Master Thesis

Electrical Engineering with
Emphasis on signal processing

Implementation and evaluation of spectral subtraction (SS) with minimum statistics and wiener beamformer combination.

VENKATA RAMI REDDY DATLA

Supervisor: Dr. Nedelko Grbic
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:

Authors:
Venkata Rami Reddy Datla
E-mail: vedb10@student.bth.se

Supervisor:
Dr. Nedelko Grbic
School of Engineering (ING)

Examiner:
Dr. Sven Johansson
School of Engineering (ING)

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

In mobile communication systems, speech signal can be distorted by acoustic noise such as babble speech noise. Such kind of noises can distort the quality of speech signal to a high extent. All speech processing equipments like the noise cancelling headphones and hearing aids should be able to filter different kinds of interfering signals and present a clear sound to the listener. Eliminating the noise present in the signal without affecting the original speech is a challenge and a topic of research in these days.

This paper concentrates on enhancing the quality of speech and reducing the noise on same hand by making use of Wiener Beamformer and Spectral Subtraction based on minimum statistics. Wiener beam former is implemented based on sub-band approach and Direction of Arrival capabilities. Both the methods are cascaded in a unique fashion so maximum speech intelligibility is obtained. All the systems are tested with various positions of noise and speech source with reference to the microphone array. All proposed methods are implemented, evaluated and results are simulated using MATLAB. The performance of the method is analyzed using Signal-to-Noise Ratio improvement (SNRI).
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and thanks to my thesis supervisor Dr. Nedelko Grbic for providing a chance to do my thesis research work under his supervision and Dr. Sven Johansson as examiner in the field of Speech Processing. I would like to thank them for the persistent help throughout the thesis work. With their deep knowledge in this field which helped us to learn new things in order to complete master thesis successfully. The continuous feedback and encouragement helped us in doing this thesis work.

I extend my appreciation and thanks to my fellow student Siva kumar Chikkala for his suggestions and discussions regarding solving different problems in doing this research thesis.

I would like to thank BTH for providing us a good educational environment where we can gain the knowledge and learn about new technologies that help us to move forward with the thesis work.

Finally, I would like to extend my immense gratitude and wholehearted thanks to my parents for their moral support and financial support throughout my educational career. They have motivated and helped us for the successful completion of thesis work. We also thank my pals for their support and encouragement during the thesis work. We take an opportunity to thank all the signal processing staff at BTH.
# TABLE OF CONTENTS

ABSTRACT .......................................................................................................................... iii
ACKNOWLEDGEMENTS ........................................................................................................ iv

List of Figures ........................................................................................................................ viii
List of Tables .......................................................................................................................... x

1 INTRODUCTION ............................................................................................................. 1
   1.1 Problem Statement ...................................................................................................... 1
   1.2 Scope of the Thesis .................................................................................................... 1
   1.3 Thesis Outline .......................................................................................................... 2

2 Background and Related information .............................................................................. 3
   2.1 Auditory Scene Analysis .......................................................................................... 4
   2.2 Source Localization ................................................................................................. 5
   2.3 Noise Reduction ....................................................................................................... 5
   2.4 Hands-free speech communication problem ............................................................ 7
   2.4.1 Background noise ............................................................................................... 8
   2.4.2 Acoustic coupling ............................................................................................... 8
   2.5 Applications ............................................................................................................ 9
      a) Voice control ........................................................................................................ 9
      b) Hearing aids .......................................................................................................... 9
   2.6 Microphone array processing ................................................................................... 10
   2.7 Single Channel Speech Enhancement ..................................................................... 12
   2.8 Fractional Delay Filters .......................................................................................... 13
      2.8.1 Ideal Fractional delay ...................................................................................... 13
      2.8.2 FIR approximation of fractional delay .............................................................. 15
      2.8.3 IIR approximation of fractional delay ............................................................... 16
List of Figures

Figure 1: Generalized Communication System ------------------------------------------3
Figure 2: Typical hands-free speech communication environment------------------------7
Figure 3: Sensitivity patterns of ideal omnidirectional and directional microphones ----11
Figure 4: Structure of microphone array system --------------------------------------12
Figure 5: Sampled(dots) impulse response of an ideal fractional delay filter when the delay is (a) D = 2 samples and (b) D = 2.7 samples. The vertical line indicates the midpoint of the continuous time impulse response in each case --------------------------15
Figure 6: Microphone Array with θ as Direction of Arrival and τ as Time Delay of Arrival -----------------------------------------------------------------------------------18
Figure 7: Overview of the extreme positions that source can occupy ------------------20
Figure 8: Block diagram showing Estimation of DOA with two Microphones [6] ------21
Figure 9: Model of Wiener Beamformer with two Microphones------------------------23
Figure 10: Linear microphone array with all microphones equidistant at distance------24
Figure 11: Representing general form of Spectral subtraction-------------------------27
Figure 12: A bank of eight band pass filters hk(n), with Fourier transforms k(f)=F{hk(n)}, comprise filter bank --------------------------------------------------28
Figure 13: Analysis-synthesis filter bank --------------------------------------------29
Figure 14: an example of noisy speech power spectrum and local minimum noise estimation -------------------------------------------------------------------------------------31
Figure 15: short time sub-band power and estimated noise floor [11] -----------------33
Figure 16: Block diagram showing the spectral subtraction approach used in the thesis 33
Figure 17: Representing cascaded implementation -------------------------------------37
Figure 18: Long-term magnitude spectrum of the speech-shaped noise-----------------38
Figure 19: Clean speech signal sampled at 16 kHz-----------------------------------39
Figure 20: Power spectrum of clean speech signal sampled at 16 kHz-------------------39
Figure 21: Speech signal corrupted with speech-shaped noise at 5 dB SNR

Figure 22: Speech signal corrupted with speech-shaped noise at 0 dB SNR

Figure 23: Output from the wiener beamformer for speech at 90° and noise at 0°

Figure 24: Output of wiener beamformer at speech at 345° and noise at 180°

Figure 25: SNR improvement for noise corrupted speech

Figure 26: SNR improvement for noise corrupted speech

Figure 27: SNR improvement for noise corrupted speech

Figure 28: SNR improvement for noise corrupted speech

Figure 29: Clean speech corrupted by babble noise (above) and output of the spectral subtraction (below)

Figure 30: SNR improvement for noise corrupted speech

Figure 31: Clean speech corrupted by noise (above) and output of cascaded implementation (below)

Figure 32: SNR improvement for noise corrupted speech

Figure 33: SNR improvement for noise corrupted speech

Figure 34: SNR improvement for noise corrupted speech

Figure 35: SNR improvement for noise corrupted speech
List of Tables

Table.1: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 90° and noise at 0°---------------------------------------------------------------44

Table.2: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 135° and noise at 225°-----------------------------------------------------------------------------------------------45

Table.3: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 210° and noise at 330°-------------------------------------------------------------------------------46

Table.4: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 345° and noise at 180°----------------------------------------------------------------------------------------47

Table.5: Tabular form for SNRI for spectral subtraction------------------------------------------------------------------------------------------------------------------------49

Table.6: Tabular form showing values cascaded implementation in terms of SNRI for speech at 90° and noise at 0°-------------------------------------------------------------51

Table.7: Tabular form showing values for cascaded implementation in terms of SNRI for speech at 135° and noise at 225°------------------------------------------------------------------------------------52

Table.8: Tabular form showing values for cascaded implementation in terms of SNRI for speech at 210° and noise at 330°-------------------------------------------------------------------------------------53

Table.9: Tabular form showing values for cascaded implementation in terms of SNRI for speech at 345° and noise at 180°--------------------------------------------------------------------------------------54
CHAPTER – 1

INTRODUCTION

In many speech related systems like mobile communication in an adverse environment, the desired signal is not available directly; rather it is mostly contaminated with some interference sources of noise. These background noise signals degrade the quality and intelligibility of the original speech, resulting in a severe drop in the performance of the applications. The degradation of the speech signal due to the background noise is a severe problem in speech related systems and therefore should be eliminated through speech enhancement algorithms [1].

1.1 Problem Statement:

An effective speech enhancement mechanism requires advanced noise cancellation methods combined with Direction of Arrival (DOA) estimation and source tracking. For robust and accurate speech enhancement scenario, two concepts are to be considered. Firstly, directional signal reception is to be achieved. This can be done by the use of microphone arrays or more precisely, the adaptive beamforming concepts. Secondly, the type of speech processing algorithm that can be used to achieve noise reduction, like Spectral subtraction using noise estimation based on minimum statistics. Furthermore, concepts like filter banks and frequency domain analysis can also help in solving the issue.

1.2 Scope of the Thesis:

This thesis work aims at designing and implementing wiener beamformer along with spectral subtraction with noise estimation based on minimum statistics. The work flow can be categorized into four main parts:

1. Implementation of beam formers including Direction of Arrival.
2. Designing of acoustic noise cancellation system in relation to beamformers.
4. Cascaded implementation of the entire speech enhancement setup which includes all the above mentioned systems.
1.3 Thesis Outline:

This thesis is a part of the complete work done and mainly concentrates on implementing a method which can improve quality of speech to desired level. This report is organized into four chapters. The introduction chapter deals with the problem and its evaluation. Background and related work is dealt in chapter 2 where information about speech enhancement techniques is available. Chapter 3 deals with implementation and chapter 4 deals with implementation and evaluation. Finally, the paper is concluded with an insight to the future work in the last chapter.
CHAPTER – 2

Background and Related information

Background

All the modern hand-held communication devices such as mobile handsets have an inherent digital signal processing system which processes a speaker’s voice, or more commonly called the speech signal to be transmitted. This speech processing mechanism is developed from observing functionality of human ear. Human ear is a sensory organ that has the amazing capability of recognizing a sound. It has an ear drum inside it which starts vibrating when a sound wave impinges on it. Comparing it with an electrical system, it actually acts as a transducer which converts sound waves into a signal of nerve impulses and transmit it to the brain. As mentioned previously, ear basically does the primary objective of recognizing a sound but the actual processing is performed in the brain [1].

The exact behavior of human ear is replicated as the sound generated from a source is recorded with a microphone which has a diaphragm inside it just as the ear has ear drum. Both are sensitive to the sound wave pressure and acts like transducers. The microphone output is connected to a signal processing system which processes the sound similar to the brain.

![Generalized Communication System](image)

**Figure 1: Generalized Communication System [2]**

Figure 1 shows a typical communication system in which a speaker’s voice is recorded with the help of a microphone. The electrical signal from the microphone output is then digitized and transmitted to the receiving end of the system. The receiver section can be of any form such as a
broadcasting system (radio), a mobile phone, a noise cancelling headset or a hearing aid of a listener. The overall system may differ depending upon the application. For example, a hearing aid device does not have an additional receiving unit or channel but the digitized input is directly processed with an extremely high sampling rate and an amplifier, this kind of sampling is known as one bit over sampled converter. Over sampling is a mathematical technique in which the signal is sampled at rate much higher than Nyquist frequency. The effect of this type of sampling is to shift circuit noise that can occur during the quantization process into very high frequencies, where it can be easily filtered out [1]

If we try to adapt the entire functionality of the ear into various topics related to speech processing, we end up to three main fields of study as follows:

- Auditory Scene Analysis [1]
- Source Localization [1]
- Noise Reduction techniques [1]

2.1 Auditory Scene Analysis:

Brain has the inherent capability of whether to integrate or segregate the complex sound recognized by the ear. The signal received by the human ear has to be analyzed (in some way). Auditory scene analysis proposes that sounds will either be heard as "integrated" (heard as a whole -- much like harmony in music), or "segregated" into individual components (which leads to counterpoint). For example, a bell can be heard as a 'single' sound (integrated), or some people are able to hear the individual components -- they are able to segregate the sound. This can be done with chords where it can be heard as a 'color', or as the individual notes.

In many circumstances the segregated elements can be linked together in time, producing an auditory stream. This ability of auditory streaming can be demonstrated by the so-called cocktail party effect. Up to a point, with a number of voices speaking at the same time or with background sounds, one is able to follow a particular voice even though other voices and background sounds are present. In this example, the ear is segregating this voice from other sounds (which are integrated), and the mind "streams" these segregated sounds into an auditory stream. [14]

Segregation of sound is nothing but separating a complex sound into individual components. This can be seen in situations like a person trying to make a conversation with another person in an open place surrounded by automobile and many other noises.

Now considering the famous cocktail party effect [15], the ear integrates and segregates at the same time. The segregation is utilized to listen to a particular person inside the party hall while the integration is used to integrate all the remaining sounds including music. This natural behavior of ear is termed as “Auditory Scene Analysis” (ASA). ASA is the key concept of speech processing. Any speech processing experts aims at gaining knowledge on these concepts and then proceed with advanced topics like speech enhancement, source localization and blind source separation. However, the ability of intentional selectivity in a cocktail party effect is attributed by the fact that we have two ears and our perception of speech is based on binaural hearing. This suggests the use of multiple microphones in the development of modern practical speech acquisition systems [4].
2.2 Source Localization:

When someone calls out our name, we will immediately turn around and answer to that particular person who called us. This is possible because our ears has the inherent capability to locate the person who is talking and then brain matches these signals to the memory database so that we can identify the person who is talking [Acoustic source localization based on independent component analysis]. This property of ears is called as “Sound Source Localization”. The human auditory system makes use of many different cues for the purpose of sound source localization such as time and level differences between both ears, spectral information, timing analysis, correlation analysis and finally pattern matching [12].

2.3 Noise Reduction:

Speech Signal Enhancement has a gained lot of importance in the present day real world situations. Research on speech enhancement techniques started more than 40 years ago at AT&T Bell Laboratories by Schroeder as mentioned in [3].

Speech Enhancement algorithms have a very diverse range of applications. Examples for such applications are: Mobile communications, Intelligent Hearing Protectors in construction and mining, conference telephony, Laptop microphones including Dictaphones and in medical field such as Hearing aids [J. Benesty, S. Makino, and J. Chen, “Speech Enhancement”, Springer-Verlag Berlin Heidelberg, 2005.]. All these applications require the speech enhancement algorithms for improving the speech quality, the speech intelligibility or damage control to the ear. Reducing the inherent noise improves the speech quality whereas speech intelligibility refers to the degree of ability to understand speech. Typical interfering signals can be of any form, for example - the sounds coming from the traffic when a speaker is delivering a speech, two persons speaking at the same time, the whistle of a train when someone is on a call inside and so on. All these examples can be mentioned in one name called background noise. There are other kinds of degradation possible like the distortion of microphone and reverberation. Reverberation is the persistence of sound in a particular space after the original sound is produced. Reverberation, is created when a sound is produced in an enclosed space causing a large number of echoes to build up and then slowly decay as the sound is absorbed by the walls and air.[16]

Many algorithms have been proposed for the purpose of speech enhancement over the last few decades. These algorithms depend upon the type of the corrupted speech signal which can be broadly classified into two types namely [3]:

- Environmentally Disturbed speech signals.
- Acoustically disturbed speech signals.

Environmentally disturbed speech signals are the signals that are affected by noisy surroundings such as a cocktail party effect, the whistle of a train, reflections from trees etc. These kinds of disturbances affect the original speech signal considerably and hence are to be eliminated. Typical algorithms that are used to suppress these kinds of disturbances include: Adaptive Filters [1], Adaptive Noise Cancellers [1] and Filter Bank topologies [3].
Acoustically disturbed speech signals are the ones that are affected by echoes and reflections. Multiple reflections in a closed room that reach a listener causes some kind of disturbance or delay in the original signal called Reverberation. Development of algorithms for the enhancement of these kinds of signals is the latest area of research. Basic beamformers and microphone arrays along with many spectral processing techniques come in to use for this enhancement [3].

Hands-free communication is the area which has undergone tremendous advancement in the recent past. It covers many things such as mobile telephony, hearing aids and automatic information systems i.e. voice controlled systems, video conferencing systems and many of the multimedia applications. More and more people are using personal communication devices, personal computers and wireless mobile telephones which in turn transforming into advanced personal communication systems. The advancements in interpersonal communication systems are realized by continuous effect for improving and extending the interaction between individuals, which are not only provides user safety and quality but it is user friendly too. The combination of telephone technologies and computers are making way for convenient hands-free communication. The advancement in wireless communication technology has provided ease of usage for voice connectivity in cellular communication and personal computer devices in order to enabling the natural communication in different environments such as cars, restaurants and offices. In hand-controlled automobile applications, the functionalities are processed with voice controls, the signal degradations in this field are same as that of distant-talkers speech recognition applications. [4]

Audio conference plays a key role in communication systems for small scale and a large scale firm which is cost effective and also aimed for user comforts. In present generations, the demand for voice controlled systems is high as the hand-controlled functions are replaced with voice controls which are efficient and also robust. The importance of speech processing techniques have been analyzed for capability of preventing damage to hearing in high-noise environments and also improving speech intelligibility in noise for hearing impaired listeners. [8]

Hands-free speech acquirement plays a vital role in all above mentioned applications. In automated speech system design the microphone is placed far away from the user (speech transmitter and receiver are installed at remote places with certain distances in between them) due to which problem like poor sound quality and acoustic echo arise from far-end side. The poor sound quality is because of the microphone placed near to the speaker due to which it suffers from unwanted disturbances caused by environment caused by environmental noise, interfering sounds and reverberation of speech signal from loudspeaker corrupts the actual speech signal. In full-duplex hands-free communication acoustic echo is generated at the near end side at microphone causes disturbance to the speaker at the far end side in which listener hear his own voice with 100-200 ms delay [4]. This leads to reduce intelligibility of the received speech in a noisy conditions and also degrading the speech in speech recognition systems. The degradation in the received speech signals makes conversation between the users difficult. For improvement in the quality of the hands-free mobile telephones, the major tasks to be considered are background noise suppression, interference reduction and acoustic echo cancellation. For the improvement of the speech quality and reducing unwanted disturbances several speech enhancement methods are implemented for robust speech communication system. Microphone arrays are widely used
technology for speech enhancement in communication systems were speech quality and speech intelligibility is being degraded due to a noisy environment and room reverberations [8].

The perception of speech signal is measured in terms of quality and intelligibility. The “Quality” is a subjective measure which reflects on the individual preferences of listeners [1]. The “Intelligibility” is an objective measure which predicts the percentage of words that can be correctly identified by listeners [1]. Speech enhancement is to improve noisy speech signals. The received speech signals in automated speech are mainly corrupted by background noise. In general, the background noise can be non-stationary and the signal to noise ratio (SNR) decreases if the noise level increases. Since a few decades the research in speech enhancement methods of acoustically distributed signals has been performed widely and the contribution of digital hearing aids has significantly improved the research in hands-free communication systems.

The acoustic echo cancellation plays a key role in acoustically coupled environments. The acoustic echo plays a major role in degrading the speech intelligibility in speech communication systems like hearing aids and telecommunication systems. In this thesis, adaptive methods like LMS, NLMS and APA algorithms are used to cancel the acoustic echo.

2.4 Hands-free speech communication problem.

Different hands free communication application and their surrounding environments were described. The major problems challenged in each application are background noise, room reverberation and acoustic echo. Typical hands-free communication environment is as shown in Figure 2

![Figure 2: Typical hands-free speech communication environment [17]](image-url)
2.4.1 Background noise

Noise is present in any type of environment. Background noise is mostly due to automobile traffic, engines, fan noise, background sound in public places, vibration noise from heavy industries, and aircrafts. In hands-free speech communication, background noises degrade the performance of speech recognition systems which is a severe problem for hearing aid users and also suppress the intelligibility of the speech. Acoustic disturbances arrive from different directions and are said to be background noise containing higher levels of low frequency components when compared to speech signal therefore to extract speech signal spectral based methods are used. In general, speech is characterized by a laplacian distribution whereas background noise is characterized by Gaussian distribution and by considering a certain class of distribution techniques can be developed for extracting speech or background noise. [17]

2.4.2 Acoustic coupling

In hands-free duplex communication, the reflected transmission path between loud speaker and microphone is the echo path. In full duplex communication, the far-end signal which is emitted by the speaker, propagates in the environment and is picked up by the microphones in the same way as other interfering signals[1]. The acoustic echo occurred during the full duplex hands free communication degrade the speech intelligibility, which disturb the user like listening his own speech after some delay. In hands-free communication system the SNR is reduced due to large distance between the microphone and the speaker as it is disturbed by ambient noises.

Hands-free speech enhancement is defined as the ability to improve the discrimination between speech and background noise, reverberation and other types of interferences colliding on microphones [1]. In hands-free communication systems perceptual aspects such as quality and intelligibility are necessary for speech enhancement. The quality and intelligibility are uncorrelated and can be achieved simultaneously. Improvement in intelligibility can be achieved by emphasizing the high frequency content of the noisy speech signal [1]. Therefore, for intelligibility improvement quality should be neglected. In other words quality and intelligibility performance is said to be inversely proportional in the noisy speech signal. Human hearing system has the capability of discrimination of speech in noisy reverberant environments.

Speech enhancement in hand-free mobile communication is possible by spectral subtraction[3] or temporal filtering such as wiener filtering, noise cancellation and microphone methods using different array techniques [3]. Different array techniques are used to handle room reverberations. Hands-free speech communication is generally characterized by reduction in speech naturalness and intelligibility resulting from the corruption of the speech sound field during data capture by microphones, as well as speech distortion generated by data transmission and reproduction [1].
2.5 Applications

Based on frequency selectivity, focused hearing and spatial sound's location, many speech enhancement systems try to substitute and analyses in accordance with the human hearing mechanism [3]. There are numerous applications of hands-free speech enhancement. A few important applications are explained briefly below.

a) Voice control

The advancement in the electrical technology made a huge demand for consumer products, telephones and personal devices and these are rapidly adapting to allow voice control. In order to provide convenience and easy use, a large number of systems is controlled by voice, a few of the applications are lights and heating systems, powering, opening window and curtains and adjusting home entertainment system[1].

The main aim of the voice control and speech recognition systems is to replace hand-controlled functions with voice controls to progress in efficiency and optimized speech automated methods. In the process of speech enhancement in ASR (Automatic Speech Recognition) method it avoids degrading the quality of speech due to the ambient noise and room reverberations. The ASR increases the quality of received speech signal and is based on statistical pattern recognition. The degradation of the signal is calculated based on the amount of similarity between clean speech recognizer and noise speech signal. In order to get improved SNR of the received noisy speech signal and also to increase the speech intelligibility microphone array technique can be used.

b) Hearing aids

Hearing aids is concerned with the remedies for the hearing problems that are caused due to unwanted disturbances. Nearly 25 per cent of the present human population is suffering from hearing impairment by damaging the inner ear hair cells of humans in the process of exposure to loud noise. The exposure to loud noise is mainly in the environments of industries, cooling systems, automobiles, engines and by listening to loud music using headsets. Human hearing system exposing to these types of environments may lead to temporary or permanent hearing loss. The hearing aid system amplifies the received signal, If the signal consists of noise, it is also amplified along with speech signal as hearing impaired people are in capable of distinguishing the speech signals and noise. The main problem for hearing aid is acoustic echo due to the small distance between microphone and speaker. To overcome the above situations, microphone arrays for speech enhancement and an acoustic echo cancellation are used[6].

In this thesis, hearing aids is considered as one of the application in order to make the hearing impaired person more comfortable in hearing the received speech signal and reducing the noise and echo caused due to various environments. During the communication, the speech signal is reverberated in the room from reflection of the wall. Therefore speech signal is corrupted by ambient noise in the environment to the far-end user.
2.6 Microphone array processing

The need for capturing sound and converting it to electric signals came with the first telephones. This is why the first microphones were designed. If we have multiple microphones arranged in a device called a microphone array, we can combine the signals from the microphones such that the microphone array will listen in the direction of the desired sound source, suppressing the sounds coming from other directions. This process is referred to as beam forming. The microphone array is electronically capable of changing the listening direction and following the movements of a human speaker. To do this it employs algorithms for beam steering and sound source localization [1].

A microphone array consists of multiple microphones placed at different spatial locations. Moreover, this enhancement is based purely on knowledge of the source location, and so microphone array techniques are applicable to a wide variety of noise types. Microphone arrays have great potential in practical applications of speech processing, due to their ability to provide both noise robustness and hands-free signal acquisition [1].

Array processing involves the use of multiple sensors to receive or transmit a signal carried by propagating waves. Sensor arrays have application in a diversity of fields, such as sonar, radar, seismology, radio astronomy and tomography. The focus of this thesis is the use of microphone arrays to receive acoustic signals, or more specifically, speech signals. While the use of sensor arrays for speech processing is a relatively new area of research, the fundamental theory is well established as it is common to all sensor arrays, being based on the theory of wave propagation. In general, sensor arrays can be considered as sampled versions of continuous apertures [1].

Directional microphones are best noted for their noise reduction properties in communication systems. Close-talking differential microphones are particularly useful when the noise environment disturbs the ability to communicate without error, such as in public and cellular telephony, aircraft communications, etc. These differential microphones work best when they are spaced within 1 cm, from the lips of the talker where the sound field has a large gradient [6].

Directional microphones have advantages over Omni directional microphones in noisy environments. The discrimination against both solid-borne and airborne noises, especially airborne noise that originates in the far field, are important in environments where communications must be clearly understood at all times. Second-order differential microphones using a single element piezoelectric transducer and tubes to sample the sound field have been suggested for use in the very noisy environment of aircraft communication systems [6].

Foil-electrets differential microphones of first order have been successfully used in many environments where both stationary and non-stationary noises are dominant. In all of the above cases, the sensors are designed to be positioned as close as possible to the lips of the talker. In environments where noise conditions are not as severe, but where there exist some possibilities of speech interference from far-field noise or speech sources, differential microphones may still be
appropriate. However, users in this kind of environment frequently relax the position requirement of the differential microphone and sometimes place the sensor as much as 7.5 cm from the lips. However differential microphones at more than 2 cm distant from the lips suffer from inordinate high frequency gain to the speech signal [6].

It is important to note that the microphone array output might still contain some residual noise and reverberation. The audio quality can be further improved using speech enhancement techniques such as noise suppression and de-reverberation [6].

In this thesis, a microphone array consisting of two microphones is used in near-talk mode.

A typical microphone can be designed to receive signals from any desired direction and with any kind of sensitivity patterns. Of these, the most commonly used microphones are either Omnidirectional or Unidirectional. Omnidirectional microphone is capable of receiving signals from all directions. On the other hand, unidirectional or simply directional microphone is intended to receive signals from a particular direction while blocking the signals from all other directions. Figure 3 shows the ideal cases of these two microphone setups. Left hand side of Figure 3 shows the pattern of an ideal omnidirectional microphone showing equal sensitivity to signals from all directions. Right hand side of Figure 3 consists of a directional microphone sensitive to signal from a particular direction (0° in this case) and zero sensitivity to signals from all other directions. This makes it capable of listening to signals in a particular direction.

![Figure 3 Sensitivity patterns of ideal omnidirectional and directional microphones](image)

Noise reduction techniques can be classified depending upon the number of microphones used as follows:

- Single microphone noise reduction techniques
- Multi microphones techniques.
2.7 Single Channel Speech Enhancement:

In most of the situations such as mobile communications, only one microphone is available. Now, a typical unidirectional microphone can be utilized for the single channel speech enhancement technique as its capability to block unnecessary signals from different directions seems promising. Nevertheless, it lacks the basic knowledge or information about the source or the noise signals. This makes it vulnerable to noise signals coming from the direction of the source. In such a case, noise reduction techniques rely on assumptions that are made related to speech and noise signals or need to exploit aspects of speech perception, speech production or speech model.

A common assumption is that the noise is additive and slowly varying, meaning that the noise process is stationary and uncorrelated with speech. So the noise characteristics estimated in the absence of speech can be used subsequently in the presence of speech. But in reality, this assumption is not valid or only partially valid and the system will either have less noise reduction or introduce more speech distortion[3].

Typical algorithms used for single channel speech enhancement are spectral subtraction, wiener filtering and minimum mean square error estimator. Spectral subtraction has the problem of introduction of musical noise at low SNR levels. Also, estimation of noise in these methods makes use of a voice activity detector (VAD) which extracts the noise estimate from the speech-free intervals. Even with limitations outline above, single channel noise reduction has attracted a tremendous amount of research attention because of its wide range of applications and relatively less cost.

Figure.4 Structure of microphone array system
Speech Enhancement using beamformers has been one of the most efficient ways to estimate the received speech in the presence of various interfering signals and noise. A beamformer can be described as an array of microphones arranged in a particular geometry (linear, in general) which will be directed towards the original speech.

2.8 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. 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 is 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 is an integer delay and for fractional delays we use the Fractional Delay (FD) filters [4].

The microphone arrays that are implemented for multi-channel speech enhancement should have the property of time aligning the incident signal. This means that the array setup should be able to delay the signal received by the microphones. When we use sampled data as input, it is most probable that these delays are not quite adequate to reconstruct the original signal. This is because the delays will not equate to multiples of whole sample periods. A fractional delay filter (FD) is one which implements a delay that is not an integer multiple of the sampling period. FD filters are intended for bandlimited interpolation in general. Bandlimited interpolation is a technique for evaluating a signal sample at any arbitrary point in time, even if it is located in between two sampling points [4].

2.8.1 Ideal Fractional delay:

According to Shannon’s sampling theorem [1], a sinc interpolator can be used to exactly evaluate a signal value at any point in time, as long as it is bandlimited to an upper frequency of \( F_s/2 \). Hence, the delay system must be made bandlimited by using an ideal low pass filter in the frequency domain while it merely shifts the impulse response in the time domain. Thus, the impulse response of an ideal FD filter is a shifted and sampled sinc function, that is:

\[
y(D) = \sum_{n=-\infty}^{\infty} y(n) \text{sinc}(n - D)
\]

(1)

Where \( D \) is the delay and the delayed function \( \text{sinc}(n - D) \) is referred to as an ideal fractional-delay interpolator:

\[
h_D(n) = \text{sinc}(n - D)
\]

(2)

The entire process of fractional delay using resampling is described by the following equation:
\[(x_k) \xrightarrow{h_D} (x_{k-D}) = (x_k) \otimes sinc(k - D) \quad (3)\]

Figure 5 shows the ideal impulse response when \(D = 0.0\) and \(D = 0.7\) samples. The first case is of finite length as only one sample at 0 is present. In the latter case, the impulse response is of infinite length and hence corresponds to a non-causal filter which cannot be made causal by a finite shift in time. Also, the filter is not stable as the impulse response is not absolutely summable. The ideal filter is thus non-realizable. Hence, real fractional delay systems are at best approximations of the ideal impulse response[4]. There are many approximations to FD filters in which both FIR and IIR implementations exist.

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 \(d \cos \theta\) after reaching. The delay in time is given as \(d \cos \theta / v\) seconds assuming that speed of sound is \(v\) meters/seconds. So we have to add time delay of \(d \cos \theta / v\) to signal received by and this time delay can be expressed in terms of samples is given as \((d \cos \theta / v) \times f_s\), where \(f_s\) is sample frequency.
2.8.2 FIR approximation of fractional delay:

The two common methods used for FIR approximation of FD are the windowed sinc function and the Lagrange’s interpolation. [4]

Windowed sinc function uses an asymmetric window function like the hamming window or a binomial window with a fractional offset. [4] It is quite easy to implement but the approximation is quite weak and the coefficients must be scaled to obtain the best approximation.
Lagrange’s interpolation belongs to a class of filters called the maximally flat filters since they have a constant (flat) magnitude around a particular frequency of interest \[4\]. The response of Lagrange interpolator is made identical to that of ideal interpolator at zero frequency. A closed form representation of Lagrange FIR filter coefficients are given by the following equation \[4\]:

\[
h(n) = \prod_{k=0, k \neq n}^{N} \frac{D-k}{n-k} \quad \text{for } n = 0, 1, 2, \ldots, N
\]

where \(N\) is order of filter and \(D\) is positive delay. The ease of computing filter taps is an important feature of Lagrange interpolators. This is a useful technique for narrow-band cases but is not competent in case of attaining minimal error power in the approximation band since it merely minimizes the error in the vicinity of \(\omega = 0\)[4].

### 2.8.3 IIR approximation of fractional delay:

All pass IIR filters are particularly suitable for FD approximation since their magnitude response is exactly unity at all frequencies. Hence the IIR approximation always deals with allpass filters and their design. Unlike FIR approximations, allpass filters can be implemented using lower order which translates into computational speed[4].

The transfer function of a general discrete time allpass filter is given by the form:

\[
H(z) = \frac{a_n z^{-1} a_{n-1} + \cdots + z^{-(N-1)} a_1 + z^{-N}}{1 + z^{-1} a_1 + \cdots + z^{-(N-1)} a_{n-1} + z^{-N} a_n}
\]

where \(N\) is the order of the filter with the real valued coefficients \(a_k\). The design of allpass fractional delay filters is usually based on solving a set of equations, such as the least squares method, or an iterative optimization algorithm, such as pseudo-equiripple design techniques[4]. 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 real-time applications[4]. The largest allpass filter order that is possible with current design programs is about 20 or less, depending on specifications[4].

### 2.8.4 Thiran allpass filter:

Thiran allpass filter[4] is the only one maximally flat FD approximation method that can be implemented with closed-loop formulas. 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 This is a simplest design with the closed loop coefficients given by:

\[
a_k = (-1)^k \binom{N}{k} \prod_{n=0}^{N} \frac{N-n-D}{N-k-n-D}
\]

where \(D\) is the group delay and \(k = 1, 2, 3, \ldots, N\). The useful range of delay \(D\) is different for allpass filters from that of FIR filters. The error curves of allpass filters are asymmetric and thus their stability must be taken into account. The Thiran interpolator is stable only for the values \(D > N-1\).
which implies that there is a lower limit for the delay to be approximated. In general, the optimal range of \( D \) is about \( N-0.5 \) to \( N+0.5 \) [4]. The error does not decrease even when the order is increased. As the coefficient values of the Thiran allpass filter typically decay fast with index \( k \), it is convenient to neglect extremely small coefficient values as their contribution cannot be significant on the properties of the filter. Hence the design neglect the smallest values by truncating the coefficient vector of the filter. For practical purposes, a prototype filter order is chosen that corresponds to the order of the original all pass filter before truncation. The truncation facilitates for the design of higher order FD filters and overcomes the problem of narrow approximation band [4].

2.9 Direction of Arrival

Microphone arrays have been implemented in large number of applications which includes speech recognition, teleconferencing, and position location of source for laptops etc. Direction of arrival (DOA) of speech signals using a set of microphone arrays has many practical applications in everyday activities. The set of microphones can be used automatically to attenuate the speech signal coming from other than concentrated direction. DOA estimation process uses the phase information present in the signals received by the microphone array. So the speech signals with desired direction are processed and other signals are attenuated based on the DOA estimation values [6].

Depending on the values of arrival angles we choose a scalar parameter to attenuate the signals from other direction by using the Gaussian distribution function. Since the microphones are spatially separated, the speech signal arrives at each microphone at different instance which cause difference in time of arrival. Using this time delay and known geometry we can estimate the Direction of arrival in most cases. Figure 6 shows the far field case for which Time Difference of Arrival (TDOA) is denoted as \( \tau \) and Direction of arrival is the angle between the plane of incident signal and the plane perpendicular to microphone array plane. Many applications like vehicle location, vehicle monitoring systems, sensor networks, distributed robotics use DOA techniques.

There has been extensive research resulting in number of proposals for development of direction of arrival estimations [6]. The estimation methods are mainly classified in to two categories i.e. spectral based and parametric approaches. In this thesis cross correlation and time of arrival of signal at microphones are used to estimate the direction of arrival.
2.10 Noise suppression using spectral subtraction

Spectral subtraction is a speech enhancement method [5]. The reduction of noise is accomplished by subtracting an estimate of the power spectrum of the noise from the power spectrum of the input signal while maintaining the noisy speech signal phase information. This can also be viewed as a filtering operation. The filter is shaped so that frequencies with high SNR may pass, but frequencies with low SNR are proportionally reduced. This results in a speech enhanced signal with a higher overall SNR. Spectral subtraction can be improved from conventional spectral subtraction method by exchanging the non-causal circular convolution with a purely linear causal filtering [5].

The filter used in spectral subtraction is applied in the frequency-domain and changes over time for each segment of the microphone signal. The signal segment is transformed to the frequency-domain by a Fast Fourier Transform (FFT) resulting in a frequency-domain segment. Conventionally the filter is of the same order as the signal segment and the
frequency-domain segment, yielding a circular convolution. Since the segments and the filter are of the same length, causal properties cannot be imposed on the filter. By using a shorter filter length, the filter can be estimated with less variance. A shorter filter length also has the benefit of providing a purely linear filtering possibility in the frequency-domain when the added length of the signal segment and the filter is shorter than the FFT length. The purely linear filtering does not introduce sharp edges between segments as with conventional spectral subtraction. These discontinuities can sometimes be heard as click sounds at segment boundaries. In conventional spectral subtraction these click sounds can be reduced by introducing a time-window, which gradually mutes the ends of a signal segment, in combination with an overlapping between segments. The introduced overlap generally results in a delay of the processed signal of approximately 10 milliseconds [5].

One important consideration in noise suppression applications is the characteristics of the noise estimator, bias and variance, mainly. Another consideration is on how to observe the noise signal. One of the early methods assumed the first second to be noise only and estimated the noise from this signal [9]. This approach had the disadvantage, that was the signal in fact speech, the method would erroneously remove speech components. Secondly, was the first second in fact noise, the assumption meant that the noise had to be stationary for the remainder period of operation. Clearly this is a very unrealistic assumption in most applications. As an alternative, it has often been proposed to use a voice activity detector (VAD) in order to detect speech-dominated and noise-dominated time frames. The noise-dominated time frames can be used to estimate the (possibly non-stationary) background noise, however, the reader would likely have realized, that this merely shifts the problem of noise estimation to developing reliable VADs which successfully detects speech-dominated time frames. This has led to extensive research in good VADs. For noise estimation, alternative more exotic noise estimation procedures have recently reported promising results in non-stationary environments. Most notable the method by Rainer Martin has been accepted as a state-of-the-art method, which works without the use of an explicit VAD [7].

This Theses makes use of method proposed in [7] which is used to noise estimation in spectral subtraction.
CHAPTER – 3

IMPLEMENTATION

3.1 Direction of arrival algorithm

Assume that the source is located opposite of both the microphones. The DOA algorithm has to detect the direction of arrival of desired speech signal. The origin is considered at centre between two microphones $M_1$ and $M_2$. The distance between the elements is $d$ meters. An orthogonal line to the microphone axis passing through origin is considered as $O$ and $S$ is the point of source, line joining the origin and point of source is represented as $OS$. Direction of arrival $\theta$ always refer to angle between the orthogonal line and $OS$ as shown in Figure 7

Figure 7 shows the angles of extreme positions that are considered and source is considered to lie between $+90^0$ to $-90^0$. A DOA estimation algorithm uses the correlation of amplitude and phase information from the microphone array in order to estimate the arrival direction $\theta$. Only single source is considered in estimating the DOA. Based on the phase difference the DOA algorithm is developed in the subsequent paragraph[6].

Let the desired source signal is denoted by $s(n)$ and the output of microphone signals can be expressed as
\[ x_1(n) = s_1(n) + v_1(n) \]  
\[ x_2(n) = s_2(n) + v_2(n) \]

Where \( s_1(n) \) and \( s_2(n) \) are the delayed versions of \( s(n) \), \( v_1(n) \) and \( v_2(n) \) represents noise.

As shown in the Figure 8, the signals \( x_1(n) \) and \( x_2(n) \) are passed through WOLA analysis stage independently. Where the signals are decomposed into sub bands and these Fourier transformed signal can be represented as

\[ X_1^k(w_n) = S_1^k(w_n) + V_1^k(w_n) \]  
\[ X_2^k(w_n) = S_2^k(w_n) + V_2^k(w_n) \]

Figure 8: Block diagram showing Estimation of DOA with two Microphones [6]
Where $\omega_n$ represents the angular frequency. $X(\omega)$, $S(\omega)$, $V(\omega)$ are the Fourier transforms of $x(n)$, $s(n)$ and $v(n)$ respectively. The phase difference between the two microphone sub bands can be calculated with cross correlation of both the signals. A correlation between the outputs of each microphone has information about direction of arrival of Speech signal. The phase difference is given as

$$\Phi_k^\omega(n) = \angle X_1^k(\omega_n) - \angle X_2^k(\omega_n) \tag{12}$$

and is limited to be in range $[-\pi, \pi]$. $\angle X_1^k(\omega_n)$ and $\angle X_2^k(\omega_n)$ represents phase of $X_1^k(\omega_n)$ and $X_2^k(\omega_n)$. Next step is phase un wrapping where the radian phases are un wrapped by changing absolute jumps greater than or equal to $\pi$ to their $2\pi$ complement. The line fitting is applied using least square sense. The Slope is obtained from the line fitted data and this slope is nothing but the time difference of arrival (TDOA). The direction of arrival can be calculated from the time delay of arrival (TDOA) by using the following equation [6]

$$\tau = dsin\theta / c \tag{13}$$

Where $d$ is distance between the elements, $\theta$ is direction of arrival and $c$ is velocity of sound.

### 3.2 Wiener Beamformer

Let the input signal to each microphone be $x_m(n)$ which is considered as the mixture of desired speech signal $S_m(n)$ and noise $v_m(n)$. The microphone array consists of two microphones and space between the microphones is $d$ meters. Let the signal received by the first microphone is denoted as $x_1(n)$ and second microphone is $x_2(n)$ where $n$ represent sample index. Each microphone signal is divided in to $K$ (in our case $K=65$) number of sub bands by WOLA analysis Stage and beamforming is performed independently on each of these sub bands. The input vector is given as

$$x_m(n) = [x_{m1}^k(n)x_{m2}^k(n)x_{m3}^k(n)x_{m4}^k(n)...x_{mk}^k(n)]^T \tag{14}$$

Where $K$ corresponds to the sub band index and $x_m(n)$ represent the microphone signal and in our case $m=2$ and $x_m^k(n)$ defined according to equation 15

$$x_m^k(n) = [x_m^k(0)x_m^k(1)x_m^k(2)x_m^k(3)...x_m^k(N)]^T \text{ for } k = 0,1,2,...,k-1 \tag{15}$$

Here for given input signal the maximum sample index is assumed as $N$. The obtained sub band input signals are filtered according to a wiener beamforming topology. Here $T$ denotes transposition, * denotes complex conjugate, $[\cdot]^H$ denote the Hermitian transposition. The optimal weights which minimizes the mean square error between the output and the reference signal is computed to each sub band.
In this thesis, Filter-and-sum beamforming technique implemented with wiener filter in time domain and implementation of sub band beamforming in frequency domain, shown in the Figure 9. The filter-and-sum beamforming is one of the simplest beamforming techniques but still gives a very good performance [18]. It is based on the fact that applying different phase weights to the input channels the main lobe of the directivity pattern can be steered to a desired location, where the acoustic input comes from. It differs from the simpler delay-and-sum beamformer in that an independent weight is applied to each of the channels before summing them as shown in the Figure 10.
Noise reduction can be done using wiener filter. The unwanted disturbances and other interferences are considered as noise. Using wiener filter into beamforming technique is one of the good solution to increase the speech intelligibility. The linear microphone array is used in the wiener filter in this case is referred to be a multichannel wiener filter.

For the input vector $x(n)$ at discrete-time instant $n$, containing mainly frequency components around the centre frequency $\Omega$, the spatial correlation matrix is given by

$$R_{xx}(n) = E [x(n)x^H(n)]$$  \hspace{1cm} (16)

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

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

$$R_{xx}(n) = R_{ss}(n) + R_{ii}(n) + R_{nn}(n)$$  \hspace{1cm} (17)

Where $R_{ss}(n)$ is the source correlation matrix,

$R_{ii}(n)$ is the interference correlation matrix and

$R_{nn}(n)$ is the noise correlation matrix defined by the following equations.

$$R_{ss}(n) = E [x_s(n)x_s^H(n)]$$  \hspace{1cm} (18)

$$R_{ii}(t) = E [x_i(t)x_i^H(t)]$$  \hspace{1cm} (19)
\[ R_{nn}(n) = E[ x_n(n)x_n^H(n)] \]  

Wiener solution for the time domain beamforming

The optimal filter weight vector

\[ W_{opt} = [R_{xx}]^{-1}r_{sx} \]  

Where the array weight vector, \( W_{opt} \) is arranged as

\[ W_{opt} = [w_1, w_2, ..., w_N]^T \]  

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

\[ r_{sx} = E[x_s(n)s^H(n)] \]  

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

\[ y(t) = W_{opt}^H X(n) \]  

Where \( X(n) \) corresponds to \( x_1(n) + x_2(n) \).

3.3 SPECTRAL SUBTRACTION

The spectral subtraction is the one of the most common single microphone speech enhancement techniques for additive noise reduction, low complexity and owing to its simplicity. Pioneering work by Berouti and Boll [5], using power magnitude spectral subtraction and spectral subtraction, respectively, led to a generalized spectral subtraction frame work as the empirical parameters can adjust in order to finite-time the algorithm. Spectral subtraction has been capable of removing stationary background noise, however, at the expense of speech distortion.

3.3.1 Spectral Subtraction algorithm basic principle

Spectral subtraction method for restoration of the power or the magnitude spectrum of a signal observed in additive noise through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. These are the systems form a category of algorithms that operate in the time-frequency domain. The noise spectrum is estimated, and updated, from the periods when the desired signal is absent and only the noise is present. The assumption being that noise is stationary or a slowly varying process, and that
the noise spectrum does not change significantly between the updating periods. For
restoration of the time-domain signal, an estimate of the instantaneous magnitude spectrum
is combined with the phase of the noisy signal, and then transformed via a inverse discrete
Fourier transform back to the time domain. The phase of the noisy signal is not modified,
as not only is it hard to get an estimation of the phase as compared to the magnitude
spectrum, but it has also been shown that from perceptual point of view the phase does not
carry so much useful information [10].

Let \( y(n) \) be the noise corrupted input speech signal, composed of the clean speech signal
\( x(n) \) and the additive noise signal \( d(n) \). i.e.

\[
y(n) = x(n) + d(n)
\]

Many speech enhancement algorithms operate in the Discrete Fourier Transform (DFT)
domain. In fourier domain, we can write

\[
Y(\omega) = X(\omega) + D(\omega)
\]

\( Y(\omega) \) can be expressed in terms of magnitude and phase as

\[
Y(\omega) = |Y(\omega)| e^{j\phi_y}
\]

where \( |Y(\omega)| \) is the magnitude spectrum and \( \phi_y \) is the phase spectrum of the corrupted
noisy speech signal. The noise spectrum in terms of magnitude and phase is

\[
D(\omega) = |D(\omega)| e^{j\phi_y}
\]

The magnitude of the noise spectrum \( |D(\omega)| \) is unknown but can be replaced by its average
value or estimated noise \( |D_e(\omega)| \) computed during non-speech activity, that is during
speech pauses. The noise phase is replaced by the noisy speech phase \( \phi_y \) that does not
affect speech ineligibility. We can estimate the clean speech signal simply by subtracting
noise spectrum from noisy speech spectrum in equation form

\[
X_e(\omega) = (|Y(\omega)| - |D_e(\omega)|) e^{-j\phi_y}
\]

where \( X_e(\omega) \) is estimated clean speech signal. Many spectral subtractive algorithm
variations have been proposed such as magnitude spectral subtraction, power spectral
subtraction and autocorrelation subtraction [6]. The estimation of clean speech for
magnitude signal spectrum is

\[
|X_e(\omega)| = (|Y(\omega)| - |D_e(\omega)|)
\]

similarly for power spectrum subtraction is
\[ |X_e(\omega)|^2 = |Y(\omega)|^2 - |D_e(\omega)|^2 \]  

(31)

The enhanced speech signal is finally obtained by computing the inverse Fourier transform of the estimated clean speech \(|X_e(\omega)|\) for magnitude spectrum subtractions and \(|X_e(\omega)|^2\) for power spectrum subtraction, using the phase of the noisy speech signal. The more general version of the spectral subtraction algorithms is

\[ |X_e(\omega)|^p = |Y(\omega)|^p - |D_e(\omega)|^p \]  

(32)

Where \(p\) is the power exponent, the general form of the spectral subtraction, when \(p=1\) yielding the magnitude spectral subtraction algorithm and \(p=2\) yielding the power spectral subtraction algorithm. The general form of the spectral subtraction algorithm is shown in Figure 11.

The main disadvantage using conventional spectral subtraction algorithms is sometimes audible “musical tones” which can disturb not only the listener but also speech coding algorithms. The musical tones are mainly due to high variance in the spectrum estimates [10]. To solve this problem spectral smoothing has been suggested, resulting in reduced variance and resolution [10]. Other suggested improvements of conventional spectral subtraction are to use the Bartlett method on fairly long sequences of samples to reduce the variance of the spectrum, or to hide the musical tones phenomenon using the masking properties of the auditory system [10]. A more detailed description of some of the different methods along with some other prominent short-time amplitude estimation based methods is presented in the sections below.
3.3.2 Spectral subtraction approach used in this thesis

This thesis removes non-stationary noise by using spectral subtraction algorithm. The spectral subtraction algorithm in turn uses weighted overlap add (WOLA) filter bank to carry out entire operation in time-frequency domain and minimum statistics to estimate the noise spectrum.

Modern speech processing methods are usually implemented in the time-frequency domain. The time-frequency domain means that a signal is represented not only as a function of time, which is the “normal” representation, but it is also a function of frequency. The time-frequency domain representation can be achieved by filtering the input time signal by a bank of band-pass filters, where the band-pass filters have very little mutual overlap in frequency, see example in Figure.12 Methods that transform a signal from the time- to the time-frequency domain are usually referred to as filter banks. The transformed filter bank signals are denoted sub-band signals. Signal processing methods are generally more efficient using filter banks, since the processing load can be reduced owing to down sampling. Filter bank processing can also be considered as a divide-and-conquer approach within signal processing, since larger problems are sub-divided into many smaller problems.

Figure .12: A bank of eight band pass filters $h_k(n)$, with Fourier transforms $k(f)=\mathcal{F}(h_k(n))$, comprise filter bank [11]
3.4 Filter banks

To understand the concept of filter banks, assume a sampled (and quantized) input signal \( x(n_t) \), such as the signal from the Analog-to-Digital Converter (ADC) of a sound codec connected to a DSP. The signal is uniformly sampled using the frequency \( F_s \) Hz, and we shall assume, that filter bank uses a set of K band pass filters, where the impulse response of each band pass filter is \( g_k(n_t) \), for \( k = 0 \ldots K - 1 \), where index \( k = 0 \) corresponds to DC frequency and where index \( k = K/2 \) corresponds to the Nyquist frequency [19]. The filter bank output signals are the convolution of the input signal with each of the band pass filters, i.e. \( x_k(n_t) = (g_k \ast x)(n_t) \), where \( \ast \) denotes linear convolution. Each filter bank output signal \( x_k(n_t) \) is referred to as a sub-band signal, since it describes one sub-band of the signal. It shall be clear that the signal set \( \{x_1(n_t), \ldots, x_K(n_t)\} \) is a time-frequency signal since the index \( n_t \) represents time and the index \( k \) represents the sub-band index which corresponds to a unique frequency band.

The sub band signals \( x_k(n_t) \), from the analysis, are processed by sub band processors \( g_k \) to yield sub band output signals \( y_k(n_t) = g_k\{x_k(n_t)\} \), to be reconstructed in the synthesis. Note that if the sub band processors are filters, the sub band output signals are computed using the ordinary convolution \( y_k(n_t) = (g_k \ast x_k)(n_t) \).

3.4.1 Analysis – synthesis

The part of the filter bank that transforms a time-signal into a corresponding time-frequency representation is referred to as an analysis filter bank. There exists a corresponding inverse transform that transforms a time-frequency signal, e.g. \( x_k(n_t) \), into a time-signal, e.g. \( x(n_t) \). The part of the filter bank that implements this inverse transform (which is also known as reconstruction) is referred to as a synthesis filter bank.

![Figure 13: Analysis-synthesis filter bank](image-url)
3.4.2 Decimation

It may be noted that, in the definition above, we have a factor \( K \) over-representation of the signal since the sub-band signals have the same sampling frequency but they, at the same time, represent only \( 1/K \) of the frequencies of the original signal. Hence, it is common practice to decimate the sub-band signals by a certain factor \( D \), i.e., the analysis filter bank output is decimated as \( x_k(n_tD) = (g_k * x)(n_tD) \). The factor \( D \) is referred to as the decimation rate. The decimation rate is related to the number of sub-bands \( K \) by the over-sampling ratio \( O = K/D \). The over-sampling ratio describes the redundancy in the sub-band data in relation to the original input time-signal data. If the over-sampling is equal to one, which means that \( K=D \), then we say the filter bank is critically sampled or critically decimated. Normally, we allow a certain redundancy in the filter bank, e.g., by a factor of \( O = 2 \) which we would refer to as a factor two over-sampled filter bank. The decimation in the analysis filter bank is naturally compensated by a corresponding interpolation in the synthesis filter bank. Note that after the interpolation, it is required to filter the interpolated signal in order to suppress aliasing terms due to the interpolation process.

3.4.3 WOLA filter bank

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 \( w(n_t) \), define the WOLA filter bank, namely \( L \) the length of the analysis window, \( D \) the decimation rate (block rate), \( K \) the number of sub-bands, and \( D_f \) the synthesis window decimation rate. The analysis stage, see Figure 14, accepts a block of \( D \) new input data samples and each new block is fed into an input FIFO buffer \( u(n_t) \), 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 \( t_1(n_t) = u(n_t)w(n_t) \), of length \( L \) samples. The temporary buffer is time-folded into another temporary vector \( t_2(n_t) \), of length \( K \) samples. The time-folding means that the elements of \( t_1(n) \) are modulo-\( K \)-added to \( t_2(n_t) \), according to

\[
    t_2(n_t) = \sum_{m=0}^{L-1} t_1(n_t + mK)
\]

The temporary buffer \( t_2(n_t) \) 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 \( t_2(n_t) \) is swapped place with the lower half of \( t_2(n_t) \). The circularly shifted buffer \( t_2(n_t) \) is then fed into a \( K \)-sized FFT to compute the sub-band signals \( x_k(n_t) \). It should be noted that the synthesis stage of the WOLA filter bank implements the actual weighted overlap-add procedure. The synthesis stage, starts by applying a size-\( K \) IFFT to the processed sub-band signals \( y_k(n_t) \). 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 \( t_3(n_t) \), of
size $K$ samples. The buffer $t_3(n_t)$ is then stacked, by repetition, in the buffer $t_4(n_t)$ of length $L/DF$. The buffer $t_4(n_t)$ is weighted by a synthesis window function $z(n_t)$ of size $L/DF$, defined as $z(n_t) = w(n_t D_f)$, i.e., a factor $DF$ decimated analysis window function. The weighted data is summed with the data in the output FIFO, $t_5(n_t)$ of length $L/DF$, and the output FIFO data is over-written with the summation result, i.e., $t_5(n_t) \leftarrow t_5(n_t) + z(n_t) t_4(n_t)$. The output FIFO is then shifted left by $D$ samples, i.e., zeros are filled from the FIFO’s rear, and the out-shifted data is the actual output data block, $y(n_t D)$.

The WOLA filter bank is popular due to its simplicity and since many DSP technologies have built in support for circular buffers and FIFOs, which of course helps when implementing the WOLA filter bank. Also, DSPs have usually direct support for FFT/IFFT which renders furthermore efficient implementations [12].

3.5 MINIMUM STATISTICS BASED NOISE ESTIMATION

In noise reduction of speech signals the objective is to estimate the speech component from an observed signal containing degraded speech. The dual problem of estimating the noise component is often inherently part of the noise reduction formulation. Some sort of noise estimation is required in order to apply the method under consideration. For spectral subtraction, it is assumed that a frequency-dependent noise estimate is known, but spectral subtraction formulates no specific method on how to obtain this estimate. The problem is, so to say, just shifted from speech estimation to noise estimation.

Figure.14: an example of noisy speech power spectrum and local minimum noise estimation [10]
The noise spectral density estimation based on minimum statistics, proposed by Martin [11], rests on two basic assumptions: (1) the speech and interfering noise are statistically independent, (2) the power of the noisy speech signal often decays to the power level of the interfering noise. The latter is true for speech signals, as they are highly non-stationary and take on varying stochastic characteristics over time. Speech signals are known not to be continuously present and often decay to a noise floor. This is what is exploited in this method. Minimum statistics noise estimation utilizes the premise that the spectrum of the noise signal generally exhibits lower magnitudes than the underlying speech signal. As a result, noise estimates can be derived by tracking the minimum spectral magnitudes over finite time windows for each frequency [11]. In order to reduce the effects of outlying spectral values, the minimum statistics algorithm typically smoothes the signal spectrum prior to calculating the noise estimate.

Figure 14 shows two main drawbacks of Minimum Statics (a) the estimated noise level is consistently lower than the true noise level and (b) the algorithm fails to respond rapidly to increases in the noise spectrum. To counteract (a) in spectral subtractive speech enhancement, the subtraction values are typically greater than 1. The slow response to increase in the noise levels is attributed to use of infinite time windows [11].

3.6 Minimum Statistics Based Noise Estimation Applied to Spectral Subtraction

In this thesis, minimum statistics is used for estimating noise estimation which essentially eliminates the need for explicit speech pause detection, without a substantial increase in computational complexity. While the conventional approach to spectral subtraction employs a speech activity detector, we here use the minimum of the sub-band noise power within a finite window to estimate the noise floor. The algorithm is based on the observation that a short time sub-band power estimate of a noisy speech signal exhibits distinct peaks and valleys as discussed in previous section.
The peaks correspond to speech activity the valleys of the smoothed noise estimate can be used to obtain an estimate of sub-band noise power as shown in the Figure 15. To obtain reliable noise power estimates, the data window for the minimum search must be large enough to bridge any peak of speech activity.

A block diagram of the basic spectral subtraction method is shown in Figure 16. The algorithm appropriately modifies the short time spectral magnitude of the disturbed speech signal such that the synthesized signal is perceptually as close as possible to the undisturbed speech signal. The optimal weighting of spectral magnitudes is computed using a noise power estimate and a subtraction rule [11].
3.7 Spectral Analysis/Synthesis

We assume that the band limited and sampled disturbed signal \( x(i) \) is a sum of a zero mean speech signal \( s(i) \) and a zero mean noise signal \( n_s(i) \), \( x(i) = s(i) + n_s(i) \), where \( i \) denotes the time index as analysis is carried out by block by block. So we consider a block of signal i.e. \( x(i) \). We further assume that \( s(i) \) and \( n_s(i) \), are statistically independent, hence \( E[x^2(i)] = E[s^2(i)] + E[n_s^2(i)] \) [11]. Spectral processing is based on a DFT filter bank with \( W_{DFT} \) sub-bands and with decimation/interpolation ratio \( R \). The phase of the disturbed signal is not modified. We denote the data window by \( h(i) \) and the \( k \)th sub-band signal of \( x(i) \) [11] as

\[
X(\lambda, k) = \sum_{\mu=0}^{W_{DFT}^{-1}} (x(\lambda R + \mu) h(\mu) e^{j \omega}) - j \frac{2\pi\mu k}{W_{DFT}} \tag{34}
\]

\( \lambda \) and \( k \) refer to the decimated time index and the DFT frequency bins \( \Omega_k = \frac{2\pi k}{W_{DFT}}, k = 0, 1, ... W_{DFT} - 1 \), respectively. Typically we use a DFT length of \( W_{DFT} = 256 \) and decimation ratio \( R = 64 \). The improved sub-band signals are converted back to the time domain using an inverse DFT. The synthesized improved speech signal is denoted by \( y(i) \), the corresponding spectral magnitude by \( |Y(\lambda, k)| \).

3.8 Subtraction Rule

Let \( P_n(\lambda, k) \) and \( \overline{X(\lambda, k)}^2 \) denote the estimated sub-band noise power and short time signal power, respectively. To obtain the short time signal power subsequent magnitude squared input spectra are smoothed with a first order recursive network (\( \gamma <= 0.9 \))

\[
\overline{X(\lambda, k)}^2 = \gamma |\overline{X(\lambda - 1, k)}|^2 + (1 - \gamma) |X(\lambda, k)|^2 \tag{35}
\]

Following the proposal of Berouti [11], we subtract spectral magnitudes with an over subtraction factor \( O_{sub}(\lambda, k) \) [Subtraction factor is a parameter that controls the amount of noise subtracted from the noisy signal. For full noise subtraction, subtraction factor=1 and for over-subtraction Subtraction factor>1.] and a limitation of the maximum subtraction by a spectral floor constant \( s_f (0.01 <= sub_f <= 0.05) \).

\[
|Y(\lambda, k)| = \begin{cases} 
\sqrt{s_f P_n(\lambda, k)} & \text{if } |X(\lambda, k)| Q(\lambda, k) \leq \sqrt{s_f P_n(\lambda, k)} \\
|X(\lambda, k)| Q(\lambda, k) & \text{else where} 
\end{cases}
\]

where \( Q(\lambda, k) = \left( 1 - \sqrt{O_{sub}(\lambda, k) \frac{P_n(\lambda, k)}{|X(\lambda, k)|^2}} \right) \tag{36} \)
While a large over subtraction factor \(O_{\text{sub}}(\lambda, k)\) essentially eliminates residual spectral peaks (‘musical noise’) it also affects speech quality such that some of the low energy phonemes are suppressed. To limit this undesirable effect, the over subtraction factor is computed as a function of the sub-band signal-to-noise ratio \(\text{SNR}_x(\lambda, k)\) and the frequency bin \(k\), i.e. \(O_{\text{sub}}(\lambda, k) = f(\lambda, k, \text{SNR}_x(\lambda, k))\). In general we use less over subtraction for high SNR conditions and for high frequencies than for low SNR conditions and for low frequencies.

### 3.9 Sub-band Noise Power and SNR Estimation

Let us first compute the short time sub-band signal power \(P_x(\lambda, k)\) using recursively smoothed periodograms [11]. The update recursion is given by equation (37). The smoothing constant is typically set to values between \(\alpha = 0.9 \ldots 0.95\).

\[
P_x(\lambda, k) = \alpha P_n(\lambda - 1, k) + (1 - \alpha)|X(\lambda, k)|^2
\]

(37)

The noise power estimate \(P_n(\lambda, k)\) is obtained as a weighted minimum of the short time power estimate \(P_x(\lambda, k)\) within a window of \(D\) sub-band power samples, i.e.

\[
P_n(\lambda, k) = O_{\text{min}} P_{\text{min}}(\lambda, k)
\]

(38)

\(P_{\text{min}}(\lambda, k)\) is the estimated minimum power and \(O_{\text{min}}\) is a factor to compensate the bias of the minimum estimate. For reasons of computational complexity and delay the data window of length \(D\) is decomposed into \(W\) windows of length \(M\) such that \(M W = D\). For a sampling rate of \(f_s = 8\) kHz and a decimation ratio \(R=64\) typical window parameters are \(M=25\) and \(W=4\), thus \(D=100\) corresponding to a time window of \(\frac{(D-1) R + W_{\text{DFT}}}{f_s} = 0.824s\).

To determine the minimum of \(M\) consecutive sub-band power samples at time, we initialize a variable \(P_{\text{Max}}(\lambda = \lambda_1, k)\) to the first of the \(M\) samples \(P_{\text{Max}}(\lambda = \lambda_1, k) = P_x(\lambda = \lambda_1, k)\). The minimum of the \(M\) samples \(P_{\text{Min}}(\lambda, k)\) is then found by a sample wise comparison of the actual minimum with the short time power \(P_x(\lambda, k)\). Whenever \(M\) samples have been read, i.e. when \(\lambda = \lambda_1 + M - 1\), we store the minimum power of the last \(M\) samples \(P_{\text{Min}}(\lambda = \lambda_1 + M - 1, k) = P_{\text{Max}}(\lambda = \lambda_1 + M - 1, k)\) and the search for the minimum begins over again. The minimum power of the length \(D\) window is now easily obtained as the minimum of the last \(M\) minimum power estimates \(P_{\text{Min}}(\lambda = \lambda_1 + q M - 1, k)\) with \(q = 1, 0, \ldots (2 - W)\). The decomposition of the full length window into sub windows has the advantage that a new minimum estimate is available after already \(M\) samples without an substantial increase in compare operations.
If the actual sub-band power $P_X(\lambda, k)$ is smaller than the estimated minimum noise power $P_{\text{min}}(\lambda, k)$ the noise power is updated immediately independent of window adjustment: $P_{\text{min}}(\lambda, k) = \min(P_X(\lambda, k), P_{\text{min}}(\lambda, k))$. Thus in case of decreasing noise power we achieve a fast update of the minimum power estimate. In case of increasing noise power the update of noise estimates is delayed by $D+M$ samples. Finally, to control the over subtraction factor $O_{\text{sub}}(\lambda, k)$, we compute the SNR in each sub-band

$$
(\lambda, k) = 10 \log_e \left( \frac{P_X(\lambda,k) - \min(P_n(\lambda,k), P_X(\lambda,k))}{P_n(\lambda,k)} \right)
$$

(39)

The window length $D = MW$ must be large enough to bridge any peak of speech activity, but short enough to follow non stationary noise variations.
CHAPTER - 4

IMPLEMENTATION AND EVALUATION

For evaluation, clean speech signals are corrupted with non-stationary noise in order to simulate real world environments. The evaluated systems were: wiener beamformer, spectral subtraction algorithm and a cascaded implementation of both wiener beamformer and spectral subtraction algorithm, see Figure 16.

![Diagram](image)

Figure 16: Representing cascaded implementation.

4.1 Implementation details

The complete implementation and analysis of the proposed algorithm was carried out on Windows 7 operating system and simulation is carried out using Matlab7.12.0.635 (R2011a). MATLAB allows offline implementation of signal-processing based algorithms with relative ease in preliminary investigation reducing constraints of time and memory and use of computational power of the workstation’s processor. This is recommended before implementing any algorithm on a real-time system.

To evaluate the implemented algorithm, 4 different speech signals are used. All speech signals are sampled with 16 kHz sampling frequency. These speech signals were recorded following ITU standards and contain both male and female voices so as to incorporate for testing on both the genders.

Speech signal 1 = “It’s easy to tell the depth of a well“ - female utterance.

Speech signal 2 = “Kick the ball straight and follow through“ - male utterance.

Speech signal 3 = “Glue the sheet to dark blue background“ - female utterance.
Speech signal 4 = “Pot of tea helps to pass the evening” - male utterance.

Speech-shaped noise at 5 dB and 0 dB SNR was added to the speech signals. This noise is in general stationary and was computed from the long-term spectrum of all the speeches used and resembles the spectral characteristics of the male and female speakers, as illustrated in Figure 17. Using speech-shaped noise makes the task of the enhancement algorithm more challenging as compared to using white Gaussian noise as commonly found in the literature. This is because the corrupting signal has its main energy concentrated towards the lower frequencies or, essentially, towards significant frequencies of the speech signals itself, making it more immune to the application of enhancement.

Figure 17: Long-term magnitude spectrum of the speech-shaped noise.

Figure 18 shows the time waveform of a clean speech signal. Figure 19 shows power spectral density of a clean speech signal. Figure 20 shows the speech sentence depicted in Figure 18 corrupted by the speech shaped noise, at 5 dB SNR and Figure 21 shows the same signal corrupted by speech shaped noise with noise energy equal to speech energy (at 0 dB).
Figure.18: Clean speech signal sampled at 16 kHz

Figure.19: Power spectrum density of clean speech signal sampled at 16 kHz
Figure 20: Speech signal corrupted with speech-shaped noise at 5 dB SNR

Figure 21: Speech signal corrupted with speech-shaped noise at 0 dB SNR.
4.2 WIENER BEAMFORMER

Speech enhancement is performed by processing the noise corrupted speech signals through wiener beamformer. Implementation considerations for wiener beamformer are as follows.

The distance between two microphones was set to 0.021 meters.

As speech and noise are coming from two directions, wiener beamformer removes the unwanted noise and improve the speech quality in noisy environments.

Noise corrupted speech in below figure 22 refers to signal $x_1(n)$ in the figure 9.

![Noise corrupted speech](image1)

![output of wiener beamformer](image2)

Figure 22: Output from the wiener beamformer for speech at 90° and noise at 0° in dB.
Noise corrupted speech in below figure 23 refers to signal \( x_2(n) \) in the figure 9.

![Noise corrupted speech](image)

**Figure 23: Output of wiener beamformer at speech at 345° and noise at 180° in dB.**

From figure 22 shows the noise corrupted speech i.e., \( x_1(n) \) and output of beamformer when speech coming from 90° and noise coming from 0°. From figure 23 shows the noise corrupted speech i.e., \( x_2(n) \) and output of beamformer when speech coming from 345° and noise coming from 180°.
4.3 Evaluation of wiener beamformer

Wiener beamformer is evaluated by using objective metric SNRI (signal to noise ratio improvement).

4.3.1 Signal to Noise Ratio Improvement (SNRI)

\( SNRI \) is difference between input SNR and output SNR. It is given by

\[
SNRI = 10 \log_{10} \left( \frac{\text{var}(\text{output speech})}{\text{var}(\text{output noise})} \right) - 10 \log_{10} \left( \frac{\text{var}(\text{input speech})}{\text{var}(\text{input noise})} \right)
\]

Where \( \text{var}(\text{output speech}) \) is variance of the output speech and \( \text{var}(\text{output noise}) \) is variance of output noise respectively.

Evaluation is carried out by assuming that speech and noise coming from 2 different positions i.e. position 1 - speech at 90\(^\circ\) and noise at 0\(^\circ\) respectively and Position 2 - speech at 345\(^\circ\) and noise at 180\(^\circ\) respectively.

SNR of Noise signal can be calculated as

\[
\text{SNR in dB} = (\text{length of input signal}) \times \sqrt{A} \quad [9]
\]

Where \( A = \left( \frac{\text{input signal}^2}{\text{length of input signal}} \right) \times 10^{\left( \text{desired SNR in dB} \right) / 10} \quad [9]. \)
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>14.3337</td>
<td>14.3099</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>15.6034</td>
<td>10.6017</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>19.0664</td>
<td>9.0584</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>28.0618</td>
<td>13.0576</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>33.1588</td>
<td>13.1577</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>38.5678</td>
<td>13.5670</td>
</tr>
</tbody>
</table>

Table.1: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 90° and noise at 0°.

Figure.24: SNR improvement for noise corrupted speech.
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0723</td>
<td>8.2417</td>
<td>8.1614</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>11.7769</td>
<td>6.7752</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>14.7293</td>
<td>4.7213</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>19.8401</td>
<td>4.8359</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>25.5118</td>
<td>5.5107</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>29.9901</td>
<td>4.9882</td>
</tr>
</tbody>
</table>

Table.2: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 135° and noise at 225°.

Figure.25: SNR improvement for noise corrupted speech
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>14.0746</td>
<td>14.0508</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>18.4329</td>
<td>13.4312</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>23.1425</td>
<td>13.1345</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>28.0380</td>
<td>13.0038</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>32.8644</td>
<td>12.8633</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>37.8376</td>
<td>12.8358</td>
</tr>
</tbody>
</table>

Table 3: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 210° and noise at 330°.

Figure 26: SNR improvement for noise corrupted speech.
From the Table 1 and Figure 24, i.e. speech at 90° and noise at 0°, that the SNR is improved by 11.5 dB on average. From the above table 2, Figure 25, i.e. speech at 240° and noise at 220°, it can be seen that the SNR is improved by 6.45 dB on average. The SNR improvement is high in table 1 than table 2 because in the former case the noise is coming...
from 0°, which is in alternate direction to the microphone so all the noise is removed. From the Table 3 and Figure 26, i.e. speech at 210° and noise at 330°, that the SNR is improved by 13.4 dB on average. From the above table 4, Figure 27, i.e. speech at 345° and noise at 180°, it can be seen that the SNR is improved by 12.7 dB on average. The SNR improvement is high in table 3 than table 4 because in the former case the noise is coming from the angle far away when compared to speech, hence the improvement is higher in both cases. So by above evaluation one can say that wiener beamformer succeeds in removing noise in almost all directions.

4.4 SPECTRAL SUBTRACTION

The speech signals were corrupted with different types of noises like Babble noise, Brown noise, White Noise and AC noise. The noise corrupted speech signals were processed by spectral subtraction and the output signals were analyzed in terms of Signal to Noise Ratio Improvement (SNRI).

Implementation considerations for spectral subtraction algorithm are as follows. WOLA filter bank considerations are

Length of the window \( L=256 \), number of sub-bands \( M=128 \), oversampling factor \( K=L/M \), Decimation ratio \( R=M/K=64 \).

The spectral subtraction parameters were over subtraction factor \( O_{\text{sub}} = 4 \), smoothing constant alpha = 0.9, spectral floor constant \( sub_f = 0.03 \) and factor to compensate bias \( O_{\text{min}}=1.1 \).

Figure.28: clean speech corrupted by babble noise (above) and output of the spectral subtraction (below).
The output of implemented spectral algorithm is shown above figure 28.

**Evaluation of Spectral Subtraction**

From the table 5 and Figure 29, it is evident that SNR is improved by 7.25 on average and by examining the above results, we can say that spectral subtraction suppress the noise successfully.

<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>6.0917</td>
<td>6.0917</td>
</tr>
<tr>
<td>5</td>
<td>4.9929</td>
<td>12.6042</td>
<td>7.6113</td>
</tr>
<tr>
<td>10</td>
<td>9.9882</td>
<td>17.0375</td>
<td>7.0493</td>
</tr>
<tr>
<td>15</td>
<td>15.0234</td>
<td>22.3651</td>
<td>7.3417</td>
</tr>
<tr>
<td>20</td>
<td>19.9869</td>
<td>28.4836</td>
<td>8.4968</td>
</tr>
<tr>
<td>25</td>
<td>25.0238</td>
<td>32.6159</td>
<td>7.5921</td>
</tr>
</tbody>
</table>

Table 5: Tabular form for SNRI for spectral subtraction.

![Figure 29: SNR improvement for noise corrupted speech.](image)
As for the cascaded implementation, we combine all the mentioned algorithms i.e. the wiener beamformer and spectral subtraction algorithm in orderly fashion as shown in Figure 16.

With all the parameters that are considered above, a cascaded implementation of wiener beamformer and spectral subtraction algorithm in reverberant environment is realized by considering the following assumptions.

Cascaded implementation can be realized by connecting wiener beamformer with spectral subtraction in an orderly fashion. As wiener beamformer is implemented with 2 microphones correspondingly, spectral subtraction algorithm is implemented for each microphone.

The evaluation is made with 0 dB speech shaped noise and speech at 90° and noise at 0° and their response is shown in Figure 30.

Figure.30: clean speech corrupted by noise (above) and output of cascaded implementation (below).
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>10.3407</td>
<td>10.3169</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>16.6208</td>
<td>11.6191</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>19.0443</td>
<td>9.0363</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>26.0360</td>
<td>11.0318</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>29.8341</td>
<td>9.8303</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>33.1598</td>
<td>8.0158</td>
</tr>
</tbody>
</table>

Table 6: Tabular form showing values cascaded implementation in terms of SNRI for speech at 90° and noise at 0°.

Figure 31: SNR improvement for noise corrupted speech.
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>7.9087</td>
<td>7.8849</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>13.9232</td>
<td>8.9215</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>17.6678</td>
<td>7.6598</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>23.3679</td>
<td>8.3659</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>26.3336</td>
<td>6.3325</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>31.9376</td>
<td>6.9358</td>
</tr>
</tbody>
</table>

Table 7: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 135° and noise at 225°.

Figure 32: SNR improvement for noise corrupted speech.
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>11.1972</td>
<td>11.1734</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>16.9375</td>
<td>11.9358</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>20.3739</td>
<td>10.3659</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>24.3408</td>
<td>9.3366</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>32.9867</td>
<td>12.9856</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>36.4535</td>
<td>11.4517</td>
</tr>
</tbody>
</table>

Table.8: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 210° and noise at 330°.

Figure.33: SNR improvement for noise corrupted speech.
<table>
<thead>
<tr>
<th>Noise in dB</th>
<th>SNR input</th>
<th>SNR output</th>
<th>SNRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0238</td>
<td>9.7067</td>
<td>9.6829</td>
</tr>
<tr>
<td>5</td>
<td>5.0017</td>
<td>14.3383</td>
<td>9.3366</td>
</tr>
<tr>
<td>10</td>
<td>10.008</td>
<td>22.5232</td>
<td>12.5214</td>
</tr>
<tr>
<td>15</td>
<td>15.0042</td>
<td>25.4559</td>
<td>10.4517</td>
</tr>
<tr>
<td>20</td>
<td>20.0011</td>
<td>29.7465</td>
<td>9.7454</td>
</tr>
<tr>
<td>25</td>
<td>25.0018</td>
<td>34.9195</td>
<td>9.9177</td>
</tr>
</tbody>
</table>

Table 9: Tabular form showing values for wiener beamformer in terms of SNRI for speech at 345° and noise at 180°.

Figure 34: SNR improvement for noise corrupted speech.

In noisy environment, it can be seen from the above table 6, Figure 31, i.e. speech at 90° and noise at 0°, that the SNR is improved by 9.8 dB on average. From the above table 7, Figure 32, i.e. speech at 135° and noise at 225°, it can be seen that the SNR is improved by 7.6 dB on average. The SNR improvement is high in table 6 because in the former case the noise is coming from 0°, which is in alternate direction to the microphone so all the noise is removed. From the table 8, and Figure 33, i.e. speech at 210° and noise at
330°, SNR is improved by 11.1 on average. From above table 9 and Figure 34, i.e. speech at 345° and noise at 180°, it can be seen that the SNR is improved by 10.9 dB on average. SNR improvement is higher in table 8 because the noise is coming from angle far away from 0°, which is in alternate direction to the microphone.

From the graphs, we can analyze that combined system is the best choice in terms of SNR improvement. The performance variations of the combined system is almost same as wiener beamformer i.e. performance of system depends on the relative positions of the source and noise. SNR improvement is high when the noise and speech lies in different quadrants and improvement is low when the speech and noise lies in same quadrant and performs worse when the speech source and noise lies very near to each other. Performance of the combined system also depends on the distance from source to microphone array, shorter the distance of the source better the performance.

By observing the above values for SNRI, it is clear that the implemented cascaded model has suppressed the noise in noisy environments successfully.
CHAPTER – 5

CONCLUSION

Noise suppression is successfully achieved by using both Wiener Beamformer and Spectral subtraction. The speech signal which needs to be enhanced is corrupted by the acoustic noise. The purpose of this thesis is to suppress the interference, acoustic noise. The corrupted speech signal is decomposed into number of sub bands using analysis bank of the WOLA filter bank. These sub bands are then processed using the enhancement algorithms i.e. Wiener Beamformer, Spectral subtraction and unique combination of both. The performance of the each system is measured in terms of the SNR improvement, to measure quality of output desired signal. The input SNR is taken from 0dB to 25dB. The performance is compared between 3 systems and best choice is indicated. The wiener beamformer provides the efficient enhancement of the corrupted speech signal when compared to the spectral subtraction. The performance of the wiener beamformer mainly depends on the relative positions of the source and noise. The high performance is obtained from the wiener beamformer when source is away from noise and performance decreases as the noise approaches near to source.

Wiener beamformer proves to be efficient in reducing noise and Spectral subtraction enhances the quality of speech. Wiener beamformer is proved to be more successful than Spectral subtraction algorithm as there is a significant rise in SNR improvement when compared to the spectral subtraction. The combined system of wiener beamformer and spectral subtraction provides better SNR improvement than individual implementations, but the quality of speech decreases when compared to wiener beamformer. The combined system also behaves same with the positions of the source and noise as wiener beamformer behaves.

So, one can conclude that cascaded system succeeds in reducing noise as well as enhancing the speech quality.

Scope of the future work can be extended by implementing the above model with larger microphone arrays like, 8 or 16, so as further increase the speech quality.
CHAPTER – 6

REFERENCES

12. Siva Kumar Chikkala, --Mobile Phone Acoustics – In Near Talk Mode with emphasis on ELKO Beamformer and Spectral Subtraction, BTH, Karlskrona, 2013 (Master’s Thesis).


18. ROBUST SPEAKER DIARIZATION FOR MEETINGS by Xavier Anguera.