recognition algorithms are not very robust to hands-free speech. The significant loss in recognition is performing due to acoustic reverberation and SNR loss that are typical in hand-free applications. It is not clear, however, that this metric is optimal for speech recognition algorithm. The attempts to increase the accuracy of hand-free speech recognition optimize the beamformer design to minimize the perturbation of the measured feature vectors in reverberation and noise. However, it is not likely that this is equivalent to maximizing the input SNR.

Since microphone arrays have the ability to generate multiple outputs of spatially filtered input signals, it would seem apparent that speech recognition algorithm in the future will incorporate the idea of multiple acoustic input, even to the point that the speech recognizer is part of beamforming processing. One obvious application of this idea would be to have the beamformers give many outputs: one for the desired signal plus noise, and the others are the essentially representative of the spatial noise field. The recognizer would then have an estimate of the background noise field to utilize for robust speech end-point detection and to update the word models and statistics. Similar synergistic possibilities exist in the design of speech and audio coders for the operation in the noisy environments. Finally the use of a spatially segmented acoustic field by beamforming could increase the cancellation depth and bandwidth of the active noise cancellation hearing systems.

Reference

[1] M.Brandstein and D.Ward (Eds.),"Micro Phone Arrays".

[2] J. Allen and D. Berkley, “Image Method for Efficiently Simulating Small Room Acoustics," Journal of the Acoustical Society of America, vol. 65, no. 4, pp. 943-950, 1979.

[3] Stephen G. McGovern, "A Model for Room Acoustics", http://sgm-audio.com/research/rir/rir.html.

[4] "Room Impulse Response Generator", dr.ir. Emanuel A.P. Habets, September 20, 2010.

[5] Vesa Välimäki, "Simple Design of Fractional Delay All pass Filters”, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, P.O. Box 3000, FIN-02015 HUT, Espoo, Finland.

[6] Benny Sallberg, Ph.D, "Digital Signal Processors ET1304", Department of Electrical Engineering, Blekinge Institute of Technology, SE-371 75 Karlskrona, Sweden.

[7] Zohra Yermeche, "Subband Beamforming for Speech Enhancement in Hands-Free Communication", Blekinge Institute of Technology, Sweden, December 2004.

[6] Zohra Yermeche, "A Calibrated Constrained Subband Beamforming Algorithm for Speech Enhancement”, Blekinge Institute of Technology, Sweden, December 2004.

[8] Ramasamy Venkatasubramanian Master's Thesis"Beamforming for MC-CDMA" Virginia Polytechnic Institute and State University, Blacksburg, Virginia, January 31 2003.

[9] Alberto Abad Gareta, "A multi-microphone approach to speech processing in a smart-room environment", PhD Thesis, Universitat Polit`ecnica de Catalunya, Barcelona, February 2007.

[10] 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”, IEEE Transactions on Speech and Audio Processing, 4(13):593-606, May 2005.

[11]. C. H. Knapp and G. C. Carter, “The generalized cross correlation method for estimation of time delay,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[12] Yasir Masood Malik, "Speaker Localization, tracking and remote speech pickup in a conference room", Blekinge Institute of Technology, November 2009.

[13] Adaptive beamformer [Online]. Available: http://en.wikipedia.org/wiki/Adaptive_beamformer

[14] Beamforming [Online]. Available: http://en.wikipedia.org/wiki/Beamforming

[15] Microphone Array [Online]. Available: http://en.wikipedia.org/wiki/Microphone_array

[16] Amit Munjal, Vibha Aggarwal, Gurpal Singh, “RLS Algorithm for Acoustic Echo Cancellation,” *Nat. Conf. Challenges and Inform. Technology COIT,* Mar 2008.

[17] 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.html

[18] Z. Yermeche, “Soft-Constrained Subband Beamforming for Speech Enhancement,” Ph.D. dissertation, Dept. of Signal Processing, Blekinge Institute of Technology, Karlskrona, SW, 2007.