INTELLIGENT CAMERA TRACKING USING SRP-PHAT BASED SOUND SOURCE LOCALIZATION IN FREQUENCY DOMAIN

DAMODAR REDDY NATURI

Supervisor: Dr. Nedelko Grbic, Dr. Benny Sällberg & Dr. Ingvar Claesson
Examiner: Dr. Sven Johansson
Department of Signal Processing
School of Engineering (ING)
Blekinge Institute of Technology
This thesis is submitted to the School of Engineering at Blekinge Institute of Technology (BTH) in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering with emphasis on Signal Processing.

Contact Information:

Author:

Damodar Reddy Naturi

E-mail: dana10@student.bth.se, damodar.reddy77@gmail.com

Supervisors:

**Dr. Nedelko Grbic**

School of Engineering (ING)

E-mail: nedelko.grbic@bth.se

Phone: +46 455 38 57 27

**Dr. Benny Sällberg**

School of Engineering (ING)

E-mail: benny.sallberg@bth.se

Phone: +46 455 38 55 87

**Dr. Ingvar Claesson**

School of Engineering (ING)

E-mail: ingvar.claesson@bth.se

Phone: +46 455 38 57 14

Examiner:

**Dr. Sven Johansson**

School of Engineering (ING)

E-mail: sven.johansson@bth.se

Phone: +46 455 38 57 10

**School of Engineering**

Blekinge Institute of Technology

371 79, Karlskrona

Sweden.

Internet: www.bth.se/ing

Phone: +46 455 38 50 00

Fax: +46 455 38 50 57
ABSTRACT

The Steered Response Power Phase Transform (SRP-PHAT) is one of the most robust methods among sound source localization operating in noisy and reverberant environments. Direction of Arrival (DOA) Estimation has important applications in human computer interfaces such as video conferencing, speech enhancement and speech recognition.

In this thesis work, SRP-PHAT method has been implemented for 16 element microphone array arranged into 4 rows and 4 columns in the presence of noise and reverberation. Computation of TDOA for each pair of microphones in a row setup or a column setup, generalized cross correlation estimates are calculated and thereby computing the source position and then by averaging the row wise obtained TDOA values and column wise obtained TDOA values, best accurate source position can be determined.

Weighted Overlap and Add (WOLA) filter bank is used in SRP-PHAT method to find the TDOA in frequency domain. Original TDOA's and estimated TDOA's obtained from SRP-PHAT are compared to analyse the performance of the SRP-PHAT method. Mean estimation error and Standard deviation are calculated to find the accuracy of the estimated values of TDOA.
ACKNOWLEDGEMENTS

First and foremost deeply thankful to our Professor Dr. Nedelko Grbic, for his wonderful guidance during this thesis-work in field of speech processing, at Blekinge Institute of Technology. I am thankful for his continuous feedback and encouragement throughout this thesis work. His broad knowledge and hardworking attitude has left us with very deep impressions and they will greatly benefit us for our life.

I would like to thank all of our family members and for their affection, guidance and encouragement throughout the study period.

Finally, I would like to extend thanks to our close friends for their continuous support, discussions and suggestions in solving all issues during the thesis work.
TABLE OF CONTENTS

Abstract ................................................................................................................................. iii
Acknowledgements ............................................................................................................. iv
List of Figures: ....................................................................................................................... v

CHAPTER-1

1. Introduction: ..................................................................................................................... 1
   1.1 Motivation for Research: ............................................................................................ 1
      1.1.1 Time Delay Estimation (TDE) ........................................................................... 2
      1.1.2 Methods of Spectral Estimation ....................................................................... 3
      1.1.3 One Stage Steered Beam-formers ..................................................................... 4
   1.2 Research Question: ..................................................................................................... 6
   1.3 Thesis Organization: .................................................................................................. 6

CHAPTER-2

2. Background Theories ...................................................................................................... 7
   2.1 Microphone Array ........................................................................................................ 7
      2.1.1 Concept of Microphone Array ............................................................................ 7
         2.1.1.1 Representation of Microphone Array Input ................................................. 8
      2.1.2 Spatial Aliasing ..................................................................................................... 9
      2.1.3 Microphone Spacing Considerations .................................................................... 10
   2.2 Fractional Delay Filters ............................................................................................. 10
   2.3 Filter Bank System ..................................................................................................... 11
      2.3.1 Basic idea of Filter Bank ..................................................................................... 12
      2.3.2 Analysis and Synthesis Bank ............................................................................. 13
      2.3.3 Weighted-overlap Add Filter Bank .................................................................... 14
         2.3.3.1 Implementation of WOLA Filter Bank ......................................................... 14
   2.4 Room Modelling ......................................................................................................... 17
      2.4.1 Introduction .......................................................................................................... 17
      2.4.2 Reverberations ..................................................................................................... 17
      2.4.3 Image Source Model ............................................................................................ 19
         2.4.3.1 Set-Up ............................................................................................................ 20
2.4.3.2 Generation of discrete frequency domain room impulse response 21
2.4.3.2.1 Locating of virtual sources ........................................... 21
2.4.3.2.2 Unit impulse response and Propagation Attenuation .... 22
2.4.3.2.3 Thiran all pass filter ..................................................... 22
2.4.3.2.4 Propagation attenuation .............................................. 23
2.4.3.2.5 Reflection Coefficient ............................................... 23

2.5 Generalized Cross-Correlation (GCC) using the Phase Transform

(GCC-PHAT) .............................................................................. 24
2.5.1 Generalized Cross-Correlation (GCC) ............................... 24
  2.5.1.1 Derivation of the GCC .................................................. 25
2.5.2 The Phase Transform (PHAT) ............................................ 26
2.5.3 GCC-PHAT ........................................................................ 27

2.6 SRP-PHAT ........................................................................... 27

CHAPTER-3

3. Design and Implementation in MATLAB ........................................ 29
  3.1 Microphone Array Setup ............................................................ 29
  3.2 Fractional Delay Filter ............................................................. 30
  3.3 Filter Bank ........................................................................... 31
  3.4 Room Impulse Response ......................................................... 31
  3.5 SRP-PHAT Implementation ..................................................... 32

CHAPTER-4

4. Evaluation of SRP-PHAT Using 16 Mics in Different Environments .... 36
  4.1 Test data .............................................................................. 36
     4.1.1 Speech Signal ................................................................. 36
     4.1.2 Noise Signal ................................................................. 36
  4.2.1 ONLY SPEECH SIGNAL ................................................... 37
  4.2.2 SPEECH AND NOISE SIGNALS ....................................... 40
CHAPTER-5
5. Conclusion and future work: ................................................................. 53
References........................................................................................................ 53

LIST OF FIGURES

Figure 2.1.1.1.1 Linear Microphone Array Layout............................................. 8
Figure 2.2.1 Two Microphones with counter Delay elements............................ 11
Figure 2.3.1.1 Frequency Response of an Ideal Bandpass Filter......................... 12
Figure 2.3.1.2 Three Band Filter Bank, with adjacent band edges not overlapping... 13
Figure 2.3.2.1 Analysis and Synthesis Filter Bank – M Band............................ 13
Figure 2.3.2.2 Frequency Response of Prototype Analysis Filters...................... 14
Figure 2.3.3.1.1 WOLA Filter Bank: Block Diagram for Analysis Stage............. 15
Figure 2.3.3.1.2 WOLA Filter Bank: Block Diagram of Synthesis stage.......... 16
Figure 2.4.2.1 Schematic diagram of a typical speaker to receiver room impulse response three parts can be easily distinguished: direct wave, early reflections and late reflections..... 19
Figure 2.4.3.1.1 Mapping of different virtual sources with mirror method........... 20
Figure 2.4.3.1.2 Path involving two reflections with two virtual sources........... 20
Figure 2.5.1: TDOA Between two microphones.............................................. 24
Figure 3.2.1 The original signal and the delayed signals with respect to Mic1, Mic2, Mic3 and Mic4 ......................................................................................................................... 30
Figure 3.4.1 The RIR filter coefficients for $R=0.65$ & Source position $Src = [5.5,4]$ ....... 31
Figure 3.5.1 Overview of 4 element microphone setup (Row-Wise setup (Type 1), $R_1$) showing that speech source is moved at a distance from reference point of microphone array from -90’ to 90’ and noise source is fixed at a location at 45’................................. 33
Figure 3.5.2 Overview of 4 element microphone setup (Column-Wise setup (Type 2), $C_1$) showing that speech source is moved at a distance from reference point of microphone array from -90’ to 90’ and noise source is fixed at a location at 45’................................. 33
Figure 3.5.3: Implementation of SRP-PHAT for 4 Microphones Setup, Only Speech........ 34
Figure 3.5.4: Implementation of SRP-PHAT for 4 Microphones Setup, Speech and Noise... 35

Figure 4.1.1 The Test speech signal benny_radio.................................................... 36
Figure 4.1.2 The Test Noise signal brown_noise.................................................... 37

Figure 4.2.1.1 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-1 Column-wise................................................................. 38
Figure 4.2.1.2 shows the mean estimation error of the estimated DOA for 256 sub-bands for Setup-2 Column-wise................................................................. 38
Figure 4.2.1.3 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-2 Row-wise................................................................. 39
Figure 4.2.1.4 shows the mean estimation error of the estimated DOA for 256 sub-bands for Setup-2 Row-wise................................................................. 39

Figure 4.2.2.1 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-1 Column-wise....................... 41
Figure 4.2.2.2 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-1 Column-wise....................... 41
Figure 4.2.2.3 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-1 Column-wise....................... 42
Figure 4.2.2.4 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-1 Column-wise........... 42
Figure 4.2.2.5 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-1 Column-wise....... 43
Figure 4.2.2.6 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-1 Column-wise....... 43
Figure 4.2.2.7 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-1 Column-wise....... 44
Figure 4.2.2.8 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-1 Column-wise…… 44
Figure 4.2.2.9 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-1 Column-wise…… 45
Figure 4.2.2.10 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-1 Column-wise…… 45
Figure 4.2.2.11 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-1 Column-wise…… 46
Figure 4.2.2.12 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-1 Column-wise…… 46
Figure 4.2.2.13 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-2 Row-wise………… 47
Figure 4.2.2.14 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-2 Row-wise………… 47
Figure 4.2.2.15 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-2 Row-wise………… 48
Figure 4.2.2.16 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-2 Row-wise………… 48
Figure 4.2.2.17 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-2 Row-wise………… 49
Figure 4.2.2.18 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-2 Row-wise………… 49
Figure 4.2.2.19 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-2 Row-wise………… 50
Figure 4.2.2.20 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-2 Row-wise........... 50
Figure 4.2.2.21 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-2 Row-wise........... 51
Figure 4.2.2.22 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-2 Row-wise........... 51
Figure 4.2.2.23 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-2 Row-wise........... 52
Figure 4.2.2.24 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-2 Row-wise........... 52

**LIST OF NOTATIONS**

<table>
<thead>
<tr>
<th>Acronyms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOA</td>
<td>Direction Of Arrival</td>
</tr>
<tr>
<td>TDOA</td>
<td>Time Difference Of Arrival</td>
</tr>
<tr>
<td>TDE</td>
<td>Time Delay Estimation</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>SCOT</td>
<td>Smooth Coherence Transform</td>
</tr>
<tr>
<td>FD</td>
<td>Fractional Delay</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transforms</td>
</tr>
<tr>
<td>WOLA</td>
<td>Weight Overlap Add</td>
</tr>
<tr>
<td>FIFO</td>
<td>First In First Out</td>
</tr>
<tr>
<td>RIR</td>
<td>Room Impulse Response</td>
</tr>
<tr>
<td>ISM</td>
<td>Image Source Model</td>
</tr>
<tr>
<td>Mics</td>
<td>Microphones</td>
</tr>
<tr>
<td>GCC</td>
<td>Generalized Cross Correlation</td>
</tr>
<tr>
<td>PHAT</td>
<td>Phase Transform</td>
</tr>
<tr>
<td>SRP</td>
<td>Steered Response Power</td>
</tr>
<tr>
<td>OS</td>
<td>Over sampling ratio</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
</tbody>
</table>
Chapter-1

INTRODUCTION

1.1 Motivation

Microphone array consists of number of microphones placed at different spatial locations which act as an Omni-directional acoustic antenna. There are several ways to design a microphone array, like linear arrays, spherical arrays, circular microphone arrays and super directive arrays. The spatial location of the source sound can be determined by the correlation of the signals received through different individual microphones. Microphone array majorly used to enhance or suppress the signal, noise reduction, optimal filtering, source separation and speaker tracking using different methods [19].

With the increased development in speech processing technologies, effective speech communication devices are getting more importance in recent research. This need is accomplished by microphone array, allowing user a hands free environment without wearing microphone or speaking exactly in front of microphone in a conference or while giving lecture’s. Other applications of microphone array are video conferencing, speech recognition and other human machine interfaces.

Although there are some disadvantages in array processing, mainly processing time and costly hardware due to large number of microphones, but still microphone array is being preferred due to its high and accurate performance in recent research.

There are different algorithms used for the speaker localization, which are generally divided into three categories, one stage procedure, two stage procedure and spectral estimation. Performance of each algorithm varies in different environmental conditions.
One stage procedure basically exploits the magnitude of the input acoustic speech signal, by steering the array data to different locations and finding the peak in the output power using beamforming technique. This peak will indicate the location of the sound source, and the technique is known as Steered Response Power (SRP) [21].

The two stage procedure consists of two steps. First step is to estimates the Time Difference of Arrival (TDOA) of all the microphone pairs using generalized cross correlation function. And in second step the estimated TDOA is used to generate the hyperbolic curves, and the speaker location is then estimated using the intersection of these curves.

In third procedure, localization is done by high resolution spectral estimation of input signal correlation matrix. The estimation in this case is done by assuming the noise and source signal statistically stationary. For speaker localization, sensitivity of this technique is reduced due to a lot of assumptions made for an ideal environment.

Due to the sensitivity of the spectral estimation for the speaker localization in a reverberant environment, recent research is more focused on the Steered Response Power (SRP) and on Time Delay Estimation (TDE).

The brief descriptions of all three procedures are explained below, describing the further research done to improve and enhance the performance to localize the speaker using microphone array.

1.1.1 Time Delay Estimation (TDE)

Time Delay Estimation (TDE) is used for the speaker localization by estimating the TDOA between two spatially separated microphones in a pair. TDE is most affected by two major problems while localizing the speaker, background noise and room reverberation. A good TDE can be done by taking long input data packets, which reduces the background noise affect. For moving source localization, TDE needs to be done with high data rate which could be achieved by taking short input data segments.
The most common TDE based method is generalized cross correlation (GCC) function [4]. However, GCC is not efficient in the noisy and reverberant environment for speaker localization. The weighting functions have been introduced in GCC for the improved and accurate performance. Various weighting functions have been developed such as maximum likelihood (ML), PHAT and smooth coherence transform (SCOT). GCC-ML weighting functions was developed and applied in a low reverberant environment for speaker localization in. But its performance starts degrading when multipath affect increases in a reverberant environment. For reverberant environment, PHAT weighting functions shows better performance in a realistic environment.

Another short time TDE method involves the minimization of weighted least square phase data for speaker localization. In outperforms both GCC-ML and GCC-PHAT in reverberant environment. But there is a trade off in this method as it takes extensive time for a complex searching algorithm. But a little improvement in the performance cannot be accepted on the cost of complex computations.

There are many applications in a speech processing which need high update rate, which eventually limits the size of input data segment. For automatic camera steering video conference may need an update rate of 200-300 ms for reliable speaker localization. However, in the case of multiple speaker localization using TDOA based methods, a high update rate which may consist of 20-30 ms data segment is required.

The TDOA based methods are more likely used for single speaker localization under low noisy conditions. When these TDOA based methods are applied for tracking of multiple talkers in a conference room using short data segment, it results in a poor TDE.

1.1.2 Methods of Spectral Estimation

Speaker localization using high resolution spectral estimation is also gaining a considerable attention in recent research. A number of spectral estimation beamformers have developed, such
as autoregressive modeling (AR) modeling, minimum variance (MV) spectral estimation, and multiple signal classification (MUSIC) algorithm. Basically these methods were developed for the spectral estimation of the frequencies and there characteristics. But for narrow band signal’s, spectral estimation is also used for estimating DOA using microphone array.

The localization in this case is done by assuming a statistically stationary noise and speaker source. A signal correlation matrix is derived from the input signal of each microphone. Eigen value decomposition is used to separate the signal correlation matrix into noise and signal subspace. Usually a sharp peak indicates the point of DOA in the input data. This sharp peak is estimated by matching either the noise subspace or signal subspace to all possible values of DOA.

The search method to estimate the accurate DOA needs complex computations and situation become worse, when there is a complex source model. The other disadvantage of spectral estimation is the assumptions in which it performs better such as, ideal source radiators, exact knowledge of microphone spacing and stationary channel characteristics. These assumptions and complexity of searching algorithm makes it less robust as compare to other beamforming techniques i.e. TDE and SRP.

1.1.3 One Stage Steered Beam-formers

A sound source can be localize by steering all possible locations through microphone array in a reverberant environment. The method for steering the sound source is known as beamforming. The beamformers can be divided into two categories, fixed beamformers and adaptive beamformers.

Adaptive beamformers steer different locations using microphone array to locate the active sound source. Mathematical Modeling and implementation of adaptive beamformers are more complex than fixed beamformers but it does not put more constraints on the arrangement and the spacing of microphone array, and it is adjustable to any kind of environmental conditions. Omni directional microphones are more suitable for the accurate localization of the speaker.
In contrast, fixed beamformers steers the desired direction of sound and does not change with the varying sound source, but attenuates the sound from all other direction. Generally microphones used for fixed beamforming are inherently directive and closely spaced. The performance of fixed beamformers degrades when the sound source is non stationary. This problem can be resolve by increasing the beamwidth. But by increasing the beamwidth, noise level will increase as well.

The steered response will be the output data extracted by simple beamformer, known as delay-and-sum beamformer. The sound source when originates will reached to all microphones with different time intervals. The delay-and-sum beamformer will align all received signals at individual microphones by giving them appropriate delay, and form a single steered response signal. The peak of that steered response will point the location of the sound source and the peak will be the Steered Response Power (SRP) [18].

SRP perform well in low reverberant environments, but once the reverberation increases from the minimal level, its performance starts degrading dramatically. The Phase Transform (PHAT) is receiving much more attention in the recent research for speaker localization, and is more robust in reverberant and low noise environments. Combining both, SRP-PHAT can be a robust speaker localizer in a reverberant environment.

SRP-PHAT has also quite remarkable performance for multiple speaker localization. Here the steered response power will have a number of peaks, pointing towards the direction of multiple speakers. But this method has a limitation for extensive computation to find a lot of local maxima in SRP. There are a lot of efficient search methods which can be used to find the optimum peak values in SRP, in the case of multi speaker localization [19].

Therefore, in this thesis work, only SRP-PHAT model is implemented using 16 microphone array is performed in sub-band domain in the presence of noise and reverberation effect and the performance of the system is analyzed by Mean estimation error and Standard deviation.
1.2 Research Question

Effective speech communication using conference telephones requires advanced speech acquisition methods combined with speaker localization and tracking. 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.

1.3 Outline of Thesis

This thesis is organized into 5 chapters, background theories including microphone array, fractional delay, room impulse response, filter bank, SRP-PHAT have been explained in detail in chapter 2. In chapter 3, implementation of the SRP-PHAT have been discussed and in chapter 4, the simulation results have been presented for all the possible conditions and then thesis work is concluded with future work and conclusion part in chapter 5.
CHAPTER-2

Background and Related Works

2.1 Microphone Array

Array Signal processing involves the use of multiple sensors in order to receive a signal carried by propagating waves. Sensor arrays are commonly considered as spatially sampled versions of continuous sensors often referred as apertures. Arrays consisting of sensors have broad range of applications in various fields like RADAR, SONAR, radio astronomy and seismology [23]. In this section we mainly focus on the use of microphone arrays to receive acoustic signals or more specifically, speech signals. Out of many types of microphone arrays, in this thesis the scope is limited to only linear microphone arrays.

2.1.1 Concept of Microphone Array

A microphone array consist of a set of microphones placed at different spatial locations in order to spatially sample the acoustic signal carried by propagating wave. Based up on principles of sound propagation, the multiple inputs can be processed to enhance or attenuate signals arriving from particular direction. In this way microphone arrays provide the enhancement of desired signal in the presence of corrupting noise sources. In frequency terms microphone arrays allows us to separate signals that have overlapping frequency content but are originated from different spatial locations. Enhancement by microphone array is purely based on the source location.

Using appropriate algorithms, each and every microphone signal in an array is processed and combined resulting in beamforming or spatial filtering. Which creates the microphones array polar response patterns having the ability of pointing to or away from particular source direction, thus speech from certain direction can be enhanced or suppressed, in similar way the correlation data of signals in microphone arrays allow us to determine the direction of arrival (DOA) of desired speech.
These microphone array techniques have great potential in practical applications of speech processing due to their ability to provide noise robustness. The linear array of microphones with source location is shown in figure 2.1.1.1

2.1.1.1 Representation of Microphone Array Input

The layout of a linear microphone array consisting of four microphones with four incident signals is shown in figure 2.1.1.1. We will assume that incident waves are planar. Each of the microphones in the array will receive each signal denoted by $x_k$. As the signals will arrive at different times i.e. these will be time delayed or time advanced versions of each other. Here origin is considered at the microphone $M_1$. Let us consider the spacing between individual microphones is $d$. The distance from microphone $M_1$ to any microphone $M_k$ is $k$ times of $d$. The frequencies of signals that a microphone array detects are important in determining constraints on the spacing of the microphones. According to trigonometry the extra distance traveled by incident signal from one microphone to other microphone is given as $d_k \sin \theta$ and the time delay corresponding to each of microphones $M_k$ are thus given in equation 2.1.2.2.

![Figure 2.1.1.1 Linear Microphone Array Layout](image_url)
\[ \tau_k = \frac{d_k \cos \theta_k}{v} \quad (2.1.1.1) \]

Where \( v \) is the velocity of sound (\( v = 343 \text{m/s} \)) and the input signal \( S_i \) at each microphone due to signal \( x_k \) can be represented as a phase shift of incoming signal by

\[ x_k(t) = s_i(t) e^{-j \omega_0 \tau_k} \quad (2.1.1.2) \]

and equivalently substituting for \( \tau_k \) and \( \omega_0 \) and generalizing the total received signal at the \( k \) th sensor is given as

\[ x_k(t) = \sum_{i=1}^{l} s_i(t) e^{-j 2\pi d_k \cos \theta_k / \lambda} \quad (2.1.1.2) \]

this form of representing the signal makes processing computationally easier.

### 2.1.2 Spatial Aliasing

The principle in temporal sampling is that the minimum sampling frequency required to avoid aliasing in the sampled signal is the Nyquist frequency. In the similar way the microphone arrays implement spatial sampling and an analogous requirement exists to avoid aliasing that is grating lobes in the directivity pattern [21].

The sampling theorem states that minimum sampling frequency \( f_s \) should be at least twice of the maximum frequency \( f_{\text{max}} \) in signal in order to avoid aliasing i.e.

\[ f_s = \frac{1}{T_s} \geq 2 f_{\text{max}} \quad (2.1.2.1) \]

Where, \( f_{\text{max}} \) is the maximum frequency component in the signals frequency spectrum. Similarly in order to avoid spatial aliasing the relation between the spatial sampling frequency \( f_{xs} \) (samples/metre) and highest spatial frequency \( f_{x\text{max}} \) component in the angular spectrum of signal given as

\[ f_{xs} = \frac{1}{d} \geq 2 f_{x\text{max}} \quad (2.1.2.2) \]

So, in order to avoid spatial aliasing the spatial sampling frequency should be at least twice of the maximum spatial frequency component in angular spectrum.
2.1.3 Microphone Spacing Considerations

Spacing between the microphones in an array is the most important one to consider. The distance between the sensors to avoid the spatial aliasing is given as

\[ d \leq \frac{\lambda_{\text{min}}}{2} \]  \hspace{1cm} (2.1.3.1)

An oversampling occurs when \( d < \frac{\lambda}{2} \), which allows the wavelength to be sampled more than two times and spatial under sampling occurs when \( d > \frac{\lambda}{2} \), which can result in aliasing higher frequency signals down in to the frequency band. In order to avoid aliasing a very effective microphone arrays can be built with microphone spacing’s much smaller than the minimum needed for the nyquist rate [21].

2.2 Fractional Delay Filters

Generally, the distance from the source to each microphone in an array is different; this difference is responsible for the phase shifted replicas of the signals received by each microphone in an array [13]. In order to obtain the constructive interference, the individual signals from each microphone should be in-phase with each other. So that summing all the signals together forms constructive interference which amplifies the output speech by the number of microphones in the array. To make the signals in-phase there’s a need to add an appropriate time delay to the recorded signal from microphone that is equal and opposite to the delay caused by the extra distance traveled by signal. This delay can be known with the help of geometry. These time delays can be easily handled if it’s an integer delay and would be difficult for fractional delays, hence Fractional Delay (FD) filters are used [14].

Assuming a uniform sampling, a fractional delay filter implements a delay that is a non-integer multiple of the sample period i.e. it implements the required band limited interpolation. Band limited interpolation is a technique for evaluating a signal sample at arbitrary point, even if it is located between two sampling points. Delaying sampled data by a fraction of a sample period is critical part of beamforming [10]. The figure below shows the two microphones with a plane wave approaching at angle \( \theta \). The wave will reach microphone \( M_1 \) before reaching \( M_2 \), so
to make both signals in-phase a delay is introduced at $M_1$. This can be shown in figure 2.2.1. Here $N$ may be integer or fractional number.

![Figure 2.2.1 Two Microphones with counter Delay elements](image)

To calculate the delay let us assume that distance between microphones is $d$ meters and arrival angle of plane wave is $\theta$, wave front should travel extra distance of $dcos\theta$ to reach $M_2$ after reaching $M_1$. The delay in time is given as $dcos\theta/v$ ms assuming that speed of sound is $v$. So we have to add time delay of $dcos\theta/v$ to signal received by $M_1$ and this time delay can be expressed in terms of samples is given as $(dcos\theta/v)f_s$.

### 2.3 Filter Bank System

The interest in digital filter banks has grown dramatically over last decade. Filter banks find wide variety of applications in communications, image compression, speech processing, antenna systems and digital audio industry. Filter bank allows a signal to be decomposed into several subbands, which facilitates more efficient and effective processing.

The filter banks and subband processing have been a huge success with broad range of literature covering the theory and applications. The basic blocks of filter bank are decimators and interpolators and these blocks were also efficiently applied in multirate systems. Since then a number of new developments have been taken place simultaneously in design of filter banks and multirate systems, these developments was reported in. In this thesis, the implementation of filter bank is based on the Short Time Fourier Transforms ($STFT$) regarding this an early theoretical framework for analysis and reconstruction using $STFT$ was formulated in. Since then
time-frequency systems based on decimated filter banks can be viewed as the generalization of the STFT representation.

2.3.1 Basic idea of Filter Bank

Let an input speech signal be \( x(n) \). This signal might have been captured either from a microphone or any other signal source. The Spectrum of a signal \( x(n) \) gives the distribution of signal energy as a function of frequency [12].

The property of an ideal bandpass filter is to reject all input signal energy outside the desired range of frequency, while allows the input signal energy within desired range. The range of allowed frequencies is often referred to as pass band. The edges of the band i.e. frequency boundaries are known as lower \( (f_l) \) and upper \( (f_h) \) cutoff frequencies. The difference between the upper and lower cutoff frequencies is known as the bandwidth \( (BW) \).

\[
BW = f_h - f_l
\]  

(2.3.1.1)

The midpoint of the frequency boundaries is known as center frequency \( (f_c) \) of the band pass filter and the ratio of the center frequency \( f_c \) to the bandwidth of the filter is known as quality factor \( Q \).

\[
Q = \frac{f_c}{BW}
\]  

(2.3.1.2)

The frequency response of the ideal bandpass filter is shown in figure 2.3.1.1

A filter bank is a system that decompose the input signal \( x(n) \) into a set of analysis signals \( x_1(n), x_2(n), x_3(n) \ldots x_k(n) \) each of which corresponds to a different range in the spectrum of \( x(n) \). The regions in the spectrum given by the analysis signals generally lies between audible
range i.e. approximately 20 Hz to 20 kHz and these regions do not overlap on each other, but are lined up one after the other as shown in figure 2.3.1.2. The analysis signals \( x_1(n), x_2(n), x_3(n) \ldots \) may be obtained using a collection of bandpass filters with bandwidths \( BW1, BW2, BW3 \), and center frequencies \( f_{c1}, f_{c2}, \ldots \) (respectively).

![Figure 2.3.1.2 Three Band Filter Bank, with adjacent band edges not overlapping](image)

Decimator and interpolators are basic blocks of filter bank system. Decimator tends to decrease the sampling rate of a signal whereas Interpolator tends to increase the sampling rate of the signal with chosen factor. Decimators and Interpolators are interpreted in.

### 2.3.2 Analysis and Synthesis Bank

A broad class of linear filter banks which employs analysis and synthesis banks is demonstrated by the block diagram in figure 2.3.2.1.

![Figure 2.3.2.1 Analysis and Synthesis Filter Bank – M Band](image)
An analysis bank is a set of analysis filters $H_k(Z) \ (0 \leq k \leq M - 1)$ which is a linear time varying system that decomposes a signal into $M$ subband signals $v_k(n) \ (0 \leq k \leq M - 1)$ each one of which is decimated by a factor $R$, when $R=M$, the system is said to be critically sampled or maximally decimated. Generally these maximally decimated systems have property of information preserving because of which it is considered as a best choice. The analysis bank splits the spectrum of $x(n)$ into consecutive subbands. The frequency response of prototype analysis filters can be shown in figure 2.3.2.2

![Frequency Response of Prototype Analysis Filters](image)

A synthesis bank is a set of filters $G_k(Z) \ (0 \leq k \leq M - 1)$, which combines $M$ signals $s_k(n) \ (0 \leq k \leq M - 1)$ into one signal $\hat{x}(n)$, typically called as reconstructed signal. For most common applications the filters in the analysis bank are band pass filters which are decimated at their normal nyquist rates.

2.3.3 Weighted-overlap Add Filter Bank

In this thesis a weighted overlap-add (WOLA) algorithm is implemented for subband processing applications and this WOLA filter is ideal for use in a wide range of applications which involves frequency domain processing like subband coding/decoding, echo cancellation, Voice activity detection and subband directional processing. The WOLA is well suited to applications that require low delay ($0<10ms$). WOLA has many special characteristics such as its flexibility, high performance, low cost and low complexity which altogether makes the speech enhancement algorithm more suitable to different applications [11].

2.3.3.1 Implementation of WOLA Filter Bank

The weighted Overlap add (WOLA) filter bank is an efficient filter bank technique frequently used in the signal processing application. It can be defined by four variables namely $L$ the length of analysis window, $D$ the decimation rate (block rate), $K$ the number of sub bands and
$D_F$ the synthesis window decimation rate along with the analysis window function $w(n)$. In the operation, the analysis stage accepts a block of $D$ new input data samples and the input $FIFO$ is shifted every time the new block of samples are arrived and $D$ new samples are stored in $u(n)$. The input $FIFO$ is then windowed by the analysis window function and stored in the temporary buffer $t_1(n)$. The windowing of input $FIFO$ is given by equation 2.3.3.1.1 and the length of temporary buffer is $L$.

$$t_1(n) = u(n) \cdot w(n)$$ \hspace{1cm} (2.3.3.1.1)

The resulting vector $t_1(n)$ is time folded in to another temporary vector $t_2(n)$ of length $K$ samples. The time folding is nothing but elements of $t_1(n)$ are modulo- $K$ added to $t_2(n)$ according to the equation 2.3.3.1.2.

$$t_1(n) = \sum_{m=0}^{L} t_1(n + mK)$$ \hspace{1cm} (2.3.3.1.2)

![Diagram](image)

Figure 2.3.3.1.1 WOLA Filter Bank: Block Diagram for Analysis Stage
In order to produce a zero phase signal for the FFT, the temporary buffer $t_2(n)$ is circularly shifted by $K/2$ a sample which means that the upper half of temporary buffer $t_2(n)$ is swapped place with the lower half of $t_1(n)$. The circularly shifted buffer $t_2(n)$ is fed to FFT of size $K$ to compute the subband signals $x_k(n)$. The synthesis stage as shown in figure 2.3.3.1.2, applies the $K$ size inverse FFT (IFFT) to the processed subband signals $v_k(n)$. To counter act the circular shift used in the analysis stage, the output of IFFT is circularly shifted $K/2$ samples and shifted data is stored in other buffer $t_4(n)$ of length $L/D_F$, where $L$ is the analysis window length $D_F$ is synthesis window decimation factor. The buffer $t_4(n)$ is weighted by a synthesis window function $z(n)$ of size $L/D_F$ and $z(n)$ is given in equation 2.3.3.1.3.

Figure 2.3.3.1.2 WOLA Filter Bank: Block Diagram of Synthesis stage
\[ z(n)=w[n D_f] \]  \hspace{1cm} (2.3.3.1.3)

The weighted data is summed with the data in the output FIFO, \( t_5(n) \) which has length of \( L/D_f \). The output FIFO data is overwritten with the summation result given as

\[ t_5(n) = t_5(n) + z(n) t_4(n) \]  \hspace{1cm} (2.3.3.1.4)

The output FIFO is then shifted left by \( D \) samples i.e. FIFO’s rear was filled with \( D \) zeros and out-shifted data is actual output data block \( y[nD] \).

### 2.4 Room Modeling

#### 2.4.1 Introduction

In order to achieve successful communication system it is compulsory that the emitter and receiver use same conditions. In other words the channel between the emitter and transmitter must be suitable for transmitting message.

In any communication system the channel is affected by several sources of distortion that can affect the success of the communication. Some of the distortion affects are acoustic noise, reverberations and interference speech of the other speaker instead of active speaker [25].

In far field microphone environments like meeting rooms where sensors are placed far apart from the source leads to the decay in the speech signal. Decay in the speech signal in far field environment is caused by two factors one is acoustic noise and the second one is reverberation.

#### 2.4.2 Reverberations

The acoustic signals are propagated in multi path in a closed room that means sound signal generated from a speaker will flow in direct path and multiple paths to the receiver (human or microphone). The microphone at the receiver will record the data of the direct path sound wave and multiple path waves those are generated by the reflection of the walls in the closed room. Theses multiple paths will depend on various constraints namely reflection coefficient,
absorption coefficient, surface of the room and size of the room. The multiple paths sound waves are different when compared to distant echo that is generated after 50 to 100 ms after initial sound wave. These multiple paths sound waves are commonly referred as reverberations [22].

In real world the existence of reverberation after the sound wave is removed is clearly observe in large closed rooms. This phenomenon is observed by reverberation time in a room. Reverberation time is time taken for sound in a room to decay by 60 dB. The reverberation time is regulated by two factors those are size of the room and surface of the room, there are so many other factors like objects in the room, temperature etc. which are also effect the reverberation but these are two most important. Surface of the room plays a crucial role because if the surface is reflective material then the number of reflections increase or else it is an absorption material then the number of reflections decreases. The Sabine’s equation illustrate the relation among the volume of the room \((V(m^3))\), the area of each surface \((S_i(m^2))\) and its absorption coefficient \(\alpha_i\) is shown in below equation (2.4.2.1)

\[
T_{60} = \frac{0.161V}{\sum_i S_i \alpha_i}
\] (2.4.2.1)

The nature of reverberation in a closed room is characterized by the speaker to receiver room impulse response. The below figure (2.4.2.1) display below indicates the room impulse response of a closed room where RIR is union of three things direct path, early reflection and late reflections. Early reflections are just some delayed and attenuated from direct sound wave. They are generated from the nearby surface and they are used for increasing the strong audibility of sound and special impression. Early reflections are essential in most of the applications. In the case of late reflections the rate of reflections increases and the interval time between the reflections decreases. These are dense pack of echoes and travel in all directions. In most of the cases late reflections causes spreading the speech spectra which leads to lack of audibility in the speech. In some of the application the presence of reverberation increases the quality of the hearing experience so the effect of reverberation depends on the applicable field. In the field of microphone array reverberation causes inaudibility and causes problem in identifying the speaker position.
Figure 2.4.2.1 Schematic diagram of a typical speaker to receiver room impulse response three parts can be easily distinguished: direct wave, early reflections and late reflections.

So in experimental set of speech processing, the reverberation makes an important role for their existence whether in the form of noise or usefulness. Experimentally reverberations of a room are generated from different methods like ray and beam tracing methods, image source method is used in this thesis for their generation.

2.4.3 Image Source Model

Image source method becomes a pre dominant tool in most of the research fields of speech processing [17]. Because of its simplicity and flexibility for achieving results it is widely implemented in acoustic and signal processing. ISM is implemented to generate impulse responses for a virtual source in a room. ISM’s one of the main important implementation is, performance assistant of various signal processing algorithms in reverberation environment. Few examples are ISM is implemented to validation the blind source separation, channel identification, equalization, acoustic source localization and many others. In such cases ISM is participate to test the specific algorithm in order to decide its endurance in different environmental reverberation.
2.4.3.1 Set-Up

Image source model approximate the room impulse response of multiple reflections from a single source in a room by transferring them into direct paths from multiple virtual sources. The virtual sources are the mirror images of the original source in the room. The figure 2.4.3.1.1 is a two dimensional structure of an original source and their mirror images of it.

![Figure 2.4.3.1.1 Mapping of different virtual sources with mirror method.](image)

In the above figure 2.4.3.1.1 Black star is microphone position, green circle is source and black circles are virtual sources. For simple illustrate the direct path and reverberated path here simple figure 2.4.3.1.2

![Figure 2.4.3.1.2 Path involving two reflections with two virtual sources](image)
In the above figure 2.4.3.1.2 path between the source and microphone without any deflection is direct path. Remaining two paths are reverberated paths as both are having deflections. Black path indicates the original sound waves and the blue path indicated the reverberation of the sound wave. The perceived path and the original path are same with respect to distance, absorption coefficient when the sound wave hits the wall the only difference is transferring multiple reflections of a single source into direct paths of multiple virtual sources [21].

2.4.3.2 Generation of discrete frequency domain room impulse response

The below equation 2.4.3.2.1 is the frequency domain representation of room impulse response

\[ H[k] = \sum_{i,j} \alpha_{i,j} \cdot \beta_{i,j} e^{i\omega k \tau_{i,j}} \]  

(2.4.3.2.1)

Variables in the equation 2.4.3.2.1 are described as follow; \( \alpha \) is the reflection coefficient, \( \beta \) is the propagation attenuation and \( \tau \) is the propagation delay. \( \alpha \) and \( \beta \) Have the capacity to alter the magnitude of the each impulse of the room impulses response.

2.4.3.2.1 Locating of virtual sources

Image source model is a well-established model for simulating RIR in a given room. Here Cartesian coordinates system \((x, y)\) are assumed for an enclosed room. \( i, j \) are the reflection indexes [1]. The locating of the virtual source is formulated with respect to one of the coordinates of the room is

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

(2.4.3.2.1.1)

Where \( x_i \) is the x coordinate of the virtual source and \( x_s \) is the x coordinate of the sound source and \( x_r \) is the length of the room in x dimension. Similarly it can be calculated for y coordinate of the virtual source. Reflection indices \( i, j \) are along the spatial axes of the room. As a part these indices can be positive or negative, if the indices \( i, j \) equals to \([0, 0]\) then the virtual
source is original sound source. Virtual source with reflection indices \( i, j \) is shown is equation 2.4.3.2.1.2

\[
V_{i,j} = (x, y)_{ij} \quad \tag{2.4.3.2.1.2}
\]

### 2.4.3.2.2 Unit impulse response and Propagation Attenuation

The direct path propagation vector between the virtual source and microphone is expressed in equation 2.4.3.2.2.1

\[
d_{i,j} = (x_m, y_m) - V_{i,j} \quad \tag{2.4.3.2.2.1}
\]

\((x_m, y_m)\) is the microphone coordinates and \( v_{i,j} \) is virtual source location. Before generating unit impulse function, it is important to find out the time delay for each echo let commence the time delay with respect to a function for generating unit impulse response. This time delay also known propagation delay is defined as,

\[
T_{i,j} = \| d_{i,j} \| \ast F_s/c \quad \tag{2.4.3.2.2.2}
\]

In the above equation 2.4.3.2.2.2, \( F_s \) is sampling frequency. This is the time taken in smaples, to propagate the length of the virtual direct path propagation vector.

### 2.4.3.2.3 Thiran all pass filter

In the simulation environments generation of fraction values is not as easy as in the case of theoretical expressions. In generating impulse response for each echo, time delay plays an important role [13]. This time delay may be round off value or fraction value. For obtaining fractional time delays, Thiran all pass filter is implemented. Thiran is popular because of it flat magnitude response and more concentration on the phase response. Transfer function for digital IIR all pass filter is formulated in below equation 2.4.3.2.3.1

\[
H(Z) = \frac{Z^{-N} D(z^{-1})}{D(z)} \quad \tag{2.4.3.2.3.1}
\]

Where \( N \) is the order of the filter and \( D(z) = 1 + a_1z^{-1} + a_2z^{-2} + \cdots + a_Nz^{-N} \) is the denominator polynomials with real value coefficient \( a_k \) and the numerator polynomials are the reverse version of the denominator ones. Thiran formula for all pass filters is
\[ a_k = (-1)^k \binom{N}{k} \prod_{n=0}^{N} \frac{d+n}{d+k+n} \quad (2.4.3.2.3.2) \]

\( d \) is the delay parameter, \( k=1, 2, 3, 4 \ldots N \), \( a_k \) will generate the coefficients for the IIR all pass filter in above equation 2.4.3.2.3.2

### 2.4.3.2.4 Propagation attenuation

Propagation is one of the factors that reduce the magnitude of the echoes. Not all the echoes have same magnitude; their magnitude will depend on a factor called as propagation attenuation.

Propagation attenuation is expressed in below equation 2.4.3.2.4.1

\[ \beta_{i,j} = \alpha^{\frac{1}{\|d_{i,j}\|}} \quad (2.4.3.2.4.1) \]

### 2.4.3.2.5 Reflection Coefficient

Reflection coefficient is another factor which will directly affect the magnitude of the echoes. Sound wave experiences partial reflection and partial transmission when it hit to walls of a room. Reflection coefficient is the amount of sound wave reflects when it hits a surface (walls, objects in the room). In the thesis concern coefficient is not the ratio between reflection and transmission sound wave, here coefficient means amount of sound wave reflection after sound wave absorb by the surface. The most important factor in the Image source model is to determine the total number of the reflections in the room. Indexing schema \(|i|, |j|\) are important for find out the total number of reflections. Let’s take wall reflection coefficient as \( \alpha \) and raise exponent \( n \) where \( n=|i|+|j| \) will give the total reflection of the sound wave is given in below equation 2.4.3.2.5.1

\[ \alpha_{i,j} = \alpha^{|i|+|j|} \quad (2.4.3.2.5.1) \]

Here \( \alpha < 1 \), reflection coefficient never greater than 1. If each wall has different reflection coefficient then it will lead to some more complex equations. If \( \alpha_{x=0} \) is the reflection coefficient of the wall perpendicular to the x-axis near the origin and \( \alpha_{x=xr} \) is reflection of the wall opposite
that, then the combined reflection coefficient by the virtual source is in below equation

\[ \alpha_{xi} = \alpha_{x=0} \left| \frac{1}{2} i - \frac{1}{4} + \frac{1}{4} (-1)^i \right| \alpha_{x=rr} \left| \frac{1}{2} i + \frac{1}{4} - \frac{1}{4} (-1)^i \right| \quad (2.4.3.2.5.2) \]

Similarly we can calculate the \( \alpha_{yj} \) and with the index \( j \) in order to generate the total reflection coefficient in below equation 2.4.3.2.5.3

\[ \alpha_{i,j} = \alpha_{xi}, \alpha_{yj} \quad (2.4.3.2.5.3) \]

The propagation attenuation and the reflection coefficient are directly proportional to the total number of reflections. As the total number of reflections increases then the both attenuation factors increases but this should be limited with the index schema to below reference order in equation 2.4.3.2.5.4

\[ |i| + |j| < N_{ref} \quad (2.4.3.2.5.4) \]

2.5 Generalized Cross-Correlation (GCC) using the Phase Transform (GCC-PHAT)

2.5.1 Generalized Cross-Correlation (GCC)

GCC has been a popular method to determine the time-difference of arrival (TDOA) between two microphones in a pair [24]. Then from multiple TDOA values, one can estimate the source location. Take a 4-element microphone array as an example as shown in figure 2.5.1.1

![TDOA Between two microphones](image)

**Figure 2.5.1: TDOA Between two microphones**
If the distance from microphone \( m \) to the source is \( r_m \) \( (m=1,2,3,4) \), the time delay (traveling time) of the signal from the source to that microphone is,

\[
\tau_m = \frac{r_m}{c}
\]  

(2.5.1.1)

Then the time-difference of arrival, TDOA between two microphones \( m \) and \( n \) can be defined as,

\[
\tau_{mn} = \tau_m - \tau_n = \frac{r_m - r_n}{c}
\]  

(2.5.1.2)

From this relation between the TDOA and the distances from the source to the microphones, \( r_m \), one can estimate the source location from multiple TDOA’s [2].

2.5.1.1 Derivation of the GCC

Let us consider a microphone signal at microphone \( k \) be.

\[
x_k(t) = s(t) * h(d_s, t) + n_k(t)
\]  

(2.5.1.1.1)

Consider a signal at another microphone \( l \) be,

\[
x_l(t) = s(t - \tau_{kl}) * h(d_s, t) + n_l(t)
\]  

(2.5.1.1.2)

Note that to be accurate, we would have to include the time delay \( \tau_k \) into the source signal \( s(t) \), i.e. \( s(t-\tau_k) \) in above equation 2.5.1.1.2 to show the signal received at microphone \( k \) is a delayed version of the source signal. However, for simplicity, here we normalized so that the time delay from the source to microphone \( k \), \( \tau_k \) is 0. In other words, we are only concerned with the relative time-difference of arrival, \( \tau_{kl} \) between these two microphones \( k \) and \( l \) [20].

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, \( \tau_{kl} \). The cross-correlation of \( x_k(t) \) and \( x_l(t) \) is defined as,

\[
C_{kl}(\tau) = \int_{-\infty}^{\infty} x_k(k)x_l(t + \tau)dt
\]  

(2.5.1.1.3)

Taking the Fourier transform of the cross-correlation results in a cross spectrum,
Applying convolution properties of the Fourier Transform for 2.5.1.1.3 when substituting it into 2.5.1.1.4, we have,

\[ C_{kl}(\omega) = X_k(\omega) X_l^*(\omega) \]  

(2.5.1.1.5)

Where \( X_l(\omega) \) is the Fourier Transform of signal \( x_l(t) \), and \( '*' \) denotes the complex conjugate.

The inverse Fourier transform of 2.5.1.1.5 gives us the cross-correlation function in terms of the Fourier Transform of the microphone signals:

\[ C_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_k(\omega) X_l^*(\omega) e^{i\omega \tau} d\omega \]  

(2.5.1.1.6)

The generalized cross-correlation (GCC) of \( x_k(t) \) and \( x_l(t) \) is the cross-correlation of their two filtered versions. Denoting the Fourier transforms of these two filters as \( W_k(\omega) \) and \( W_l(\omega) \), we have the GCC, \( R_{kl}(T) \) is defined as,

\[ R_{kl}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (W_k(\omega) X_k(\omega))(W_l(\omega) X_l(\omega))^* e^{i\omega \tau} d\omega \]  

(2.5.1.1.7)

A combined weighting function, \( \Psi_{kl}(\omega) \) is defined as

\[ \Psi_{kl}(\omega) \equiv W_k(\omega) W_l^*(\omega) \]  

(2.5.1.1.8)

Substituting 2.5.1.1.8 into 2.5.1.1.7, the GCC becomes

\[ R_{kl}(\tau) \equiv \int_{-\infty}^{\infty} \Psi_{kl}(\omega) X_k(\omega) X_l^*(\omega) e^{i\omega \tau} d\omega \]  

(2.5.1.1.9)

The TDOA between two microphone k and l is the time lag T that maximizes the GCC \( R_{kl}(T) \) in the real range limited by the distance between the microphones:

\[ \tau_{k,l} = \arg\max R_{k,l}(\tau) \]  

(2.5.1.1.10)
2.5.2 The Phase Transform (PHAT)

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

\[ \Psi_{kl}(\omega) \equiv \frac{1}{|X_k(\omega)X_l^*(\omega)|} \]  \hspace{1cm} (2.5.2.1)

2.5.3 GCC-PHAT

Applying the weighting function PHAT from above equation 2.5.2.1 into the expression for GCC in equation 2.5.1.1.9, the Generalized Cross-Correlation using the Phase Transform (GCC-PHAT) for two microphones \( k \) and \( l \) is defined,

\[ R_{k,l}(\tau) \equiv \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{|X(\omega)X(\omega)|}X(\omega)X^*(\omega) e^{j\omega\tau} d\omega \]  \hspace{1cm} (2.5.3.1)

2.6 SRP-PHAT

A further development of GCC is the SRP which extends the GCC to an arbitrary number of sensors and arbitrary sensor configurations \([18]\), SRP type of algorithms are methods based on steering a beam former across the search space in order to find maxima in the response power in the beam former output signal. The SRP of a filter-and-sum beamformer output signal. The SRP of a filter-and-sum beam former with \( M \) sensors in a uniformly linear sensor array can be stated as

\[ P(\tau) = \sum_{(k,l) \in P} \int_{-\infty}^{\infty} \Psi(\Omega)G_{kl}(\Omega) e^{j\Omega(\tau-\tau)} d\Omega \]  \hspace{1cm} (2.6.1)

Where \( P \) is the set of all combinations of sensor pairs within the array. The TDOA for an acoustic source is estimated as

\[ \hat{\tau} = \arg\max_{\tau} P(\tau) \]  \hspace{1cm} (2.6.2)

The defined SRP is equivalent to the sum of the GCC for all combinations of sensor pairs in the sensor array, which means that for a two-sensor array the SRP is equivalent to the GCC. The
SRP-PHAT is obtained by using the weighting function $\psi(\Omega)$, which combines the increased accuracy in the SRP over the GCC and the robust weighting by the PHAT.

DOA’s are calculated from TDOA for a uniform linear microphone array as

$$\hat{\alpha} = \arcsin\left(\frac{c}{f_s d} \cdot \frac{T}{\Delta f}\right)$$

(2.6.3)

Where $T$ is the estimated TDOA, $c$ is the propagation speed of sound, $F_s$ is the sampling frequency and $d$ is the distance between microphones [9].
CHAPTER-3

Design and Implementation in MATLAB

3.1 Microphone Array Setup

In my thesis, a microphone array of 16 microphones is modeled in a 4x4 uniform linear array, separated by a distance \( d \) in such a way that to avoid aliasing effect. The speech signal used here is sampled at 8 \( KHz \) so the distance between microphones i.e \( d \) should be less than 4.2 \( cm \) to avoid the aliasing effect. So the total distance between microphones in the array should not exceed 4.2 \( cm \) and microphone positions are taken for a 16 sensor array in 2D coordinates as,

\[
\text{Mic1} = [3.98 \quad 3.98]; \quad \text{Mic2} = [3.99 \quad 3.98]; \quad \text{Mic3} = [4.00 \quad 3.98]; \quad \text{Mic4} = [4.01 \quad 3.98]; \\
\text{Mic5} = [3.98 \quad 3.99]; \quad \text{Mic6} = [3.99 \quad 3.99]; \quad \text{Mic7} = [4.00 \quad 3.99]; \quad \text{Mic8} = [4.01 \quad 3.99]; \\
\text{Mic9} = [3.98 \quad 4.00]; \quad \text{Mic10} = [3.99 \quad 4.00]; \quad \text{Mic11} = [4.00 \quad 4.00]; \quad \text{Mic12} = [4.01 \quad 4.00]; \\
\text{Mic13} = [3.98 \quad 4.01]; \quad \text{Mic14} = [3.99 \quad 4.01]; \quad \text{Mic15} = [4.00 \quad 4.01]; \quad \text{Mic16} = [4.01 \quad 4.01];
\]

Total distance of Microphone Array = 4 \( cm \)

So the distance between the microphones is modeled as 1 \( cm \) and the total distance of the microphone array also known as aperture is 4\( cm \). For testing the algorithm the speech and noise source positions are placed at different angles, in detail speech source is moved in a semicircle with radius of 1.5 \( m \) from the midpoint between the microphones i.e. [3.995, 3.995] at an angle of -90 to +90 degrees and the noise source is placed at a fixed location at a distance 4 \( m \) from the midpoint of the microphone array at an angle of 45 degrees.
3.2 Fractional Delay Filter

The amounts of signal delayed in reaching to the 16 microphones are performed by using Thiran FD filter. Let us consider a White Gaussian Noise signal is given as input and its position at [8,8], Microphone positions are at Mic1=[3.98 3.98]; Mic2=[3.99 3.98]; Mic3=[4.00 3.98] and Mic4=[4.01 3.98] then the signal gets a time delay in samples of D1= 265.04, D2= 264.72, D3= 264.38 and D4= 264.05 in reaching the microphones Mic1, Mic2, Mic3 and Mic4 respectively. The delays D1, D2, D3 and D4 contains both integer part and fractional part, with this delay we construct FD filter using Thiran. This amount of delays in the signal varies as the distance between the signal and microphone varies. The figure 3.2.1 shows the original signal and the delayed signals with respect to Mic1, Mic2, Mic3 and Mic4.

![Figure 3.2.1 The original signal and the delayed signals with respect to Mic1, Mic2, Mic3 and Mic4](image_url)
3.3 Filter Bank

In my thesis for SRP-PHAT source localization method, WOLA filter bank is used to transform the signal into sub-band signals. The parameters which plays key role in modeling a filter bank are prototype filter length $N$, number of subbands $M$, oversampling ratio $OS$ and decimation rate $D$. In my thesis $OS=2$, $M=256$, $N=1024$, $D=256$ are used. Here flow of simulation is, we assume a block of samples at rate $D=256$, process this set of samples through analysis bank and instead of reconstructing this samples of data using synthesis bank, we perform a search for maximum peak value in the range of $-T_{max}$ to $T_{max}$ where $T_{max}$ is defined as the maximum time lag.

3.4 Room Impulse Response

To perform a real implementation scenario, $RIR$ setup is used in my thesis. This setup filters the signals using room impulse response filter and the resulted signal would consist of direct path signal and also reverberated signals. The parameters which plays key role in modeling a $RIR$ are Number of virtual sources $N$, reflection coefficient of the wall $R$ and size of the room $Rm$. In this thesis $N=6$, i.e, the model will count for $(2*6+1)^2 = 169$ virtual sources, $R=0.65$, $Rm=[8,8]$ are used. Figure 3.4.1 shows the RIR coefficients for source position at $Src=[5.5,4]$.

![RIR filter coefficients](image)

Figure 3.4.1 The RIR filter coefficients for $R=0.65$ & Source position $Src=[5.5,4]$
3.5 SRP-PHAT Implementation

In my thesis, I have done implementation of SRP-PHAT for 16 Mics in 2 types, i.e : 1. Row-wise Setup, 2. Column-wise Setup. Type 1: Row-wise implementation is done by following way,

\[
\begin{align*}
\text{Mic1} & = [3.98 \ 3.98]; \quad \text{Mic2} = [3.99 \ 3.98]; \quad \text{Mic3} = [4.00 \ 3.98]; \quad \text{Mic4} = [4.01 \ 3.98]; \quad \rightarrow \text{Row wise} = R_1 \\
\text{Mic5} & = [3.98 \ 3.99]; \quad \text{Mic6} = [3.99 \ 3.99]; \quad \text{Mic7} = [4.00 \ 3.99]; \quad \text{Mic8} = [4.01 \ 3.99]; \quad \rightarrow \text{Row wise} = R_2 \\
\text{Mic9} & = [3.98 \ 4.00]; \quad \text{Mic10} = [3.99 \ 4.00]; \quad \text{Mic11} = [4.00 \ 4.00]; \quad \text{Mic12} = [4.01 \ 4.00]; \quad \rightarrow \text{Row wise} = R_3 \\
\text{Mic13} & = [3.98 \ 4.01]; \quad \text{Mic14} = [3.99 \ 4.01]; \quad \text{Mic15} = [4.00 \ 4.01]; \quad \text{Mic16} = [4.01 \ 4.01]; \quad \rightarrow \text{Row wise} = R_4
\end{align*}
\]

Row Wise SRP-PHAT Estimation \( SRP_R = (R_1 + R_2 + R_3 + R_4) / 4 \)

\( R_1 \) is evaluated for SRP-PHAT using 4 Mics in row-wise, like that \( R_2 \), \( R_3 \) and \( R_4 \) is evaluated individually and then \( SRP_R \) is calculated by taking average of all these row-wise values.

Type 2: Column – wise implementation is done by following way, \( C_1 \) is evaluated for SRP-PHAT using 4 Mics in column-wise, like that \( C_2 \), \( C_3 \) and \( C_4 \) is evaluated individually and then \( SRP_C \) is calculated by taking average of all these column-wise values.

\[
\begin{align*}
\text{Mic1} & = [3.98 \ 3.98]; \quad \text{Mic2} = [3.99 \ 3.98]; \quad \text{Mic3} = [4.00 \ 3.98]; \quad \text{Mic4} = [4.01 \ 3.98]; \\
\text{Mic5} & = [3.98 \ 3.99]; \quad \text{Mic6} = [3.99 \ 3.99]; \quad \text{Mic7} = [4.00 \ 3.99]; \quad \text{Mic8} = [4.01 \ 3.99]; \\
\text{Mic9} & = [3.98 \ 4.00]; \quad \text{Mic10} = [3.99 \ 4.00]; \quad \text{Mic11} = [4.00 \ 4.00]; \quad \text{Mic12} = [4.01 \ 4.00]; \\
\text{Mic13} & = [3.98 \ 4.01]; \quad \text{Mic14} = [3.99 \ 4.01]; \quad \text{Mic15} = [4.00 \ 4.01]; \quad \text{Mic16} = [4.01 \ 4.01];
\end{align*}
\]

Column wise=\( C_1 \) \quad Column wise=\( C_2 \) \quad Column wise=\( C_3 \) \quad Column wise=\( C_4 \)

Column Wise SRP-PHAT Estimation \( SRP_C = (C_1 + C_2 + C_3 + C_4) / 4 \)
Figure 3.5.1 Overview of 4 element microphone setup (Row-Wise setup (Type 1), \( R_1 \) ) showing that speech source is moved at a distance from reference point of microphone array from -90° to 90° and noise source is fixed at a location at 45°.

Figure 3.5.2 Overview of 4 element microphone setup (Column-Wise setup (Type 2), \( C_1 \) ) showing that speech source is moved at a distance from reference point of microphone array from -90° to 90° and noise source is fixed at a location at 45°.
Figure 3.5.3: Implementation of SRP-PHAT for 4 Microphones Setup, Only Speech
Figure 3.5.4: Implementation of SRP-PHAT for 4 Microphones Setup, Speech and Noise
CHAPTER-4

4. Evaluation of SRP-PHAT Using 16 Microphones in Different Environments

4.1 Test data

4.1.1 Speech Signal

In this thesis, a speech signal "Benny_radio.wav" contains male voice saying "One two three four five six seven eight nine ten" with some music and a female voice in the background and this file is recorded at a sampling frequency of 8KHz and it is of 8 seconds duration and contains 72077 samples of data. It is as shown in figure 4.1.1.

![Speech Signal Waveform](image)

Figure 4.1.1 The Test speech signal benny_radio

4.1.2 Noise Signal

In this thesis, a noise signal ‘Brown_Noise’ is used and this file is recorded at sampling frequency of 32KHz and contains 72077 samples of data. It is as shown in figure 4.1.2.
Performance of the SRP-Phat method, the accuracy of the estimated TDOA's and DOA’s can be measured using Standard Deviation and Mean Square Error (which is the mean of the difference between the estimated angle and the original angle).

### 4.2.1 ONLY SPEECH SIGNAL

A speech signal with no added noise other than the environment noise together with reverberation created in the recording environment called ‘Benny_radio’ with 8000 HZ sample rate is used in this test. Table 4.2.1 shows the original and estimated TDOA values in samples for 256 subbands for all positions from +90’ to -90’ for different setups. Figure 4.2.1.1 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-1 Column-wise and figure 4.2.1.3 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-2 Row-wise. From this figure, it is apparent that the variance of the DOA estimates increases as the angle approaches the end fire. The mean estimated error for 256 subbands for Row-wise and Column-wise setups are plotted in figure 4.2.1.4 and figure 4.2.1.2. The mean estimation errors in degrees are also seems to have a tendency to underestimate the angle as the source approaches the end fire.
Figure 4.2.1.1 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-1 Column-wise.

Figure 4.2.1.2 shows the mean estimation error of the estimated DOA for 256 sub-bands for Setup-2 Column-wise.
Figure 4.2.1.3 shows the standard deviation of the estimated DOA for 256 sub-bands for Setup-2 Row-wise.

Figure 4.2.1.4 shows the mean estimation error of the estimated DOA for 256 sub-bands for Setup-2 Row-wise.
4.2.2 SPEECH AND NOISE SIGNALS

The same speech signal with an added Brown noise at different Signal to Noise Ratios (SNR) from 0 to 25 dB for 256 subbands is tested. Standard Deviation of DOA estimate is calculated for speech source moving from -90’ to 90’ at a distance of 1.5 m from the reference point of the microphone array and noise played at angle of 45’ constant at a distance of 4 m from reference point of the microphone array. Figures 4.2.2.1 to 4.2.2.12 shows the standard deviation error and mean estimation error for estimated DOA from SRP-Phat for 0 to 25dB for Column-wise setup. Figures 4.2.2.13 to 4.2.2.24 shows the standard deviation error and mean estimation error for estimated DOA from SRP-Phat for 0 to 25dB for Row-wise setup. These figures show that SRP-PHAT works in noisy environments.
Figure 4.2.2.1 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-1 Column-wise

Figure 4.2.2.2 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-1 Column-wise
Figure 4.2.2.3 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-1 Column-wise.

Figure 4.2.2.4 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-1 Column-wise.
Figure 4.2.2.5 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-1 Column-wise.

Figure 4.2.2.6 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-1 Column-wise.
Figure 4.2.2.7 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-1 Column-wise.
Figure 4.2.2.8 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-1 Column-wise.

Figure 4.2.2.9 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-1 Column-wise.
Figure 4.2.2.10 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-1 Column-wise.

Figure 4.2.2.11 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-1 Column-wise.
Figure 4.2.2.12 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-1 Column-wise.

![Bar chart showing mean estimation error for a speech and noise source with input SNR=25dB.](image)

Figure 4.2.2.13 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-2 Row-wise.

![Line graph showing standard deviation for a speech and noise source with input SNR=0dB.](image)

Figure 4.2.2.14 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=0dB, for 256 sub-bands for Setup-2 Row-wise.

![Line graph showing mean estimation error for a speech and noise source with input SNR=0dB.](image)
Figure 4.2.2.15 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-2 Row-wise.

Figure 4.2.2.16 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=5dB, for 256 sub-bands for Setup-2 Row-wise.
Figure 4.2.2.17 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-2 Row-wise.
Figure 4.2.2.18 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=10dB, for 256 sub-bands for Setup-2 Row-wise.

Figure 4.2.2.19 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-2 Row-wise.
Figure 4.2.2.20 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=15dB, for 256 sub-bands for Setup-2 Row-wise.

Figure 4.2.2.21 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-2 Row-wise.
Figure 4.2.2.22 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=20dB, for 256 sub-bands for Setup-2 Row-wise.

Figure 4.2.2.23 shows the standard deviation of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-2 Row-wise.
Figure 4.2.2.24 shows the mean estimation error of the estimated DOA for a speech and noise source with input SNR=25dB, for 256 sub-bands for Setup-2 Row-wise.
CHAPTER-5

CONCLUSION AND FUTURE WORK

In this thesis work, SRP-Phat based speech source localization method is implemented using a 16 element microphone array. This 16 element microphone array has been arranged into two different setups namely, Row-setup and Column-setup where the microphones are arranged in a 4x4 matrix as 4 rows and 4 columns. Source localization using SRP-Phat works in both the row and column setups. As a huge microphone array is used in the setup, computation time is increased and SRP-Phat algorithm does estimate only the DOA of the speech source and it doesn’t determine the position of the speech source. Also, it is difficult if there are multiple talkers, especially in real-time environment. Performance of the source localization method can be changing the position of the microphones in the array. If the distance between the microphones is decreased, the performance of the setup can be improved and hence, more number of microphones could be accommodated in the microphone array for better video surveillance. In future, this work can be extended to implement other source localization methods like SRK-Phat using a huge microphone array and compare the performance of both the methods using measures like standard deviation and mean estimation error in the presence of environment noise and reverberation effect.
REFERENCES


