Audio Signal Analysis for Classification and Source Localization in e-Learning Applications

Deepti Singh

A thesis submitted to the University of Dublin, Trinity College

in fulfillment of the requirements for the degree of

Doctor of Philosophy

December 2007

Audio Signal Analysis for Classification and Source Localization in e-Learning Applications

Approved by
Dissertation Committee:

____

____

____

____

____

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this or any other University, it is entirely my own work and I agree that the Library may lend or copy the thesis upon request.

Deepti Singh

Deepti Singh

Dated: December 11, 2007

To my family.

Acknowledgements

I would like to express my gratitude to my thesis supervisor Prof. Frank Boland for his guidance, patience and help. I would also like to thank Trinity College, Dublin for giving me an opportunity to be a part of the college, for the scholarship and great facilities. I am also grateful to Enterprise Ireland and Science Foundation Ireland for their financial support.

I also would like to thank the Electronic and Electrical Engineering Department for their support. I am grateful to the School of Computer Science and Statistics for the Post-Graduate diploma in Statistics course. Special thanks to Dr. Myra O' Reagen, Dr. Cathal Walsh, Dr. Brendan Murphy and Ms. Aideen Keaney for their great teaching and support.

I would like to thank my husband, Dr. Atul Singh, for constructive criticism that improved my work and for his inputs throughout my work and thesis writing phase. I also would like to thank him for the encouragement, help and support. I also would like to thank my parents, Dr. Faujdar Singh and Dr. Anita Singh, for motivating me, listening to me, encouraging me, providing me an opportunity to grow, doing everything that is in my best interest and making me what I am. I would like to thank my sister, brother and brother-in-law for encouraging me and being there for me when I needed them. Special thanks to my niece Aarushi for her smiling photographs that encouraged me to work with a smile.

I am grateful to Angela, Damien, Darren, Denis and Gavin for their help and co-

operation. I would like to especially thank Bernedette Clerkin, Mrs. Subhadra Rao and Dr. Hanumantha Rao for their help, advice and support.

Last but not least I would like to thank everyone whom I have not thanked explicitly here. Thank you all.

Abstract

E-learning provides an opportunity for the students in a classroom to receive and participate in lectures broadcasted from a remote location. For convenience and scalability it is preferable to have a microphone array instead of an individual microphone per user to capture the signals. A Direction of Arrival (DOA) system can be used to determine the angle of arrival of the speech signal so that a camera can be steered in the direction of the signal source. In a DOA system for e-learning, Time Delay Estimation (TDE) techniques are suitable to determine the relative time difference between the signals to reach the microphones in the array. This time difference can be used to find the Direction of Arrival (DOA) of the signal. This thesis presents a DOA system called Modified yIN based Doa Estimator for e-leaRning (MINDER).

The primary challenge in the DOA estimation in an e-learning environment is to get high accuracy of the TDE in the presence of noise and reverberation. The existing algorithms for TDE (such as Cross-correlation (CC), Generalized Cross-Correlation Phase Transform (GCC-PHAT), Average Magnitude Difference Function (AMDF)) do not perform well in the presence of both high noise and reverberation. The YIN algorithm is based on the AMDF algorithm and is used to find the fundamental frequency of the music and speech signals. This thesis proposes two TDE algorithms based on the YIN algorithm. The first algorithm is called the modified YIN based TDE algorithm and the second algorithm is called the Weighted YIN based TDE algorithm. The Weighted YIN based TDE algorithm is inspired by the Weighted Cross-Correlation (WCC) algorithm proposed by Chen et al. [32]. Simulations presented in this thesis show that the modified YIN based TDE algorithm has higher accuracy for TDE when compared to WCC, Weighted YIN based TDE and GCC-PHAT algorithms in presence of high noise and reverberation. Its accuracy for TDE is comparable and marginally better than the contemporary Modified Average Magnitude Difference Function (MAMDF) algorithm proposed by Chen et al. [32]. The performance of the Weighted YIN based TDE algorithm is similar to the WCC and GCC-PHAT algorithm. MINDER uses the modified YIN based TDE algorithm for TDE.

Voice Activity Detection (VAD) is used to detect the presence of speech in a signal. It is a useful pre-processing stage to DOA as the number of DOA computations is reduced and steering of the video camera in the direction of the undesirable signal sources is eliminated. The key steps in designing a VAD system is deciding on the classification algorithm and the signal features that can be input to the classifier. This thesis proposes a new feature for VAD called the Mel-Spectrum Teager Energy (MSTE) coefficients. The performance of MSTE coefficients is on par with the Mel-Frequency Cepstrum Coefficients (MFCC) feature that is used in speech processing applications. This thesis presents a VAD system based on MSTE coefficients. Other features from the existing literature have been included in the feature set used in the VAD system to improve the classification. The feature set was optimized using forward feature selection in LDA. Simulations in the thesis show that Linear Discriminant Analysis (LDA) and radial Support Vector Machine performed well for classification. The VAD system uses LDA for classification. In the VAD system the signal is classified into four classes namely the voiced speech, unvoiced speech, stationary noise and the non-stationary noise. The VAD system is used as a pre-processing stage in MINDER.

Contents

Acknowledgements i

Summary ii

List of Figures ix

List of Tables xi

Chapter 1 Introduction 1

1.1 E-Learning Environment 2

1.2 DOA system 5

1.2.1 Voice Activity Detection 6

1.2.2 Direction of Arrival Estimation 8

1.3 Contributions 9

1.4 Thesis Outline 10

Chapter 2 State of the Art: Direction of Arrival Estimation for Source Localization. 12

2.1 Signal Model 13

2.2 Angle of arrival 16

2.3 Spatial Aliasing 18

2.4 DOA Resolution 20

2.5 Parabolic Interpolation 21

2.6 Classroom Environment 23

2.7 Direction Of Arrival Algorithm Taxonomy 24

2.8 Steered Beamformer Based Algorithms 25

2.8.1 Delay and Sum Beamformer 27

2.8.2 Frequency Domain Beamforming 29

2.9 High Resolution Spectral Estimation Based Algorithms 32

2.9.1 Subspace Techniques 32

2.10 Time Difference Of Arrival Based Algorithms 34

2.10.1 Cross-correlation 35

2.10.2 GCC based algorithms 37

2.10.3 Average Magnitude Difference Function (AMDF) 44

2.10.4 Others 46

2.11 Discussion and Conclusions 47

2.12 The YIN Algorithm 49

2.13 Summary 52

Chapter 3 State of the Art: Voice Activity Detection 53

3.1 Voice Activity Detection (VAD) System 53

3.2 Speech Signal Characteristics 54

3.3 Audio Signal Features 57

3.4 Feature Selection for Classification 68

3.4.1 Feature Selection 69

3.4.2 Backward and Forward selection 72

3.5 Classifier for Voice Activity Detection 74

3.6 Classification Using Supervised Machine Learning Algorithms 77

3.6.1 Multivariate Gaussian Distribution 77

3.6.2 Bayes Decision Theory 79
3.6.3 Supervised Machine Learning Algorithms 80
3.7 Summary 92

Chapter 4 DOA Studies **93**
4.1 YIN for TDE 93
4.2 Weighted YIN 96
4.3 Performance Evaluators for Time Delay Estimator 96
4.4 Simulation Setup 97
4.5 Simulation Results 100
4.6 Observations and Conclusion 112

Chapter 5 VAD Feature Selection Studies **116**
5.1 Mel-Spectrum Teager Energy Coefficients 117
5.2 Setup 119
5.3 Supervised Machine Learning Algorithm Selection for VAD 120
5.4 Features and Feature Selection Results 122
5.5 VAD Four Class Classification Results 124
5.6 Observations and Conclusion 130

Chapter 6 DOA System **132**
6.1 Experimental Setup 132
6.2 Direction of Arrival System Experimental Results 135
6.3 Conclusion 140

Chapter 7 Conclusion **142**
7.1 Thesis Summary 142

7.2 Future Work 147

Bibliography **149**

List of Figures

1.1 Schematic diagram of the DOA System. 5

2.1 Direction Of Arrival (DOA) estimation. 16

2.2 Example of Spatial Aliasing 19

2.3 Delay and Sum Beamformer. 28

3.1 Schematic diagram of the VAD System. 54

3.2 Voiced (left) and Unvoiced (right) Speech Signals 55

3.3 Power Spectral Density for Voiced and Unvoiced Speech Signals 57

3.4 Relation between the signal frequency in Hz and perceived signal frequency in Mel units. 66

3.5 Plot of 2-dimensional Gaussian Distribution. 78

3.6 Example of separable problem of SVM in a 2 dimensional space 87

4.1 Source locations for simulations. 98

4.2 Plot for female speaker with microphone separation (d) = 3.4 cm . . . 101

4.3 Plot for female speaker with microphone separation (d) = 6.8 cm . . . 102

4.4 Plot for female speaker with microphone separation (d) = 10.2 cm . . . 103

4.5 Plot for female speaker with microphone separation (d) = 3.4 cm . . . 104

4.6 Plot for female speaker with microphone separation (d) = 6.8 cm . . . 105

4.7 Plot for female speaker with microphone separation (d) = 10.2 cm . . . 106

4.8 Plot for male speaker with microphone separation (d) = 3.4 cm 107

4.9 Plot for male speaker with microphone separation (d) = 6.8 cm 108

4.10 Plot for male speaker with microphone separation (d) = 10.2 cm 109

4.11 Plot for male speaker with microphone separation (d) = 3.4 cm 110

4.12 Plot for male speaker with microphone separation (d) = 6.8 cm 111

4.13 Plot for male speaker with microphone separation (d) = 10.2 cm 113

6.1 Source locations for DOA system experiment. 133

6.2 VAD Experimental results classifying the audio signal into Voiced Speech (VS) and others (consisting of the unvoiced speech, stationary noise and the non-stationary noise signals). 135

6.3 Position 1: TDE results for source at 57°. 137

6.4 Position 2: TDE results for source at 90°. 138

6.5 Position 3: TDE results for source at 131°. 139

List of Tables

3.1 Center frequency of 20 critical bands 65

4.1 Delay values in signal sample number and their corresponding source angles for different separation between microphone pair (d) at sampling frequency = 16kHz 99

5.1 The Hit Rate (HR) and False Alarm Rate (FAR) for classifying the test data into four classes. 121

5.2 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 10 dB. 125

5.3 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 15 dB. 126

5.4 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 20 dB. 127

5.5 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 10 dB. 128

5.6 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 15 dB. 129

5.7 The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 20 dB. 130

6.1 Hit Rate (HR), False Alarm Rate (FAR) for each class and overall HR for the experimental signal using LDA. 134

Chapter 1

Introduction

E-learning provides an opportunity for the students in remote classrooms to actively participate in the lectures broadcasted from the main classroom. A Microphone Array (MA) is used to capture the audio signal. The knowledge of the Direction Of Arrival (DOA) of the source signal is used to steer the video camera in the direction of the signal source. The geometry of the microphone array can be used to find the source location. Detecting the presence of speech in the audio signal (voice activity detection) so that the DOA estimation is performed only on parts of audio signal containing speech reduces the number of DOA computations and also prevents the steering of the video camera in the direction of undesired signal sources.

This thesis deals with the DOA estimation in an e-learning environment. This thesis proposes a DOA system called MINDER (Modified yIN based Doa Estimator for e-leaRning), consisting of the voice activity detection and the DOA stages, for e-learning environment. The first section of the chapter provides a brief introduction to the e-learning environment. The DOA system is introduced in the following section. Finally the contribution of the thesis and its outline is presented.

1.1 E-Learning Environment

This section focuses on the e-learning environment, on which the entire work of this thesis is based. In an e-learning environment a group of classrooms are connected together through the internet. The lecturer is present in any one of the classrooms. For convenience the room where the lecturer is present is called the principal classroom and the other classrooms are called the remote classrooms. The principal classroom is connected to the remote classrooms through the Internet that transmits the audio and video data of the lecture being delivered in the principal classroom. The communication between the classrooms is typically duplex so that the students in the remote classrooms not only attend the lecture remotely but also benefit from the interaction with the lecturer in the principal classroom. This arrangement is convenient for the students as they do not have to travel or relocate to gain knowledge. It also encourages collaboration between educational institutes by providing more options for the students to explore different courses offered to them.

Existing Tools

Several e-learning tools are available to make e-learning pleasant for the students. The complexity of these tools depends on the functionality they offer. Blackboard [1] is a software tool that facilitates interaction between different groups and the lecturer through internet. The students are able to listen to the lecture on their computer and interact with the lecturer through text based chat or by voicing their questions by being in a queue where the students are allowed to speak on first come first serve (First In First Out) basis. This tool also allows to restrict access, setup material such as class notes, homework, projects, assignments, solutions and students performance (grades, marks etc) information. It also allows recording the lectures so that the student can listen to them later at their own convenience.

Echalk [2, 59] is another powerful e-learning tool that uses an electronic whiteboard instead of a chalkboard, a digitizer tablet connected to a computer and a retroprojection system in a classroom. The whiteboard is used instead of a chalkboard to write and draw directly on the screen. The digitizer tablet can also be used to write on the screen. It also provides features such as plotting mathematical functions on the screen, evaluating handwritten mathematical expression in real time using handwriting recognition, displaying pictures from internet or a computer. The lectures are recorded so that the student can listen to them later. Students in remote classrooms are provided with the live feed of the lecture that they can access using a Java enabled web browser. The students can view the lecturer on the screen and hear the lecture. It also facilitates a printable PDF version of the content of the screen so that students can concentrate on understanding the lecture instead of trying to take notes.

Classrooms for e-learning

In e-learning case where educational institutes allocate rooms at the time of the lecture so that a group of students can listen to the lecture in a classroom with live feed from the principal classroom, the challenge in such a scenario is to localize the student interacting with the lecturer. Classroom protocol requires that there is only one talker in the room at any instant. This means that when the lecturer is talking then the rest of the audience members are listening and when the lecturer is interacting with the audience members then only one member of the audience speaks at any instant. Assuming that the classroom protocol is adhered to, then at any given instant there will be only one speech source in the room.

Some of the source localization systems consist of a camera to localize the member of the audience and a microphone for each of the audience members [3]. The audience members can press a button associated with the microphone to inform that they wish to speak. The camera is steered in the direction of the talker based on the microphone location from which the cue to speak was received. This system though simple and efficient requires lot of hardware which is expensive and prone to frequent damage as it is in direct contact with the audience members.

Another approach is to use a microphone array to capture the signal and estimate the Direction Of Arrival (DOA) and the source location based on the geometry of the microphone array [27, 79]. In such a system, the DOA of the speech signal is estimated on the signals captured by a microphone array located away from the audience. The relative time difference of arrival of the signal to reach a pair of microphones in the microphone array is used to find the DOA of the signal. The DOA estimates obtained from different pairs of microphones can be used to localize the speech source and steer the camera in the direction of the speech source. This thesis concentrates on source localization using DOA for an e-learning environment.

The problem associated with finding the direction of arrival of the signal is the presence of noise and reverberation in the room that tends to give rise to errors in the Direction Of Arrival (DOA) estimates. Excessive levels of noise and reverberation also reduce the comprehensibility of speech. This problem is worsened among students with hearing disability. American Speech Language Hearing Association (ASHA) [4] and the Acoustical Society of America [100] recommend that the Signal to background Noise Ratio (SNR) in a classroom should not be less than 10 dB and for students with some hearing impairment it should not be less than 15 dB. Similarly the recommended reverberation time for classrooms is 0.4 - 0.6 seconds [100].

Apart from the presence of noise and reverberation in the captured signal there are instances when the signal received by the microphones in the source localization system does not contain speech. The signal contains background noise that comprises of the noise made by the electronic equipment present in the classroom such as computers/overhead projector fan noise and noise of the Heating, Ventilating and Air Conditioning (HVAC) systems. This leads to spurious estimates of the direction of

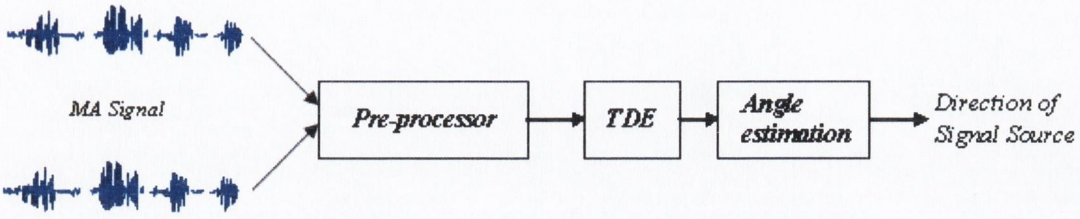


Fig. 1.1: Schematic diagram of the DOA System.

arrival and the camera will be steered in the direction of the background noise source.

Other undesirable signals include human produced sounds such as laugh, cough and sneeze and non-human produced sounds like paper shuffle, closing/opening of the door, key jangle and sounds from coins. In order to avoid computation of the DOA in the absence of speech signal a voice activity detection (VAD) system is used to recognize segments of signal that contain speech. The location of the source is estimated only on the segments of signal identified as speech. This reduces computations as well as prevents steering of camera at spurious locations.

1.2 DOA system

Direction Of Arrival (DOA) system is used to find the DOA of the speech signal and steer a camera in the direction of the speech source. The DOA information can also be used in beamforming and for source localization. The audio signal captured by the microphone array and the distance between the microphones in the microphone array is used to find the DOA of the source signal. Figure 1.1 shows a schematic diagram of the proposed DOA system.

This thesis presents a DOA system called Modified yIN based Doa Estimator for e-leaRning (MINDER). MINDER consists of two stages namely the Voice Activity Detection (VAD) stage for signal pre-processing and the DOA stage. The VAD system in this work not only detects speech frames but also detects stationary noise and non-stationary noise. The captured signal is classified into four classes namely the voiced speech, unvoiced speech (refer to section 3.2 for more information on voiced and unvoiced speech), stationary noise and the non-stationary noise. The DOA stage finds the DOA of the speech source based on the relative time difference of arrival of the speech signal received by a pair of microphones in the microphone array. As shown in figure 1.1, the DOA stage estimates the DOA of the source in two steps namely Time Delay Estimation and Angle Estimation.

1.2.1 Voice Activity Detection

Pre-processing the signals to detect the presence of speech in the audio signal is useful in speech recognition, speaker recognition and source localization applications. In some applications that involve noise cancellation identification of the background noise is of primary concern. Voice Activity Detection is used to detect the presence of speech signal in the audio signal. The characteristics of the signal called the features are employed by a VAD algorithm to distinguish between speech and non-speech signals. For VAD the audio signal is divided into small frames of fixed length. The values of the features are calculated for each frame and are input to the VAD algorithm.

The number and types of classes into which the audio signal is classified depends on the application requirements. For example if an application must distinguish between speech and other types of signals, as required in VoIP applications, then the audio signal can be classified into two classes namely speech and non-speech. Speech can be further classified into voiced speech and unvoiced speech. This is used in applications where the DOA estimation is performed on voiced segments of speech. Noise can be further classified into stationary noise (background noise) and the non-stationary noise.

The type of classes that the audio signal is classified into depends on the requirement of the application.

VAD can be considered as a pattern classification problem. This work proposes to use the supervised machine learning algorithms for VAD for e-learning. Pattern classification using supervised machine learning algorithms generally involves solving two crucial problems [87, pg. 171]. They are the selection/evaluation of features and decision rule for classification. Feature selection involves determining the signal features that help to distinguish between the signals. Speech signals can be described by some of their properties. But noise signals do not have specific properties. Thus determining a good feature set that result in an accurate classification is a challenge. Another challenge is to identify a classification algorithm that can classify the signals well into their respective classes. A brief discussion of the key steps and the associated challenges involved in desiging a VAD system are presented in [106].

In this work, the audio signals are classified into four classes namely the voiced speech, unvoiced speech, stationary noise and the non-stationary noise using the Discriminant Analysis (DA), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Classifying the signal into the four classes is useful for source localization as well as post-processing e-learning.

The signal in this work consists of four classes of signals belonging to the speech and non-speech signal category. Two of the four signal classes belong to the speech category namely the voiced and unvoiced speech signals. Non-speech signal is divided into stationary noise and non-stationary noise. Stationary noise consists of the background noise. Non-stationary noise consists of the cough, sneeze, laugh, paper shuffle, closing/opening of door, key jangle and sounds of coins.

1.2.2 Direction of Arrival Estimation

As shown in figure 1.1, the Direction Of Arrival (DOA) of the source signal is found in two steps. The first step is to find the relative time delay between the signals received by the microphone pair in the microphone array. Second step involves finding the actual DOA based on the time delay information from the first step.

Time Delay Estimation

Time Delay Estimation (TDE) is one of the popular methods for DOA estimation for single source. This method of finding the DOA is known as Time Difference Of Arrival (TDOA) based method. All the microphones in the MA do not receive the audio signal simultaneously. The relative time difference for the signal to reach the microphones in a given microphone pair is found in signal sample numbers. The time delay information is used to find the DOA of the source signal. TDE is obtained on various combinations of microphone pairs in the microphone array. Knowledge of TDE and array geometry is used for source localization. This method performs better than other methods in presence of noise and reverberation. It is also computationally practical for real time applications. The primary challenge in the DOA estimation in e-learning is to get high accuracy of the TDE in the presence of high noise and reverberation. Simulations in chapter 4 suggest that existing TDE algorithms such as the Generalized Cross-Correlation Phase Transform (GCC-PHAT) do not perform well in the presence of noise and reverberation. This work proposes a modified YIN based TDE algorithm for DOA estimation. Another YIN based algorithm called Weighted YIN based TDE algorithm is also presented.

Angle Estimation

The formula used for the angle of arrival of the signal source is based on the far-field assumption. The formula is given by [115, 91]:

$$\theta = cos^{-1}\left(\frac{\tau c}{f_s d}\right) \tag{1.1}$$

The angle estimation involves the estimated time delay (τ), the velocity of sound (342 m/s), the signal sampling frequency (f_s) and the separation (d) between a pair of microphones in the array.

1.3 Contributions

The primary contribution of this thesis is an extension of the YIN algorithm called the modified YIN based TDE algorithm that can be used for TDE. Another extension of the YIN algorithm called the Weighted YIN based TDE algorithm is also presented. The thesis presents simulations comparing the performance of the modified YIN based TDE algorithm and the Weighted Yin based TDE algorithm with other TDE algorithms like the Generalized Cross-Correlation Phase Transform (GCC-PHAT), Modified Average Magnitude Difference Function (MAMDF) and the Weighted Cross-Correlation (WCC) for different SNR and fixed reverberation time.

The thesis also proposes a Teager Energy Operator (TEO) based feature known as the Mel-Spectrum Teager Energy (MSTE) coefficients for VAD. A VAD system based on MSTE that used Linear Discriminant Analysis (LDA) supervised machine learning algorithm constitutes the secondary contribution of this thesis. The thesis presents a comparison of the performance of the MSTE coefficients and Mel Frequency Cepstrum Coefficients (MFCC) for classification along with seven other existing features to classify the signal into four classes namely the voiced speech, unvoiced speech, stationary noise and the non-stationary noise. The supervised machine learning algorithms considered for VAD for e-learning are the Discriminant Analysis (DA), Artificial Neural

Network (ANN) and Support Vector Machine (SVM). The seven features considered for VAD along with MSTE and MFCC are the log energy of the signal, zero-crossing rate, second and third Linear Predictive (LP) coefficients, absolute difference between the first two Partial Correlation or Reflection (PARCOR) coefficients, skewness of the LP residual signal and the spectral centroid of the signal. A DOA system called Modified yIN based Doa Estimator for e-leaRning (MINDER) is presented that combines the VAD system based on MSTE coefficients and the modified YIN based TDE algorithm. Experimental results for Modified yIN based Doa Estimator for e-leaRning (MINDER) are presented.

1.4 Thesis Outline

The remainder of the thesis is organized as follows:

Chapter 2 presents the state of the art of the DOA algorithms and explains the YIN algorithm for fundamental frequency estimation.

Chapter 3 explains the background associated with the supervised machine learning algorithms. It explains the features of the speech signal and the feature selection process. It also describes the DA, cascade ANN and the SVM algorithms.

Chapter 4 presents the simulation results for the TDE for DOA comparing the performance of the GCC-PHAT, contemporary Modified Average Magnitude Difference Function (MAMDF) and Weighted Cross-Correlation (WCC), proposed modified YIN based TDE algorithm and Weighted YIN based TDE algorithm.

Chapter 5 presents the VAD system for the DOA system. This chapter includes the initial feature selection and VAD results for selecting the supervised machine learning algorithms for inclusion in the VAD system. It also presents the simulation results comparing the performance of the proposed MSTE coefficients and MFCC using the LDA and the radial SVM algorithms at different SNR values. It also presents the performance of the MTSE with the features that were included in the feature set to improve the classification.

Chapter 6 combines the VAD system of chapter 5 and the modified YIN based TDE algorithms to form the DOA system. It presents the experimental results on signal with VAD and parabolic interpolation.

Chapter 7 presents the conclusion of the thesis.

Chapter 2

State of the Art: Direction of Arrival Estimation for Source Localization.

This chapter presents the theory associated with the Direction Of Arrival (DOA) algorithms and the state of art of the DOA algorithms for source localization. As discussed in section 1.2, a Direction Of Arrival (DOA) system consists of a pre-processor stage and a TDE stage that is used for the DOA estimation. In this chapter the signal model is introduced with the basic theory behind DOA estimation. It also describes and discusses the DOA algorithms for source localization. This is followed by a description of the YIN algorithm that is utilized for fundamental frequency estimation. This is required to understand the primary contribution of this thesis which is a YIN based novel approach for Time Delay Estimation (TDE).

2.1 Signal Model

The Signal model used in the DOA algorithms is provided in this section. The DOA algorithms exploit the properties of the signal to estimate the DOA of the source signal. The signal model presented in this section is utilized by the DOA algorithm categories mentioned in section 2.7.

Assuming that there is more than one signal source in the room the signal received by the i^{th} microphone ($x_i(t)$) in the microphone array is represented by [91]:

$$x_i(t) = \sum_{j=1}^{Q} h_{ij}(t) \ * \ s_j(t) + n_i(t) \tag{2.1}$$

where $s_j(t)$ is the signal from the j^{th} source, $h_{ij}(t)$ is the cascade of the impulse responses of the room and the response of the i^{th} microphone for the j^{th} source, '$*$' is the convolution operator, $n_i(t)$ is the additive noise component and Q is the number of signal sources. All the acoustic paths (direct and reflected signal components) between the source and the microphone are represented by the room impulse response. The impulse response of the room depends on the position of the microphone with respect to the source, the characteristics of the propagation medium and contents of the room. When the position of the source and the microphones is fixed, over short time interval the impulse response of the room can be considered to be time-invariant [24, pg. 165]. The microphone response includes the gain of the microphone. The additive noise component consists of the ambient noise and the noise of the microphone. The additive noise signal and the source signal are assumed to be uncorrelated [91].

For a single source the signal received by the i^{th} microphone is given by [24, pg. 165]:

$$x_i(t) = h_i(t) \ * \ s(t) + n_i(t) \tag{2.2}$$

where $s(t)$ is the source signal, $h_i(t)$ is the cascade of the impulse response of the room and the response of the i^{th} microphone for the source signal and $n_i(t)$ is the additive noise component.

For time difference of arrival computation, when interest is particularly on the direct path component, the signal received by the i^{th} microphone can be rewritten as [24, pg. 165]:

$$x_i(t) = \frac{1}{r_i}\ s(t - \tau_i) * g_i(t) + n_i(t) \tag{2.3}$$

where r_i is the distance between the source and the microphone, $g_i(t)$ is the impulse response of the room due to the other reflected components and the i^{th} microphone (consists of all the components of $h_i(t)$ of equation 2.2 except the direct path component), τ_i is the delay due to the direct path component.

Assuming that the impulse response $(g_i(t))$ in equation 2.3 is similar for all the microphone signals in the microphone array, the signal received by the microphones with respect to the reference microphone in the microphone array will consist of a scaled and time shifted version of the signal received by the reference microphone [24, pg. 166]. Thus the signals received by the microphones with respect to the reference microphone (l) in the microphone array is given by:

$$x_i(t) = a_i\ s(t - \tau_l - \tau_{l,i}) + n_i(t) \tag{2.4}$$

where $s(t - \tau_l)$ is the signal received by the reference microphone l, a_i is the scaling factor for the received signal $s(t - \tau_l)$ for the i^{th} microphone, $\tau_{l,i}$ is the time shift (time difference of arrival) of the signal between the reference microphone (l) and the i^{th} microphone $(\tau_i - \tau_l)$.

The model of the signal vector relative to the source signal $s(t - \tau_0)$ translated to the reference microphone $l = 0$, of the M element microphone array is represented by:

$$\begin{aligned}\mathbf{x}(t) = [&a_0\ s(t - \tau_0 - \tau_{0,0})\ \ a_1\ s(t - \tau_0 - \tau_{0,1})\ \ a_2\ s(t - \tau_0 - \tau_{0,2})\ \ \cdots \\ &a_{M-1}\ s(t - \tau_0 - \tau_{0,M-1})]^T\ +\ [n_0(t)\ \ n_1(t)\ \ n_2(t)\ \cdots\ n_{M-1}(t)]^T\end{aligned} \tag{2.5}$$

In the equation 2.5, the term τ_0 represents the time delay for the source signal $(s(t))$ to arrive to the reference microphone $(l = 0)$. The term $\tau_{0,M-1}$ represents the relative time difference of arrival of the signal between the reference microphone ($l = 0$) and the $(M-1)^{th}$ element of the microphone array. Since the term τ_0 is common to all the microphone and it does not affect the relative time difference of arrival estimates the term can be omitted. The term $\tau_{0,0}$ is zero.

The model of the signal vector relative to the source signal $s(t)$ translated to a reference element 0 of the array in equation 2.5 is rewritten as:

$$\begin{aligned}\mathbf{x}(t) = {} & [a_0 s(t) \;\; a_1 s(t-\tau_{0,1}) \;\; a_2 s(t-\tau_{0,2}) \;\; \cdots \;\; a_{M-1} s(t-\tau_{0,M-1})]^T \\ & + [n_0(t) \;\; n_1(t) \;\; n_2(t) \;\; \cdots \;\; n_{M-1}(t)]^T \end{aligned} \tag{2.6}$$

where $\tau_{0,i}$ is the relative time delay between the reference microphone ($l = 0$) and i^{th} microphone. Discrete Fourier transform [37, pg. 86-92] of signal vector $\mathbf{x}(t)$ gives the frequency domain representation of the signal. The signal vector in the frequency domain $X(\omega)$ is represented by:

$$\begin{aligned}\mathbf{X}(\omega) = {} & [a_0 S(\omega) \;\; a_1 S(\omega) \exp^{-j\omega\tau_{0,1}} \;\; a_2 S(\omega) \exp^{-j\omega\tau_{0,2}} \;\; \cdots \;\; S(\omega) \exp^{-j\omega\tau_{0,M-1}}]^T \\ & + \;[N_0(\omega) \;\; N_1(\omega) \;\; N_2(\omega) \;\; \cdots \;\; N_{M-1}(\omega)]^T \end{aligned} \tag{2.7}$$

Equation 2.7 can be written as:

$$\mathbf{X}(\omega) = S(\omega)\; \mathbf{A}(\omega) \; + \; \mathbf{N}(\omega) \tag{2.8}$$

where $S(\omega)$ is the Fourier transform of the source signal $s(t)$, ω is the frequency of the source signal, $\mathbf{N}(\omega) = [N_0(\omega)\; N_1(\omega)\; N_2(\omega) \; \cdots \; N_{M-1}(\omega)]^T$ is the Fourier transform of the noise vector and $\mathbf{A}(\omega)$ is the array response vector also called the steering vector or direction vector [116]. The steering vector is given by:

$$\mathbf{A}(\omega) = [a_0 \;\; a_1 \exp^{-j\omega\tau_{0,1}} \;\; a_2 \exp^{-j\omega\tau_{0,2}} \;\; \cdots \;\; a_{M-1} \exp^{-j\omega\tau_{0,M-1}}]^T \tag{2.9}$$

In far-field where the distance between the adjacent microphones in the array is very small compared to their distance from the signal source, with identical microphones in

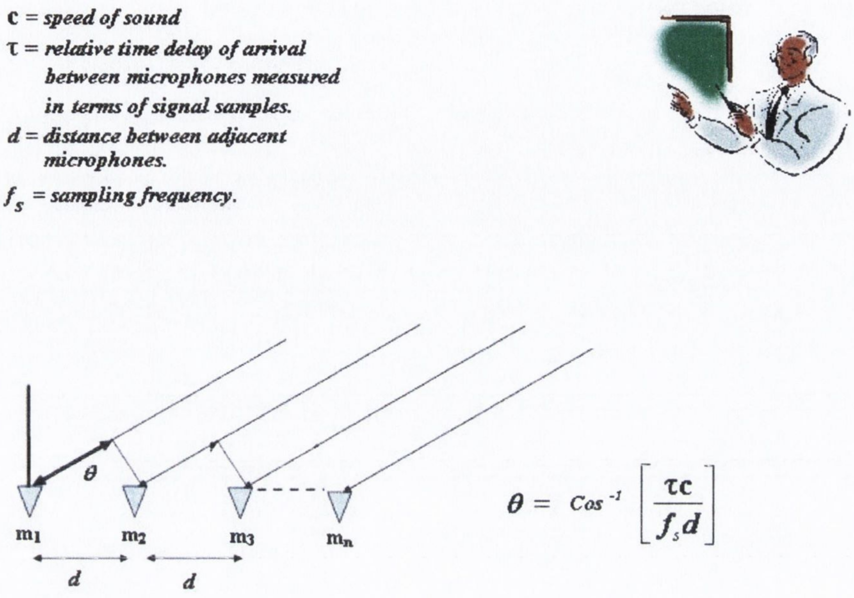

c = speed of sound
τ = relative time delay of arrival between microphones measured in terms of signal samples.
d = distance between adjacent microphones.
f_s = sampling frequency.
θ
m1
m2
m3
mn
d
d
$\theta = Cos^{-1}\left[\frac{\tau c}{f_s d}\right]$

Fig. 2.1: Direction Of Arrival (DOA) estimation.

the array the gain value can be assumed to be unity and $A(\omega)$ can be written as [115, pg. 12]:

$$\mathbf{A}(\omega) = [1 \quad \exp^{-j\omega\tau_{0,1}} \quad \exp^{-j\omega\tau_{0,2}} \quad \cdots \quad \exp^{-j\omega\tau_{0,M-1}}]^T \tag{2.10}$$

After a description of the signal model, the formula to find the direction of arrival of the signal based on the far-field assumption is discussed in the next section.

2.2 Angle of arrival

This section presents the formula employed to estimate the Direction Of Arrival (DOA) based on the far-field assumption for known value of the Time Delay Estimate (TDE). Figure 2.1 shows a signal source and a uniform linear array (ULA) consisting of M microphones d cm apart. In an ULA, all the microphones are on the same plane and adjacent microphones are equidistant spaced. The geometry of the microphones in the array plays an important role in the determination of the source location.

The signal emanating from a source is a spherical wavefront. As the wavefront move away from the source its curvature decreases and it becomes planar. ULA is referred as microphone array in this thesis. Consider a microphone array consisting of M elements spaced d cm apart. According to the far-field assumption if the distance between the microphone array and the signal source is very large compared to the distance between microphone pair (d), then the signal wavefront received by the microphones in the array can be assumed to be planar [102, pg. 13], [47, pg. 43], [86]. The additional distance the wavefront covers to reach a microphone compared to the other is $dcos(\theta)$ [109]. The relative time difference ($t_{i,j}$) for a signal to reach a pair of microphones (i, j) traveling with a velocity of $c\ m/s$ is given by:

$$t_{i,j} = \frac{d_{i,j}\ cos\ \theta}{c} \tag{2.11}$$

The value of c is approximately 342 m/s [102]. The direction of arrival $\theta_{i,j}$ of the signal for a pair of microphone based on the far-field assumption is given by:

$$\theta_{i,j} = \cos^{-1}\left(\frac{t_{i,j}\ c}{d_{i,j}}\right) \tag{2.12}$$

The estimated $\theta_{i,j}$ is an approximate value of the DOA of the source signal. Relative time difference, $t_{i,j}$, can be expressed in terms of the signal samples ($\tau_{i,j}$) and the sampling frequency (f_s) by:

$$t_{i,j} = \frac{\tau_{i,j}}{f_s} \tag{2.13}$$

By substituting equation 2.13 in equation 2.12, the DOA of the source signal is found by the equation:

$$\theta = \cos^{-1}\left(\frac{\tau_{ij}\ c}{f_s\ d_{ij}}\right) \tag{2.14}$$

As the distance between the ULA and the source increases compared to the distance between the microphone pairs in the array, the error in the DOA estimation due to the far-field assumption decreases. The error in the DOA estimate is also dependent on the closeness of the relative time difference in terms of the signal sample.

The next section presents the restriction applicable on the microphone array based on the type of signal for time delay estimation.

2.3 Spatial Aliasing

Signals can be broadly grouped into wideband signals and narrowband signals. Wideband signals are also known as broadband signals. They do not have any particular characteristic wavelength as their power is distributed over a wide frequency range. The ratio of their highest frequency to the lowest frequency component is relatively large. Narrowband signals on the other hand assume that the signal has a very narrow bandwidth [45] and thus have a nominal wavelength. Speech signal are broadband signals as their power is distributed over a wide range of frequencies [31].

When phase information is used to find the DOA of a narrowband signal the distance between the microphones in the array is restricted by the frequency of the signal. This is due to spatial aliasing. Phase lag or lead between a signal and its shifted version can be found if the phase shift is in the range of $[-\pi, +\pi]$. If the phase shift does not fall within the range, wrapping in the phase occurs that makes it impossible to distinguish whether the phase shift is a lead or lag in phase. Figure 2.2 shows a sine wave that has a phase lead of $\frac{4\pi}{3} = \pi + \frac{\pi}{3}$ and sine wave that has a lag of $\frac{2\pi}{3} = \pi - \frac{\pi}{3}$. Both the waves are same and compared to the original signal it is difficult to find if there is a lead of $\frac{4\pi}{3}$ or a lag of $\frac{2\pi}{3}$ in the phase of the shifted signal. This results in incorrect DOA estimation.

For the broadband signal let λ_{min} correspond to the wavelength associated with the maximum frequency (f_{max}) of the signal. The phase shift is restricted by the relation:

$$2\pi f_{max} t \leq \pi \qquad (2.15)$$

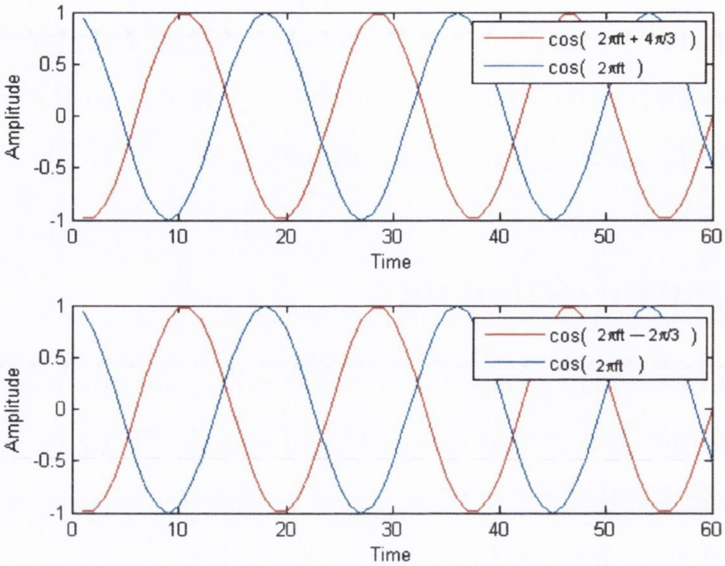


Fig. 2.2: Example of Spatial Aliasing

where t is the time delay. Using the relation between t and $cos\ \theta$ of equation 2.11, equation 2.15 can be rewritten as:

$$2f_{max}\frac{dcos\ \theta}{c} \leq 1 \tag{2.16}$$

where d is the distance between any microphone pair in the array for which the phase shift is being computed and c is the velocity of sound. The maximum value of $cos\ \theta = 1$ at $\theta = 0$. Substituting $cos\ \theta = 1$ in equation 2.16 can be rewritten as:

$$\frac{2f_{max}d}{c} \leq 1 \tag{2.17}$$

But $\frac{c}{f_{max}} = \lambda_{min}$. Hence equation 2.17 can be rewritten to provide a relation between the distance between any microphone pair in the microphone array and the wavelength of the signal. This relation is given as:

$$d \leq \frac{\lambda_{min}}{2} \tag{2.18}$$

For narrowband signals if the distance between the microphone pair, for which the phase shift is being computed for time delay estimation, in a microphone array is greater than $\frac{\lambda_{min}}{2}$ then spatial aliasing occurs [115]. Spatial aliasing leads to incorrect DOA estimation in the algorithms using TDE based on relative phase difference of the signal. To avoid this problem the distance (d) between the adjacent microphones in the array is restricted by the equation 2.18 [115, 91].

The above assumption is not applicable to the TDE on broadband signals using the cross-correlation method since the signal processing is done in the time domain [115]. Section 2.10 discusses cross-correlation based TDE algorithms for DOA estimation. Next section describes the methods applied to improve the resolution of time delay estimates for finer DOA estimates.

2.4 DOA Resolution

As shown in equation 2.14 of section 2.2, DOA estimate using far-field assumption formula involves the estimation of the relative time difference for the signal to reach a pair of microphones. The time difference is expressed in terms of the signal samples ($\tau_{i,j}$). Time Difference in signal samples is referred to as time delay ($\tau_{i,j}$) in this work. Relation between the time difference, time delay and sampling frequency is given by:

$$t_{i,j} = \frac{\tau_{i,j}}{f_s} \tag{2.19}$$

where f_s is the sampling frequency.

Thus $t_{i,j}$ is expressed in terms of the integer values of the signal sample ($\tau_{i,j}$) depending on the sampling rate f_s of the signal. The DOA formula based on the far-field assumption given by equation 2.14 in section 2.2 is computed using $cos^{-1}(.)$ operation. The number of possible DOA estimates is limited to the angle range of $[0,\ \pi]$ based on the range of cos function of $[1, -1]$. The number of time delays can be increased to obtain finer angular resolution by increasing the sampling frequency of the signal, by upsampling or by increasing the distance between the microphone pair. Increasing the sampling rate of the signal or upsampling the signal to increase the sampling rate leads to higher storage and computational requirements. The distance between the microphones is restricted to satisfy the far-field assumption which states that the distance between the microphone pair in the microphone array should be very less compared to the distance between the signal source and the microphone array.

Interpolation is also used to improve the DOA resolution [32, 26]. This method is useful in the Time Difference Of Arrival (TDOA) based algorithms for DOA estimation discussed in section 2.10. In the case of TDOA based algorithms after initial estimate of the time delay in terms of integer value of the signal sample is obtained, parabolic interpolation is applied to obtain the time delay in terms of non-integer value of the signal sample. Parabolic interpolation is not computationally intensive. In this way DOA resolution is improved without increased storage requirements and with a small increase in the computation requirement.

The next section presents Lagrange's parabolic interpolation method that is suitable for the algorithms discussed in section 2.10.

2.5 Parabolic Interpolation

Interpolation is useful in obtaining the approximate intermediate values between known points of a function by modeling it by a logical functional form [89, pg. 99]. Given N known points $(x_1, f(x_1)), (x_2, f(x_2)), \cdots, (x_N, f(x_N))$ the Lagrange's classical formula for interpolation to obtain the unknown value of a function $(f(x))$ at a point x is given by [85]:

$$f(x) = f(x_1)\frac{(x-x_2)(x-x_3)\cdots(x-x_N)}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_N)} + f(x_2)\frac{(x-x_1)(x-x_3)\cdots(x-x_N)}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_N)} + \cdots + f(x_N)\frac{(x-x_1)(x-x_2)\cdots(x-x_{N-1})}{(x_N-x_1)(x_N-x_2)\cdots(x_N-x_{N-1})} \tag{2.20}$$

A function closely related to a polynomial can be interpolated using the polynomial as the modeling function. The order of the polynomial used in interpolation is given by the number of points minus one. The parabolic interpolation or Lagrange's interpolation with three points is expressed as:

$$f(x) = f(x_1)\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} + f(x_2)\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} + f(x_3)\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)} \tag{2.21}$$

where $f(x)$ is the interpolated value of the function for the desired value of x and $f(x_1)$, $f(x_2)$, $f(x_3)$ are the already known values of the function for the values of x_1, x_2, x_3 respectively. Parabolic interpolation is also known as second order, quadratic interpolation or three point interpolation.

For time delay estimation, parabolic interpolation has been used in [32, 36] to improve the resolution of the obtained direction of arrival by finding the fractional sample delays. Parabolic interpolation is appropriate for the function involved in the time difference of arrival based algorithms for DOA estimator discussed in section 2.10 since they are close in form to a parabola around their maximum or minimum values. In this work parabolic interpolation has been used to improve the resolution of the DOA estimates by finding the delay that is a non-integer signal sample value.

Next section discusses some of the problems encountered in a classroom environment.

2.6 Classroom Environment

Classroom environment play an important role in the performance of a Direction Of Arrival (DOA) algorithms. Two main problems encountered in a classroom are the noise in the room and the room reverberation. Noise refers to the ambient (background) noise present in the room. This noise is due to the presence of mechanical equipments such as the heating system, HVAC and ventilation system in the classroom. Other equipments that contribute towards the background noise are the computers and projectors.

Presence of background noise affects the speech intelligibility. In the presence of high background noise, speech intelligibility suffers which makes it difficult for students to understand the lecture. This situation is worsened for students that have certain hearing disability.

Intensity measures the perceived loudness of the sound waves. Sound intensity is measured in decibel (dB). dB scale is logarithmic. Signal to Noise Ratio (SNR) is the ratio of the signal power to the noise power present in a classroom. It enables to estimate the understandability of the speech in a room. Both signal and noise power are measured in dB and hence SNR is the difference between the signal and noise power in dB [100].

The signal to noise ratio is given by:

$$SNR = 10 \log_{10}\left(\frac{P_s}{P_n}\right) \tag{2.22}$$

where P_s and P_n are the source signal and noise power respectively [91].

SNR of the room varies in the room due to different levels of signal and noise at different locations in the room. SNR is lowest at two locations in a classroom. One of the locations is the back of the classroom which is farthest from the lecturer's or teacher's location. The other location is nearest to the source of background noise where background noise is the maximum. A technical report by Acoustical Society of America [100] recommends a SNR to be above 10 dB for speech intelligibility and above 15 dB for children with certain hearing disability. American Speech-Language-Hearing-Association (ASHA) [4] guidelines recommend SNR to be greater than 15 dB.

Reverberation is another problem in a room. Reverberation is the echo of the signal [100]. When the signals in a room come into contact with the surfaces of the room they are absorbed, transmitted, diffused or reflected. The reflected and diffused signals interfere with the next words that are spoken. Depending on the degree of interference the speech intelligibility varies. Acceptable reverberation level in a room is measured in terms of reverberation time. Reverberation time (RT60) is a measure of the time it takes for the sound to decay to 60 dB of its original value. Volume as well as the surface materials of the room influence RT60 of the room. A small amount of reverberation is desirable as it helps in the propagation of useful sounds in the classroom. Recommended reverberation time is 0.4 - 0.6 seconds [100]. Bistafa et al. [23] recommend 0.4 - 0.5 seconds of RT60 for 100% intelligibility.

Following sections present state of the art for the Direction Of Arrival (DOA) algorithms for source localization.

2.7 Direction Of Arrival Algorithm Taxonomy

According to [24] source localization algorithms can be classified into three algorithm categories based on the signal processing approach for DOA estimation. They are:

1. Steered Beamformer based algorithms.

2. High Resolution Spectral Estimation based algorithms.

3. Time Difference of Arrival (TDOA) based algorithms.

In section 2.8, steered beamformer based algorithms are explained and a discussion of their role in the DOA estimation is presented. This is followed by an explanation and discussion of high-resolution spectral estimation based algorithms and the Time Difference Of Arrival (TDOA) algorithms.

2.8 Steered Beamformer Based Algorithms

Beamforming involves focusing the microphone array to capture signals from certain directions in presence of interfering signals. It involves spatial filtering where the spatial filter of the array is focused towards the desired direction through the signal processing algorithms rather than physically focusing the array in the desired direction [60, pg. 112], [116]. By spatial filtering two signals consisting of same frequencies are separated provided that they are generated in different locations [116]. The beam in the desired direction is formed by summing weighted signals of the microphones in the array [46, pg. 237]. Beamforming is mainly used in communications systems. Antenna arrays are used for beamforming. The concept of beamforming has been extended to the field of speech signal processing. The applications of beamforming include DOA estimation, steering a null in the direction of the undesired signal and signal enhancement [115].

Steered beamformer based algorithms are used to find the DOA of the source signal and also for signal source localization. The steered beamformer based algorithms utilize the Maximum Likelihood (ML) principle. The source localization by these algorithms depends upon maximizing the steered response power of a beamformer. A beam is steered in all the possible directions and the output power is calculated for every direction. The angle, at which maximum output power is obtained, is taken as the DOA of the signal source.

DOA estimation using beamforming is performed on narrowband signals. For DOA on broadband signals the signal is divided into smaller frequency bands. Narrowband and broadband signals are introduced in section 2.3. Each frequency band is considered to be a narrowband signal on which the DOA estimation is performed. The results of

DOA of these frequency bands is combined to get the overall DOA estimate [115].

Consider a narrowband signal similar to the one given in equation 2.6 captured through a ULA of M elements. The output of a beamformer for the signal is given by:

$$y(t) = \sum_{i=0}^{M-1} w_i^* \; x_i(t) \tag{2.23}$$

where $x_i(t)$ is the signal captured by i^{th} microphone in the microphone array and w_i is the corresponding complex weight. The weight performs the spatial filtering such that the signal from a particular direction also called the look direction is emphasized [91, 45].

Equation 2.23 can be written as:

$$y(t) = \mathbf{w}^H \; \mathbf{x}(t) \tag{2.24}$$

where $\mathbf{w}$ is the complex weighting vector and $\mathbf{x}(t)$ is the microphone array signal vector.

When the signal is a zero mean stationary process the mean output power of the beamformer for a given weight vector is [45]:

$$P(\mathbf{w}) = E[y(t) \; y^*(t)] \tag{2.25}$$

$E[.]$ is the expectation operator.

Speech signals change rapidly and hence are not stationary. The signal is divided into smaller sections called frames. The signal when divided into smaller frames can be considered to be stationary. Thus the mean output power for the signal with L frames is given by [91]:

$$\begin{aligned} P(\mathbf{w}) &= \frac{1}{L}\sum_{t=1}^{L} |y(t)|^2, \\ &= \frac{1}{L}\sum_{t=1}^{L} \mathbf{w}^H \; \mathbf{x}(t) \; \mathbf{x}^H(t) \; \mathbf{w}, \\ &= \mathbf{w}^H \; \mathbf{R} \; \mathbf{w} \end{aligned} \tag{2.26}$$

where $\mathbf{R} = \frac{1}{L}\sum_{t=1}^{L} \mathbf{x}(t)\ \mathbf{x}^H(t) = E[\mathbf{x}(t)\ \mathbf{x}^H(t)]$ is the correlation matrix of the array.

For the DOA estimation the angle at which the output power of the array is maximum is taken as the DOA of the source signal. The weights vector distinguishes between different types of beamformers.

In this section, beamforming in the time and frequency domain is presented. Initially the delay and sum beamformer is discussed. It is an example of the beamforming in the time domain. Beamforming in the frequency domain is described next.

2.8.1 Delay and Sum Beamformer

The delay and sum beamformer (DSB) is one of the oldest and the simplest beamformers [60, pg. 112]. DSB is shown in Figure 2.3. It is also known as the conventional beamformer. As the name suggests in this approach the signals captured by the microphones in the microphone array are delayed by the sample values corresponding to the relative time delay with respect to the reference microphone signal in the microphone array and added to form the DSB. The aim of DSB is to reinforce the signal arriving from a particular direction by delaying the signals captured by microphones in the microphone array by the appropriate sample values. The delay value in terms of signal samples depend on the time taken by the signal wavefront to travel from the source to the microphones in the microphone array [60, pg. 112]. The mean output power of this beamformer in the direction of the source is equal to the source power. In this beamforming process instead of mechanically steering the array in the direction of the signal source it is steered electronically by adjusting the phase of the signal. Moreover in the absence of interfering signals from other directions and the presence uncorrelated noise maximum possible signal to noise ratio (SNR) obtained using this beamformer [45]. For the DOA estimation the delayed versions of the signals captured by different microphones in the microphone array is added to the reference microphone signal to form beams focusing in different directions. The delays (in terms of signal

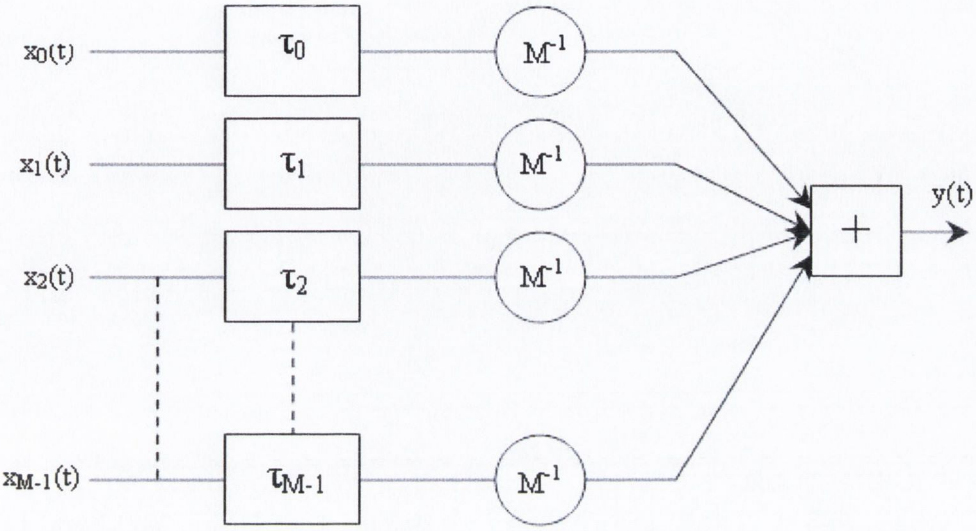


Fig. 2.3: Delay and Sum Beamformer.

sample) are dependent on the relative time difference of arrival of the signals between the microphones in the microphone array [24, pg. 159]. The output of a DSB for signal arriving from a particular direction θ is given by:

$$y(t_\theta) = \sum_{i=0}^{M-1} w_i^* \; x(t - \tau_{i,\theta}) \tag{2.27}$$

where $y(t_\theta)$ is the output of the delay and sum beamformer(DSB) for angle θ, w_i is the weight associated with the i^{th} microphone and $x(t - \tau_{i,\theta})$ is the signal received by i^{th} microphone delayed by $\tau_{i,\theta}$ with respect to the reference microphone.

The power of the DSB for a frame of length K is given by:

$$P(\theta_i) = \frac{1}{K} \sum_{t=0}^{K-1} |y(t_{\theta_i})|^2 \tag{2.28}$$

The output power vector for the possible k angles is given by:

$$\mathbf{P}(\Theta) = [P(\theta_0), P(\theta_1), \; .... \; , P(\theta_{k-1})] \tag{2.29}$$

The DOA of the source signal corresponds to the angle associated with maximum value of power in the $\mathbf{P}(\Theta)$ vector.

The DSB can be used on broadband signals. It estimates the location of the signal source by estimating the likelihood of the presence of a signal source at a particular location based on the energy measures obtained for that location. Birchfield [21] explains that the DSB equation consists of two terms. The first term is a measure of the pair-wise similarity between the signals received in the array and the second term represents the total energy of all the signals. DSB is similar to Bayesian formulation in terms of maximizing the likelihood of the signal source being present at a particular location. The difference between the two is in the weights used for the energy terms. For stationary signals, the energy term does not affect the likelihood of the presence of the source at a particular location. This results in equal DSB and Bayesian equations.

For a DSB, the resolution in the estimated angles is dependent on the sampling rate of the signal. To get a higher resolution in angle the sampling rate of the signal has to be increased. This is because the delay is in integer values of the signal sample. Higher sampling rate translates to higher storage requirement and increased processing power. To overcome the problem of improved resolution in angle without increasing the sampling frequency, frequency domain beamforming is employed [91]. Frequency domain beamforming is described in the next section.

2.8.2 Frequency Domain Beamforming

As mentioned at the beginning of section 2.8, frequency domain beamforming is performed on narrowband signals. The frequency domain representation of a narrowband signal with central frequency ω_c given in equation 2.8 is given by:

$$\mathbf{X}(\omega_c) = S(\omega_c)\mathbf{A}(\omega_c) \ + \ \mathbf{N}(\omega_c) \tag{2.30}$$

The corresponding beamformer output is:

$$\mathbf{Y}(\omega_c) = \mathbf{w}^H \mathbf{X}(\omega_c) \tag{2.31}$$

where $\mathbf{w}$ is the complex weight vector given by $\mathbf{w} = [w_0, w_1, ..., w_{M-1}]^T$ and M is the number of elements in the microphone array. The choice of weight vector is based on the requirement that all the signals for the look direction are summed coherently [91, 115] and the gain is high for signals in look direction whereas the signals from other directions are attenuated [115].

The power spectral density (PSD) of the beamformer is:

$$\begin{aligned} \Phi_{\mathbf{YY}}(\omega_c) &= \mathbf{Y}(\omega_c)\ \mathbf{Y}^*(\omega_c), \\ &= \mathbf{w}^H\ \mathbf{X}(\omega_c)\ \mathbf{w}^H\ \mathbf{X}^*(\omega_c), \\ &= \mathbf{w}^H\ \Phi_{\mathbf{XX}}\ (\omega_c)\ \mathbf{w} \end{aligned} \tag{2.32}$$

The DOA of the source signal is taken as the direction for which the value of $\Phi_{\mathbf{YY}}(\omega_c)$ is maximum. $\Phi_{\mathbf{XX}}(\omega_c)$ is the $M \times M$ PSD matrix of the input signals.

Since beamforming was developed for applications involving narrowband signals, in order to perform frequency domain beamforming on a broadband signal like the speech signal, the spectrum for the signal received by each microphone in the microphone array is obtained by Discrete Fourier Transform (DFT). The signals are then converted to narrowband signal by dividing them into smaller frequency bins. Individual narrowband spectrums with the corresponding frequency are multiplied by the suitable complex weight and summed.

The resolution in angle using a beamformer is dependent on the number of microphones in the microphone array. The maximum angle resolution that can be obtained with a ULA consisting of M elements, in radians is given by:

$$\delta_\theta = \frac{2\pi}{M}\ radians \tag{2.33}$$

While accuracy of the DOA estimates using beamforming is adequate, finer resolution in angles is obtained for higher sampling rates. Higher sampling rate requires more computational power rendering it unsuitable for real-time implementation [91]. Moreover since beamforming was developed for narrowband signals, the performance of the beamformer on broadband signals such as speech deteriorates in the presence of reverberation [91].

The performance of the beamformer for DOA estimation can be improved by using *a priori* information. Duraiswami et al. [95] have exploited the *a priori* information about the enclosed space dimensions and its relationship with the signal wavelength to perform coarse-to-fine both in the frequency and space for efficient source localization beamforming.

Discussion

Delay and sum beamformer appears to be very practical for the purpose of source localization in e-learning environment due to its simplicity. The drawback with this approach is that finer angular resolution cannot be obtained without increasing the sampling rate of the signal. Increased sampling rate leads to higher storage and computational requirements making it impractical for real time applications.

Beamforming in the frequency domain provides improved resolution in the angle estimates. Since they were proposed for narrowband signals, more computation is required to make them suitable for broadband signals such as speech. The performance of the beamforming algorithms deteriorates in the presence of reverberation [91]. The algorithm based on the coarse to fine search [95] requires *a priori information* that is not known beforehand rendering it unsuitable for the purpose of source localization.

Next section presents the second category of the Direction of Arrival (DOA) algorithms known as the high resolution spectral estimation based algorithms. The section starts with an introduction to the algorithms of this category a discussion about the advantages and disadvantages of the algorithms of this category with respect to its application for DOA estimation in an e-learning environment.

2.9 High Resolution Spectral Estimation Based Algorithms

The algorithms belonging to this category are suitable for narrowband signals. These algorithms are applied on broadband signals like speech similar to the way frequency domain beamforming is applied on the broadband signal. As mentioned in section 2.8.2 broadband signal is transformed to obtain its spectrum using the DFT. The signal spectrum is divided into several narrowband frequency bins [91]. The narrowband algorithm is applied on each narrowband frequency bin and then the results of all the narrowband frequency bins are combined to obtain an overall result for the broadband signal. Subspace based algorithms or eigen structure methods are prominent examples of the high resolution spectral estimation based algorithms category.

2.9.1 Subspace Techniques

In this section a brief introduction is provided on the subspace techniques. The algorithms based on the subspace techniques are based on two properties of the Cross-Correlation matrix of the signal. The first property states that the space spanned by the eigenvectors of the signal can be divided into two orthogonal subspaces namely the signal subspace and the noise subspace. The second property states that the steering vector corresponding to the directional sources are part of the signal subspace since they too are orthogonal to the noise subspace [45].

For a narrowband signal with center frequency ω_c the cross-correlation matrix is a

$M \times M$ matrix given by:

$$\mathbf{R}(\omega_c) = \mathbf{X}(\omega_c)\ \mathbf{X}^H(\omega_c) \tag{2.34}$$

Assuming that the rank of $\mathbf{R}$ is full and the number of sources present (P) is less than the number of microphones in the array (M). The eigen decomposition of the $\mathbf{R}$ is expressed as [91]:

$$\mathbf{R} = \sum_{i=1}^{M} \lambda_i \mathbf{e}_i \mathbf{e}_i^H \tag{2.35}$$

where λ_i is the i^{th} eigenvalue and $\mathbf{e}_i$ is the i^{th} eigenvector.

The eigen decomposition of $\mathbf{R}$ results in a P-dimensional signal subspace and $(M - P)$-dimensional noise subspace. The eigenvalues are arranged in the descending order. The highest P eigenvalues represent the signal subspace and their corresponding eigenvectors span the signal subspace. The rest of the eigenvalues represent the noise subspace which is spanned by their corresponding eigenvectors.

These algorithms search for the directions for which the corresponding steering vectors are orthogonal to the noise subspace and belong to the signal subspace [91]. The length of the signal frame should be large enough to ensure that the rank of R is full. Small frame size can lead to error in the detection of the number of sources and estimation of the DOA [91].

A discussion about the suitability of the algorithms of the high resolution spectral estimation based algorithm is presented next.

Discussion

The algorithms of this category are found to be very efficient for DOA estimation especially in the presence of multiple sources. The main drawback of these algorithms is that they are meant for narrowband signals. They can be applied on the broadband signals such as speech. The spectrum of the speech signal is divided into smaller frequency band to form narrowband signal and the algorithms are performed on the individual bands. The results of all the bands are combined to get the DOA of the speech signal. The algorithms are computational intensive. The computational requirements are further increased to facilitate their application on the speech signal. Furthermore the performance of these algorithms deteriorates in the presence of noise and reverberation. Thus the computation requirement along with a poor performance in presence of noise and reverberation makes the algorithms of this category less appealing for real time applications such as e-learning.

Next section discusses the third category of source localization algorithms known as the time difference of arrival (TDOA) based algorithms. These algorithms are suited for broadband signal processing.

2.10 Time Difference Of Arrival Based Algorithms

In case of Time Difference Of Arrival (TDOA) based algorithms the DOA is carried out in two steps. First the Time Delay Estimation (TDE) is performed on the signals received by the microphones in microphone array. TDE is computed on various combinations of microphone pairs in the microphone array. The second step involves using the TDE information and the knowledge of array geometry for source localization. Compared to the algorithms of the other two categories of DOA algorithms discussed in section 2.8, 2.9, these algorithms are not complex to implement and are computationally efficient [32]. They are also computationally practical for real time applications. The main disadvantage of the TDOA algorithms is that the presence of a single source is assumed for DOA [24, pg. 163]. Based on the e-learning scenario discussed in section 1.1 this disadvantage in not of great significance as it is expected that for communication only one person is the primary talker at any given instant. Thus they are suitable for source localization in e-learning.

From the literature Time Delay Estimation (TDE) can be divided into two prominent categories of algorithms based on their approach for TDE estimation. They are the Generalized Cross-Correlation (GCC) based algorithms and the Average Magnitude Difference Function (AMDF) based algorithms. Before describing the GCC based algorithms and AMDF based algorithms, cross-correlation that is a part of most the TDE algorithms is discussed in the next section.

2.10.1 Cross-correlation

The cross-correlation algorithm finds the relation between two different signals. Based on the signal model presented in section 2.1, signals received by a pair of microphones in the microphone array can be represented as:

$$x_1(t) = s_1(t) + n_1(t) \tag{2.36}$$

$$x_2(t) = a\ s_1(t - \tau) + n_2(t) \tag{2.37}$$

where $x_1(t)$ is the signal received by the reference microphone that consists of the source signal $s_1(t)$ corrupted by the additive noise $n_1(t)$. Similarly $x_2(t)$ is the signal received by the second microphone consisting of a time shifted version of the source signal $s_1(t)$ scaled by a factor a and corrupted by the additive noise $n_2(t)$. The noise signals $n_1(t)$ and $n_2(t)$ are assumed to be uncorrelated to each other and to the source signal.

The cross-correlation of two microphone signals $x_1(t)$ and $x_2(t)$ is given by:

$$R_{12}(\tau) = \sum_{t=0}^{K-1} x_1(t)x_2(t + \tau) \tag{2.38}$$

where K is the length of the signal frame, $\tau = 0, 1, 2, \cdots$ and $R_{12}(\tau)$ is the cross-correlation value at a delay τ. The cross-correlation ($R_{12}(\tau)$) is found for different delay τ values. All the $R_{12}(\tau)$ for different τ values form the Cross-Correlation Function

(CCF). Cross-Correlation Function (CCF) denoted by (R_{CCF}) can be found by the Fourier transform and is given by:

$$R_{CCF} = F^{-1}[X_1(k)X_2^*(k)] \tag{2.39}$$

where $X_1(k)$ and $X_2(k)$ are the Fourier Transforms of $x_1(t)$ and $x_2(t)$ respectively and $F^{-1}[.]$ is inverse Fourier transform operation. The term $X_1(k)$ $X_2^*(k)$ is known as the cross-spectrum of the two signals [114]. Maximum Likelihood (ML) principle is applied on the CCF to identify the delay (τ) at which the two signals resemble each other the most. According to the ML principle the value of the τ that corresponds to the peak value of the CCF is the estimated time delay. The Cross-Correlation Function (CCF) is sensitive to noise signals [79] and amplitude changes [36].

Since in real time implementation the signal frame size is fixed thus the overall cross-correlation function is biased towards zero delay. This is because the shift in the second signal with respect to the first signal, ends up in an overall cross-correlation term, associated with a particular τ (other than zero shift), with less number of terms that are multiplied and added. As the shift moves away from the zero delay the cross-correlation value tends to decrease as the number of terms multiplied and added to the function decreases.

The simplest method to reduce the effect of bias is to use signal frames which are long compared to the total number of delay values. However, to get an accurate estimate τ it should be ensured that the signal is statistically stationary. This is not possible for very long frames. Thus there is a trade-off between the degree of bias of the cross-correlation algorithm and signal frame length. In addition, longer signal frames cause computational load to be significantly increased since the order of computations is *O(*K^2*)* where K is the size of the signal frame. The complexity can be reduced by a frequency domain implementation that reduces the complexity to *O(K* log_2 *K)* [114].

Another way to reduce the effect of bias is to normalize the $R_{12}(\tau)$ function by dividing it by the difference of the signal frame size and the τ value associated with the particular shift. This method is known as *unbiased cross-correlation* and can be written as:

$$R'_{12}(\tau) = \frac{R_{12}(\tau)}{(K - |\tau|)} \; where \; 0 \leq |\tau| < K \tag{2.40}$$

After presenting a discussion on the cross-correlation of the signal, Generalized Cross-Correlation (GCC) based Time Delay Estimation (TDE) algorithms are discussed in the next section.

2.10.2 GCC based algorithms

One of the simplest way to find the time delay τ is to calculate the cross correlation of the signals received by pair of microphones. The delay value corresponding to the maximum value of the cross correlation function gives the relative time difference of arrival of the signal for that microphone pair. Even though two microphones are enough to estimate the DOA of the source by this approach its performance deteriorates in presence of noise and reverberation. More microphones can be used to increase the robustness and resolution of the result [114] and for source localization.

Since the adjacent speech samples are highly correlated, the cross-correlation function peak obtained is very wide [114]. A wide peak does not pose a problem in case of a single delay but when the signal has multiple delays (in presence of more than one source or reverberation) wide peaks are not desirable. The cross-correlation peak in the presence of many delays can spread into each other to produce a cross-correlation function with broad local maximum. This can result in incorrect TDE since the peak corresponding to the actual delay can be overshadowed by other peaks [70], [24, pg. 167].

For better time-delay resolution sharp peaks are desirable. Fixed length signal frames introduce errors in sharp peaks in case of low signal-to-noise ration (SNR). To overcome this problem weighting function or pre-filters for the cross-spectrum of the cross-correlation function have been proposed by Knapp et al. [70, 29]. The cross-correlation function obtained from the inverse fourier transform of the pre-filtered cross-spectrum is known as the Generalized Cross-Correlation (GCC). These weighting functions sharpen the peaks of the cross-correlation function.

The normalized cross-correlation function can be expressed as:

$$R_{12}(\tau) = F^{-1}[\psi(k)X_1(k)X_2^*(k)] \tag{2.41}$$

where $\psi(k)$ represents the weighting function.

The prominent effective weighting functions proposed for GCC are presented below:

1. **The Roth Processor**: This weighting function was proposed by Roth. It is give by:

   $$\psi(k) = \frac{1}{X_1(k)X_1^*(k)} \tag{2.42}$$

   The Roth processor suppresses the frequency regions where the noise is large as weights are assigned to the correlation function in accordance with the characteristics of the noise and signal [70]. The problem with this approach is that the *a priori* information about the noise and signal statistics is unavailable in many applications [29]. The problem of source localization in a room is an example of the application where *a priori* information regarding the signal and noise statistics is unavailable.

2. **Smoothed Coherence Transform (SCOT)**: Carter et al. [30] proposed the SCOT weighting function. It was proposed for TDE to localize an underwater acoustic source in the presence of strong undesired sinusoidal interference [98]. It is given by:

   $$\psi(k) = \frac{1}{\sqrt{X_1(k)X_1^*(k)X_2(k)X_2^*(k)}} \tag{2.43}$$

   The SCOT weighting function is used to whiten the signal before finding the cross-correlation of the signal. Whitening of the signals sharpens the peak of the cross-correlation function [114]. Kuhn [72] proves with experimental results on broadband signal with different signal to noise ratio, that the SCOT prefiltering is suitable for TDE on broadband signals. When $X_1(k)X_1^*(k)$ is equal to $X_2(k)X_2^*(k)$ then this function is equal to the Roth weighting function. The prewhitening of the signal is not enough in this case to sharpen the cross-correlation peak. SCOT function suppresses frequency bands that contain higher noise content [70]. Experimental results by Scarbrough [98] show that for TDE in presence of non-sinusoidal interfering (additive uncorrelated white Gaussian noise) signals, SCOT weighting functions performance is worse compared to the unbiased cross-correlation function.

3. **Phase Transform function (PHAT)**: The PHAT weighting function involves whitening the cross spectrum of the signals. The PHAT function is given by:

$$\psi(k) = \frac{1}{|X_1(k)X_2^*(k)|} \tag{2.44}$$

This weighting function gives equal weight to all the frequency components in the signal. The TDE is based on the phase difference of the two signals. The performance of the cross-correlation due to multi-path improves by using the PHAT weighting. The main drawback of this method is that since all the frequencies including all the ones where noise dominates are given equal weight this method is less robust to noise and makes speech detection difficult [114]. This results in poor performance of the PHAT in conditions of low reverberation and high noise [24, pg. 162].

4. **The Eckart Filter**: The Eckart filter weighting function is given by [70, 29]:

$$\psi(k) = \frac{X_1(k)X_2^*(k)}{N_1(k)N_1^*(k)N_2(k)N_2^*(k)} \tag{2.45}$$

where $N_1(k)$ and $N_2(k)$ are the Fourier transform of the noise signals $n_1(t)$ and $n_2(t)$ respectively. The ratio of the change in the mean correlation output due to the presence of the signals ($s_1(t)$ and $s_2(t)$) and the standard deviation of correlation output due to noise is known as the deflection criterion. The Eckart Filter aims to maximize the deflection criterion by maximizing the numerator and minimizing the denominator. Similar to the SCOT function it tends to give less weight to frequency bands that have high noise content and zero weight to noise only frequency bands. The main drawback of this function is that the denominator includes the spectrum of the noise signals. Thus *a priori* knowledge of the signal as well noise spectrum is required to implement it [70]. The *a priori* information of the noise spectrum and the source spectrum is not known or only approximate information is available in some applications [29].

5. **The HT Processor**: The HT Processor proposed by Knapp et al. [70] is also known as the ML estimator. This assumes that the speech signal and the noise signals are Gaussian processes [70, 29]. The HT processor is given by:

$$\psi(k) = \frac{|\gamma_{12}(k)|^2}{|X_1(k)X_2^*(k)|[1 - |\gamma_{12}(k)|^2]} \tag{2.46}$$

where

$$|\gamma_{12}(k)|^2 = \frac{|X_1(k)X_2^*(k)|^2}{X_1(k)X_1^*(k)X_2(k)X_2^*(k)} \tag{2.47}$$

The ML weighting function is proposed for uncorrelated signals whereas in real time environment in presence of reverberation the noise and signals are highly correlated [108]. Moreover the performance of this weighting function degrades in reverberant environment [25]. The performance results obtained for this weighting function suggests that in the absence of reverberation this method performs well for low SNR. It performance degrades for SNR less than 10 dB. But in reverberant environment its performance degrades even at high SNR values. Carter [29] emphasizes that in order to achieve good performance for TDE, it is necessary that the weighting functions are properly designed so that including the pre-filtering process contributes towards improving the performance of the time delay estimators.

6. **Pitch based frequency weighting function**: Another GCC weighting function for TDE, proposed by Brandstein [25, 26] involves weighting the frequencies of the signal based on the signals pitch. The frequency bands containing the voiced parts of speech are emphasized whereas the parts containing high noise are given low weights. The speech signal frames maintain their periodicity in the presence of reverberation. To find the fundamental frequency of the signal a Multi-Band Excitation speech vocoder generates excitation spectrum $E(\omega)$ for different fundamental frequencies. This excitation spectrum along with the room transfer function $H(\omega)$ is used to generate speech spectrum. The difference between the actual speech spectrum $X(\omega)$ and generated speech spectrum for different fundamental frequencies is calculated. The equation for the error is given by:

$$\varepsilon = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(\omega) - H(\omega)E(\omega)|^2 d\omega \tag{2.48}$$

The error for each harmonic associated with a fundamental frequency is found by dividing the frequency band into various bands centered about the harmonics of the fundamental frequencies. The error for each harmonic is found and added to get the overall error for that fundamental frequency band. The formula to find the error for i^{th} harmonic is given by:

$$\varepsilon_i = \frac{1}{2\pi}\int_{l_{i,1}}^{l_{i,2}} |X(\omega) - A_i E(\omega)|^2 d\omega \tag{2.49}$$

where $l_{i,1}$ and $l_{i,2}$ are the lower and upper frequency limits with the i^{th} harmonic as the center and A_i is given by:

$$A_i = \frac{\int_{l_{i,1}}^{l_{i,2}} X(\omega)E^*(\omega)d\omega}{\int_{l_{i,1}}^{l_{i,2}} |E(\omega)|^2 d\omega} \tag{2.50}$$

The error value for different fundamental frequencies of interest is found and the frequency that has the minimum error value is taken as the fundamental frequency of the signal.

The closeness of the estimated fundamental frequency spectral region to the actual spectral region is found by calculating normalized error for i^{th} harmonic by:

$$E_i = \frac{\varepsilon_i}{\frac{1}{2\pi}\int_{l_{i,1}}^{l_{i,2}} |X(\omega)|^2 d\omega} \tag{2.51}$$

Low value of E_i indicate strong voiced harmonic region and high values of E_i indicate presence of high noise and non-periodic signal. The weighting function for the i^{th} harmonic with limits $l_{i,1}$ and $l_{i,2}$ is given by:

$$\psi(\omega) = \frac{(1 - max(E_{i,1}E_{i,2}))^\gamma}{|X_1(\omega)X_2^*(\omega)|} \tag{2.52}$$

where $E_{i,1}$ and $E_{i,2}$ are the normalized error for i^{th} harmonic of signal $X_1(\omega)$ and $X_2(\omega)$ respectively. The values of γ is taken between 1 and 2. By using the above mentioned weighting the signal spectral regions containing strong voiced characteristics are emphasized while the spectral regions containing noise like characteristic are deemphasized. At low SNR values the performance of this weighting function degrades especially in the presence of high reverberation. The estimation of the fundamental frequency for this method is computationally demanding as a resolution of 1 Hz in the case of higher harmonics for the fundamental frequency is required. A coarse to fine search method can be used where coarse search of the region where the fundamental frequency may be present can be identified and a fine search for the fundamental frequency can be performed in this region.

7. **GCC-CEP**: Apart from the above weighting function using cepstral pre-filtering as a weighting function to reduce the impact of reverberation has been proposed in [108] for TDE. This weighting function is known as GCC-CEP. The complex cepstrum of a signal is obtained by taking the inverse Fourier transform of the complex logarithm of the Fourier transform (frequency spectrum) of the signal. The cepstrum of a signal can be decomposed into two components namely the minimum phase component and the all pass component. Any modifications made to the all pass component effects the TDE but slight modifications to the minimum phase component does not effect the TDE. By subtracting the minimum phase component of the channel cepstrum from the signal cepstrum the TDE can be improved, based on the assumption that minimum phase component of the signal cepstrum varies for all the frames and has zero mean whereas the minimum phase component of the channel cepstrum varies slowly.

   The channel cepstrum is calculated frame by frame. An exponential window is applied on the frame before the cepstrum pre-filtering. The minimum phase components are emphasized compared to the all pass components using the exponential window since these components can be modified to reduce the effect of reverberation on the time delay estimates. The channel cepstrum is calculated recursively which is subtracted from the signal cepstrum. The resultant is transformed into time domain signal and inverse exponential window is applied. GCC is performed on this signal to get the time delay [108]. This method requires frames to have stationary speech signal which is possible for small frames but for de-convolving the effects of reverberation using the cepstrum requires long frames as the impulse response of the room can be very long. Stephenne et al. [108] have tested the GCC-CEP algorithm on the Gaussian noise as the source signal.

Discussion

As suggested by Chen et al. in [32], GCC is a prominent algorithm used for the TDE. Among the weighting functions used in TDE through the GCC algorithms, the filtered cross-spectrum obtained using PHAT weighting function involves the channel responses for the source signal without involving the signal itself [32]. For wideband signals, GCC-PHAT weighting is independent of the characteristics of the signal [86]. The performance of GCC using the PHAT weighting is better compared to using other weighting functions for signal characteristics that change in time [32], [24, pg. 167]. The characteristics of noise and the speech signal need not be modeled for PHAT weighting function.

2.10.3 Average Magnitude Difference Function (AMDF)

AMDF is another algorithm that is used for TDE [32, 58]. This algorithm involves taking the sum of the absolute value of the difference between the microphone signals at different time delays. The delay value at which the difference between the signals is minimum is taken as the estimated time delay.

AMDF is given by:

$$R_{AMDF}(\tau) = \frac{1}{K} \sum_{n=0}^{K-1} |x_1(n) - x_2(n+\tau)| \tag{2.53}$$

where K is the signal frame length. This algorithm has low computational complexity as it does not involve any multiplications. Its performance degrades in the presence of reverberation and noise [32].

Another algorithm that is used is the Average Square Difference Function (ASDF). This is a variation of the AMDF where the sum of the squares of the difference of the microphone signals for different delays is taken for TDE. The estimated time delay is equal to the delay corresponding to the minimum value of the ASDF.

ASDF is written as:

$$R_{ASDF}(\tau) = \frac{1}{K}\sum_{n=0}^{K-1}(x_1(n) - x_2(n+\tau))^2 \tag{2.54}$$

Average Mean Sum Function (AMSF) is another algorithm that takes into consideration the sum of the signals unlike the AMDF. It is given by:

$$R_{AMSF}(\tau) = \frac{1}{K}\sum_{n=0}^{K-1}|x_1(n) + x_2(n+\tau)| \tag{2.55}$$

Like AMDF this algorithms also does not involve any multiplication operations and hence has a low computational complexity. Compared to the cross-correlation algorithm, AMDF and ASDF perform well in the presence of medium to high noise [58]. The performance of AMDF deteriorates in the presence of reverberation.

Chen et al. [32] have recently proposed two TDE algorithms based on the GCC-PHAT, AMDF and AMSF algorithms. The first algorithm namely the Weighted Cross-Correlation (WCC) estimator aims to merge the robustness of the GCC algorithm and the accuracy of the AMDF algorithms. The second algorithm known as the Modified AMDF (MAMDF) combines the AMDF and the AMSF algorithms for TDE.

The WCC algorithm is given by:

$$R_{WCC}(\tau) = \frac{R_{GCC}(\tau)}{R_{AMDF}(\tau) + \epsilon} \tag{2.56}$$

where $R_{GCC}(\tau)$ and $R_{AMDF}(\tau)$ are given by the equations 2.41 and 2.53 respectively. The value of the weighting function for R_{GCC} is the PHAT function of equation 2.44. A small positive value ϵ is added to prevent division overflow. The τ value that corresponds to the minimum value of the R_{WCC} function is the estimated time delay.

MAMDF algorithm is given by the equation:

$$R_{MAMDF}(\tau) = \frac{R_{AMDF}(\tau)}{R_{AMSF}(\tau) + \epsilon} \tag{2.57}$$

where $R_{AMDF}(\tau)$ and $R_{AMSF}(\tau)$ are given by the equations 2.53 and 2.55 respectively. ϵ is a small positive value added to the denominator to prevent division overflow. The τ value that corresponds to the maximum value of the R_{MAMDF} function is the estimated time delay.

Based on the experimental results in [32], both the algorithms perform equally well in the presence of reverberation and are better than GCC-PHAT. In the presence of noise their performance is better than the AMDF algorithm.

2.10.4 Others

Apart from the algorithms based on the GCC and AMDF, hemisphere sampling is another method proposed for DOA estimation. This algorithm employs the GCC-PHAT given in equation 2.44 for TDE. The cross-correlation vectors obtained by this method for different pair of microphones are mapped to a common coordinate system. This coordinate system is called the sampled unit hemisphere whose center is the microphone array [22]. The DOA is estimated based on the accumulated sum of the mapped correlation vectors where the hemisphere's maximum cells give both the azimuthal and the elevation angles. The authors claim that the method is robust to noise and reverberation, can handle any microphone configuration and does not suffer from blind spots. Blind spot are the locations for which the DOA cannot be found [22].

More recently Talantzis et al. [110] have proposed DOA estimation based on the information theory. This algorithm estimates the DOA for a single source in a highly reverberant environment. It is based on the finding the TDE by maximizing the mutual information that one microphone has on the other in a microphone pair. For a signal source with zero-mean Gaussian distribution, the mutual information for the signal received by a microphone pair is given by:

$$MI = \frac{-1}{2} \ln \frac{det\ [C(\tau)]}{C_{11}C_{22}} \tag{2.58}$$

where $det[.]$ is the determinant operator and $C(\tau)$ is the joint covariance matrix.

The mutual information obtained is not sufficient to obtain robust TDE in reverberant environments. Thus the information between the microphones is estimated by considering jointly 'N' neighboring samples of the signals. This gives the marginal mutual information which can be used to find the TDE. For a very long frame the diagonal elements of the $C(\tau)$ are independent of τ. The relative time delay is equal to the delay value that maximizes the value of MI given in equation 2.58. In the presence of reverberation this algorithm performs better than GCC-PHAT but its performance degrades with the increase in the room reverberation.

2.11 Discussion and Conclusions

The Time Difference Of Arrival (TDOA) based algorithms are suitable for real time implementation as they are not computationally intensive. Another advantage of these algorithms is that finer resolution in angle can be obtained by applying interpolation to obtain time delay estimates in terms of fractional signal samples.

The main drawback with using the cross-correlation algorithm for Time Delay Estimation (TDE) is that it is very sensitive to amplitude changes which results in spurious Direction Of Arrival (DOA) estimates. The cross-correlation function also has wide peaks which is not desirable when the signal consists of multiple time delays and also for finer resolution in the DOA estimate. The Generalized Cross-Correlation algorithms have been proposed to provide sharper peaks by using a weighting function with the cross-correlation function. Among the proposed weighting functions the performance of the GCC-PHAT is the best in the presence of noise and reverberation. The performance of algorithms based on the Average Magnitude Difference Function (AMDF) for TDE deteriorates in the presence of reverberation. Among the recently proposed TDE algorithms, MAMDF algorithm appears to perform better than GCC-PHAT in the presence of noise and reverberation.

Steered beamformer based algorithm such as delay and sum beamformer though simple to implement requires higher sampling rate for finer resolution in the DOA estimates. This increases storage and computation requirements making it impractical for real time implementation. Beamforming in the frequency domain increases the resolution in the DOA estimates without increasing the signal sampling rate. But this method is suitable for narrowband signals. It can be applied on broadband signals like speech but this requires increased processing power. Moreover the performance of the algorithms deteriorates in the presence of reverberation.

The high resolution spectral estimation based algorithms are efficient for DOA estimation especially in the presence of multiple sources. The algorithms are computational intensive. The main drawback of these algorithms is that are suitable for narrowband signals. They are applicable on broadband signal such as speech at additional computation requirements. Furthermore the performance of these algorithms deteriorates in the presence of noise and reverberation. Thus the computation requirement along with a poor performance in presence of noise and reverberation makes the algorithms of this category less appealing for real time applications such as e-learning.

The Time Difference Of Arrival (TDOA) based algorithms are suitable for real time applications as they are not computationally intensive. They are suitable for DOA estimates in an e-learning environment since they assume presence of a single signal source which is in agreement with the classroom environment discussed in section 1.1 that during lecture there is only a single signal source. Finer resolution in the DOA estimate can be obtained by interpolation instead of upsampling the signal or increasing the sampling rate. From the discussion in section 2.8, 2.9 and 2.10, it can be concluded that for real time source localization in an e-learning environment DOA estimation using TDOA algorithms is practical.

The Generalized Cross-Correlation (GCC) algorithms have been proposed to provide sharper peaks by using a weighting function with the cross-correlation function. In the presence of noise and reverberation the performance of the GCC-PHAT is the best among the proposed GCC weighting functions. The performance of algorithms based on the Average Magnitude Difference Function (AMDF) for TDE deteriorates in the presence of reverberation. Among the recently proposed TDE algorithms, MAMDF algorithm appears to perform better than GCC-PHAT in the presence of noise and reverberation.

2.12 The YIN Algorithm

This section describes and explains the YIN algorithm. YIN algorithm [36] is used for estimating the fundamental frequency of a signal. Modification to the YIN algorithm for TDE, primary contribution of this thesis, is presented in section 4.1 of chapter 4.

The fundamental frequency of a periodic signal is equal to the inverse of its period. YIN algorithm is used to find the period of a speech or music signal which in turn is used to find the fundamental frequency of the signal. The objective of the YIN algorithm is to find the period of the signal based on the similarity of the signal and it's shifted or delayed version. YIN utilizes the difference between the actual signal and a shifted version of the signal at different time lags to find the period of the signal. It assumes that the actual signal and its shifted version will be most similar (in ideal case will be equal) at the time lag value that is the period of the signal.

The TDE algorithm aims to find the relative time difference for a signal to reach a pair of microphones in a microphone array. The solutions to the TDE problem is based on finding the time delay by measuring the similarity between the received signals at different lag values. TDE algorithms assume that the signals will be most similar (in ideal case will be equal) at the actual time delay. Due to the similarity in operation of the YIN algorithm and the TDE algorithms, it will be shown in chapter 4 how the

YIN algorithms can be used for TDE.

YIN algorithm consists of 5 steps. They are as follows:

1. Step 1: Calculate the difference function which is equal to the sum of the squares of differences of the signal and its delayed version. The difference function for the signal is given by:

$$d(\tau) = \sum_{n=0}^{K-1} (x(n) - x(n+\tau))^2 \tag{2.59}$$

   where $x(n)$ is the input signal, $x(n+\tau)$ is the input signal delayed by an integer sample value represented by τ and K is the length of the signal frame.

   The equation 2.59 when expanded gives:

$$d(\tau) = \sum_{n=0}^{K-1} x^2(n) + \sum_{n=0}^{K-1} x^2(n+\tau) - 2\sum_{n=0}^{K-1} x(n)x(n+\tau) \tag{2.60}$$

   The first two terms on the right hand side of the equation 2.60 represent the energy of the signal. The second energy term is not constant and varies with the value of τ. The final term is the autocorrelation of the signal at the lag value [36].

   With an increase in the signal amplitude with time, the peak of the autocorrelation function of the signal increases instead of remaining constant. This results in the incorrect lag estimate that will be higher than the expected lag as the peak of the autocorrelation will move away from the expected lag value to higher lag values. This problem is known as "too low" error. "Too high" error occurs when decrease in the signal amplitude with time results in reduction of the peak of the autocorrelation function of the signal instead of remaining constant. In this case the peak of the autocorrelation function will coincide with the lag value which is smaller than the expected lag value. The difference function given by equation 2.59 is insensitive to the problem of "too low" error as change in amplitude, with lag, increases the period-to-period dissimilarity [36].

2. Step 2: A problem with the difference function given by equation 2.59 is that it tends to be zero for zero lag and non-zero at the actual period value. Thus a lower limit of one lag value is used on the period search range. Another problem encountered with difference function is that a strong resonance at the first formant (refer to section 3.2) of the speech signal might lead to incorrect identification of the formant as the period [36]. To avoid both the problems the difference function is normalized to form the cumulative mean normalized difference function. The cumulative mean normalized difference function is given by:

   $$d'(\tau) = \begin{cases} 1 & (for\ \tau = 0) \\ \frac{d(\tau)}{\frac{1}{\tau}\sum_{i=1}^{\tau} d(i)} & (otherwise) \end{cases} \tag{2.61}$$

   Due to normalization of the difference function, the cumulative mean normalized difference function reduces the "too high" error. This step also eliminates the necessity to limit the higher search range of the period. The lag value corresponding to the minimum of the cumulative mean normalized difference function is taken as the period of the signal.

3. Step 3: The calculation of $d'(\tau)$ is followed by the selection of the signal period based on absolute threshold. Absolute threshold is used to reduce the error in the estimated signal period. The absolute threshold value is not constant and is decided by the user. The lag that corresponds to the minima that is below the absolute threshold value and is closest to it is taken as the period of the signal.

4. Step 4: Parabolic interpolation follow absolute threshold step. Parabolic interpolation (discussed in section 2.5) is used to improve the resolution of the periodicity measure [36].

5. Step 5: The final step involves finding the period of the signal based on the best local estimate. This step is used to further reduce the error in period estimation. The first four steps are implemented on the signal where the upper limit on the

frequency range does not exist. In the final step the cumulative mean normalized difference function is found for the signal around the pitch estimate with a restriction on the search range which is given by the range of $[t - \frac{\tau_{max}}{2}, t + \frac{\tau_{max}}{2}]$ and comes up with the best periodicity measure. τ_{max} is the largest expected period and t is the time instant at which the period of the signal is estimated [36]

This algorithm is suited for high pitched voice and music signals as no restriction is placed on the frequency upper search range [36].

2.13 Summary

In this chapter the background literature associated with the DOA algorithms was presented. These included the signal model, the formula for DOA estimation based on far field assumption, spatial aliasing, resolution of the DOA, parabolic interpolation, problems encountered in a classroom environment that interfere with the performance of DOA estimation. This was followed by a description, advantages and disadvantages of the steered beamformer based algorithms, high resolution spectral estimation based algorithms and Time Difference Of Arrival (TDOA) based algorithms. It was concluded that for e-learning environment DOA estimation from TDE is practical as they are suitable for real time implementation. Finally the YIN algorithm that is used to find the fundamental frequency of the speech and music signals was described in detail. The YIN algorithm is the basis of the primary contribution of the thesis which is a TDE algorithm that performs well in presence of noise and reverberation.

Chapter 3

State of the Art: Voice Activity Detection

This chapter presents the background literature associated with Voice Activity Detection (VAD) and the various stages in designing of a VAD system. The chapter begins with a high level overview of a VAD system. This is followed by a description of speech signal characteristics that can be used for VAD. Audio signal features used for VAD in the existing literature are presented next. The chapter ends with a description of feature selection and supervised machine learning algorithms.

3.1 Voice Activity Detection (VAD) System

Voice Activity Detection (VAD) is an important pre-processing stage to TDE in a DOA system as shown in Figure 1.1. VAD involves distinguishing between speech and noise signals [82]. It is beneficial as a pre-processing stage for applications such as VoIP [97, 88], mobile telephony [19, 20, 39], source localization [69], speech recognition [120] and speaker recognition. In VoIP, bandwidth is saved by identifying parts of signal that is speech as only speech packets are transmitted. In mobile telephony, it saves bandwidth

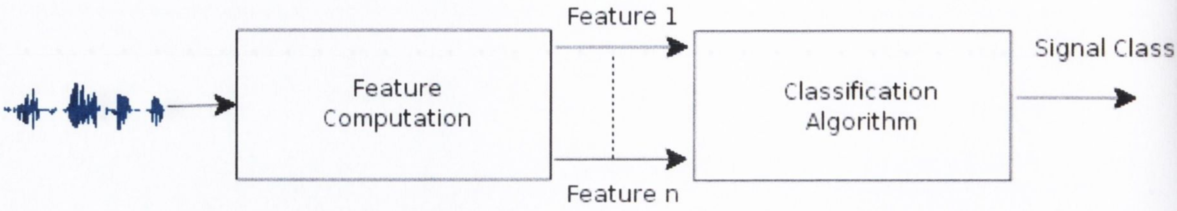


Fig. 3.1: Schematic diagram of the VAD System.

which helps to increase the number of simultaneous users and also to save power of the handset [75]. In conference applications such as e-learning it reduces the number of computations and errors by finding the DOA only on the signal segments that consist of speech.

A block diagram of a VAD system is presented in Figure 3.1. The VAD process is executed in two steps. In the first step the features of the signal frame required for VAD are computed. A feature is a characteristic of the signal that is input to the classification algorithm to determine the signal types. The second step involves the classification algorithm that determines the class of the signal frame based on the value of the features. Audio signal properties and features used in various classification algorithms are discussed in section 3.2 and section 3.3 respectively. Section 3.4 presents the theory associated with the selection of the signal features for classification. The discussion on the classification algorithms for voice activity detection is presented in section 3.6.

3.2 Speech Signal Characteristics

The speech signal received by a listener is a sound pressure wave produced by a talker and is made up of a sequence of sounds that convey the thoughts of the talker in the accepted rules of communication between humans. Speech signal waveforms change

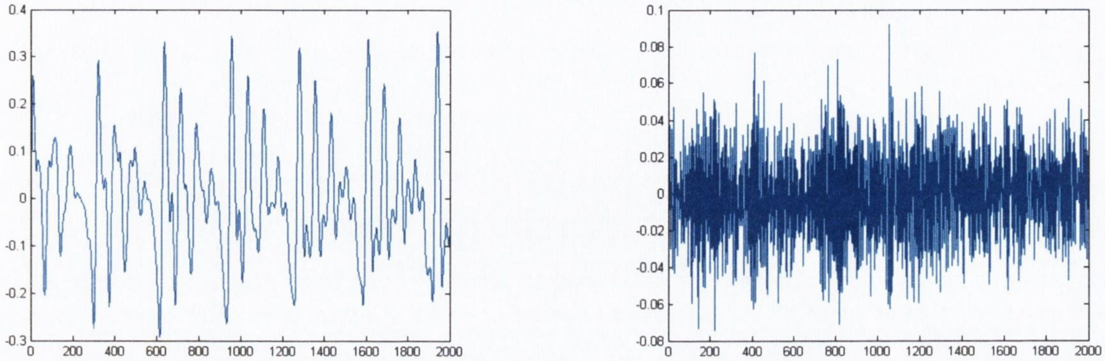


Fig. 3.2: Voiced (left) and Unvoiced (right) Speech Signals

rapidly in few milliseconds (ms) [61, pg. 225] and have non-stationary spectral characteristics [61, pg. 104]. Thus short duration frames of the speech signals are taken during which the signal is considered to be stationary.

According to the literature on human speech production in [87, 61], the lungs and trachea, the larynx and the vocal tract are the three main subsystems of the human vocal organ. The lungs provide the compressed air and influence the loudness of the speech. In the speech production process larynx plays an important role as it is the organ for producing voice. Vocal tract mainly consists of the two cavities namely the pharyngeal (throat) and the oral (mouth). Vocal cords, velum, tongue, teeth and lips are other significant anatomical features for producing speech. The resulting speech is modulated by the vocal tract.

Speech sounds can be grouped into two broad classes namely the voiced and the unvoiced (see figure 3.2). Voiced speech has characteristics of a deterministic waveform whereas unvoiced speech is more noise like [61, pg. 29 - 30]. In the English alphabets, voiced speech comprises of all the vowels and a few consonants. Unvoiced speech consists of consonants [87, pg. 66].

The Larynx contains and controls the vocal cords [87, pg. 61]. It is the source of the periodic excitation for voiced speech [61, pg. 103]. The excitation causes oscillations of the vocal cords known as the phonation. The glottal pulses produced during phonation have a repetition rate called the pitch of the pulses [87, pg 64-65, 133]. Voiced speech contain phonation due to the periodic excitation whereas unvoiced sounds do not contain phonation as their excitation is noise like [87, pg. 65], [61, pg. 109]. For a male talker pitch ranges from 80 Hz to 160 Hz and for female talker it varies from 160 Hz to 400 Hz. The rate of vibration of the vocal cords is known as fundamental frequency of the phonation. Fundamental period is the time between successive openings of the vocal cords [61, [pg. 112]. Fundamental frequency and fundamental period are inversely related. The term pitch is used interchangeably with the fundamental frequency and is usually denoted by F_0 [36], [61, pg. 114].

The vocal tract is a tube and has natural resonant frequencies known as the formants [87, pg. 66, 103]. The vocal tract has non-uniform cross section and for a male talker vocal tract is approximately 17 cm long, for female talker it is approximately 14 cm long and for children it about 10 cm long [61, pg. 102]. The shape and physical dimensions of the vocal tract determine the location of the formants in the frequency domain [61, pg. 107]. For a tube with uniform cross-section the resonant frequencies will occur at the frequency values given by:

$$f_i = \frac{c(2i-1)}{4l}, for\ i = 1, 2, 3, 4, ... \tag{3.1}$$

where f_i is the i^{th} formant frequency, c is the velocity of speech signal and l is the length of the vocal tract. However due to the non-uniform cross-section of the vocal tract the formants do not occur at the exact value. For speech recognition first three formant frequencies are used [87, 103-104]. A correspondence exists between the vowel sound and the formant frequencies [87, pg. 104]. This is useful in detecting voiced speech and also in speech/speaker recognition.

The glottal pulse train which is an excitation signal for voiced speech is periodic and contains harmonics. This pulse when applied to the vocal tract results in a speech signal that is the glottal pulse train convolved with the impulse response of the vocal

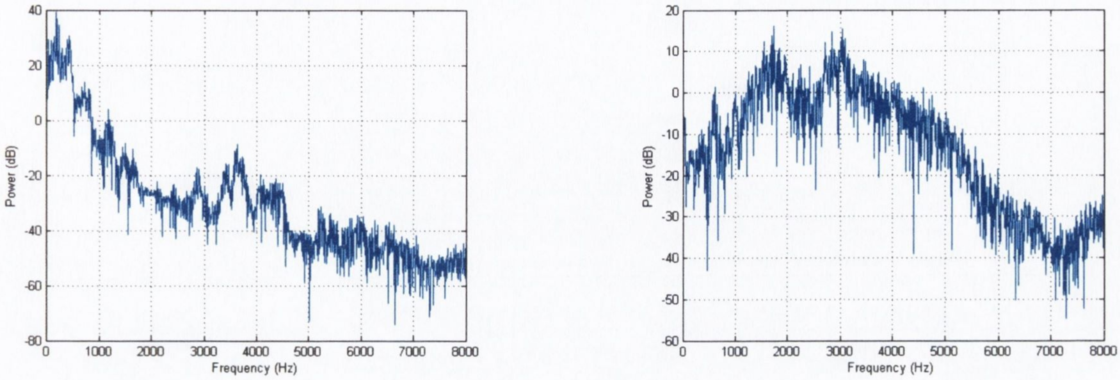


Fig. 3.3: Power Spectral Density for Voiced and Unvoiced Speech Signals

tract. As a result the output of the vocal tract appears periodic with a interspacing of glottal pulse period [87, pg. 115]. The spectrum of the glottal pulse train has a spectral roll-off of approximately 12 dB/Octave in the frequency range of 0.8 - 1 kHz. The sound spectrum undergoes a boost of 6 dB/Octave while the speech is emanated from the lips (see figure 3.3). Thus the output speech spectrum will have a net 6 dB/Octave roll-off between the first two formants for vowels [87, pg. 115-116].

3.3 Audio Signal Features

Several classification algorithms have been proposed based on the signal characteristics for different applications. The features used in the VAD classification algorithms are discussed below.

1. Short term energy and low band to full band energy ratio: One of the signal features used in the VAD algorithms is the short-term power or short-term energy [80, 82, 16], [94, pg. 120-126] of the signal. Short term energy of a signal frame is the sum of the squares of the amplitude of the signal samples within the frame. Some algorithms take the logarithm of the energy as the measure of short term energy [18, 44, 67]. Since the short term energy values obtained are very

small, the logarithm of the energy is taken as the the short term energy. Log Energy is given by:

$$E = log_{10} \sum_{i=0}^{K-1} x^2(i) \tag{3.2}$$

where $K \rightarrow$ Number of signal samples in the frame and $x(i) \rightarrow i^{th}$ signal sample in the frame.

Energy of the signal divided by the number of samples in the frame gives the short-term power of the signal. Short-time log power is given by the logarithm of the average short-term energy [68]. Both energy and power of the signal frame provide similar information about the signal but short term energy is preferred as it involves less computation [61, pg. 246].

Short-term energy helps to distinguish between vowels (usually voiced) and consonants (usually unvoiced) parts of speech [87, pg. 293]. The voiced speech has higher energy than the unvoiced speech [121]. In some scenarios it can also be used to distinguish between the background noise and other signals especially voiced speech at high SNR since the energy of the speech signal is greater than that of the background noise [111, 107]. In the presence of high Signal to Noise Ratio (SNR) it is used to detect silence periods in the speech signal. However, the short-term energy of the signal by itself is not sufficient to classify the signal in presence of low SNR as it becomes difficult to distinguish between speech and noise [107].

Other features related to the energy of the signal used in the VAD algorithms include root mean square energy of the input signal [104], peak amplitude of the signal [104], ratio of the signal energy to the noise energy [67], energy distance measure [17, 93] and different frequency band energies. Energy distance measure gives an idea of the closeness of the signal frame to the average signal of a particular class. It is the normalized Euclidean distance given by the difference between the log energy of the current frame and the known average log energy for the signal of a particular class normalized by the standard deviation of the the log energy for the signal of that class [17, 93].

Low band energy is taken in the range of 0-1 kHz and the high band energy is calculated in the frequency range of 2-4 kHz [103, 50]. High band energy helps to detect consonants [103] as they tend to have less energy in the low band [75] and high spectral concentration in the high band [74]. In contrast the low band energy of the voiced signals is high. Gaussian like noise signals have low full band energy and a mid level of low band energy [75]. Several algorithms have used different forms of the band energies in the VAD algorithms. Some of the variations include differential power/energy in 0 - 1 kHz band [18, 40, 42], differential power/energy over the whole band [19, 40, 42], variance of ratio of high frequency band to low frequency band energy [17], low band to full band energy ratio [75], ratio of the energy of the signal is greater than 4kHz to the low band energy [104].

2. Zero-crossing rate: Zero-crossing rate is a prominent feature used in the VAD [112, 93, 111, 28, 104, 94, 118] and speech recognition algorithms [62]. It's value indicates the number of times a signal has changed its sign within the frame [61, pg. 245]. Thus when the speech signal energy is high its zero-crossing rate tends to be low and vice versa [50]. It indicates the dominant frequency in the frame duration [66]. Zero-crossing rate of the signal is given by:

$$ZCR = \frac{1}{K}\sum_{i=0}^{K-1}\frac{|sign\{x(i)\} - sign\{x(i-1)\}|}{2} \tag{3.3}$$

where $sign\{x(n)\} = +1$ for $x(n) >= 0$, $sign\{x(n)\} = -1$ for $x(n) < 0$ and K is the frame length.

Voiced speech will have lower zero-crossing measurements compared to unvoiced signals [61, pg 251], [121]. Deller et al. [61] mention that short-term energy of the signal with zero-crossing is useful in identifying speech signal. Noise signals too have high zero-crossing rate. Variance of the zero-crossing rate [66], zero-crossing difference [73, 18, 42], delta zero-crossing rate [28], ratio of frame zero-crossing rate to the zero-crossing rate of the noise signal [67] are the variations of this feature utilized for VAD. Like short term energy, in low SNR, zero-crossing is not very effective for VAD [103].

3. LP coefficients, LP residual energy and PARCOR coefficients: Linear Predictive Coding (LPC) is used in speech compression [77]. The analysis of the speech signal gives the filter coefficients known as the Linear Predictor (LP) coefficients that can be used to synthesize the speech signal. The speech signal is synthesized by filtering an impulse signal with pitch period equal to the pitch of the signal for voiced speech and a zero mean, unity variance, uncorrelated noise signal for the unvoiced speech [61, pg. 267]. LP coefficients are obtained during the analysis of the speech signal. The future signal samples can be predicted using the LP coefficients, the past and the current signal samples. For classification feature sets the first two LP coefficients are considered.

   The signal sample at instant 'n' is given by:

   $$\hat{x}(n) = \sum_{i=1}^{p} a(i)x(n-i) \tag{3.4}$$

   where $x(n-i)$ is the previous $n-i$ signal samples, $a(i)$ are the LP coefficients and p is the number of LP coefficients.

   The LP coefficients can be computed from the autocorrelation function of the signal. The autocorrelation function of the signal is given by:

   $$R_{xx}(k) = \sum_{i=0}^{K-1} x(i)x(i-k) \tag{3.5}$$

   where K is the length of the signal frame and $k = 0, 1, 2, \cdots$.

The autocorrelation function (R) and the LP filter coefficient vector (A) are related by the equation $RA = M$, where

$$R = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & R_{xx}(2) & \cdots & R_{xx}(p-1) \\ R_{xx}(1) & R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(p-2) \\ R_{xx}(2) & R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(p-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{xx}(p-1) & R_{xx}(p-2) & R_{xx}(p-3) & \cdots & R_{xx}(0) \end{bmatrix}$$

$$A^T = \begin{bmatrix} a(1) & a(2) & a(3) & \cdots & a(p) \end{bmatrix}$$
$$M^T = \begin{bmatrix} R_{xx}(1) & R_{xx}(2) & R_{xx}(3) & \cdots & R_{xx}(p) \end{bmatrix}$$

The filter coefficients are determined by solving the following equation:

$$A = R^{-1}M \tag{3.6}$$

The first few LP coefficients are used for voice activity detection as they contain more information about the signal [16, 104]. They are more robust in low SNR conditions [76]. Typical filter order used is 8-10 [61, pg. 285], [82].

The difference between the signal synthesized from the LP coefficients and the original signal is called the LP residual signal. The LP residual energy for the k^{th} frame is given by:

$$x_{res(k)} = \sum_{i=0}^{K-1}(x(i) - \hat{x}(i)) \tag{3.7}$$

where K is the length of the signal frame. In [61, pg. 288-289] the LP filter order versus the LP residual energy plot shows that for voiced sounds the LP residual energy is lower compared to that of the unvoiced sounds. The LP residual energy of the signal is utilized to distinguish between voiced sounds and unvoiced sounds [16, 103, 61]. LPC distance measure [92] and log of sum of the square of the LP coefficients [68], maximum LP residual autocorrelation peak [71] are other LPC based features used for VAD.

Partial correlation or reflection (PARCOR) coefficients are generated during the analysis of the signal along with the LP coefficients [87, pg. 144]. The PARCOR coefficients are useful as the stability of the filter is ensured if their magnitude is less than 1. They are used in the lattice filter structure for the vocal tract model [87, pg. 155]. The first 'n' PARCOR coefficients values do not change if the order of the filter is increased from 'n' to a higher order. The PARCOR coefficients also contain information about the signal like the LP coefficients hence are used for VAD. The absolute difference between the first two PARCOR coefficients is another signal feature for classification [55].

$$PAR = |b(1) - b(2)| \tag{3.8}$$

where $b(1)$ and $b(2)$ are the first and second PARCOR coefficients.

4. Higher Order Statistics (HOS): Higher order statistics (HOS) like the skewness and the kurtosis of the signal can be used to distinguish between Gaussian noise and non-Gaussian signals [82]. Skewness and kurtosis are the third and fourth standardized moments respectively.

   Skewness of the signal frame consisting of K samples is given by:

$$Skew = \frac{\sqrt{(K)}\sum_{i=0}^{K-1}(x(i) - \bar{x})^3}{\{\sum_{i=0}^{K-1}(x(i) - \bar{x})^2\}^{\frac{3}{2}}} \tag{3.9}$$

   Kurtosis of the signal frame of K samples is given by:

$$Kurt = \frac{K\sum_{i=0}^{K-1}(x(i) - \bar{x})^4}{\{\sum_{i=0}^{K-1}(x(i) - \bar{x})^2\}^2} - 3 \tag{3.10}$$

   where $\bar{x}$ is the mean of the signal frame given by:

$$\bar{x} = \frac{1}{K}\sum_{i=0}^{K-1} x(i) \tag{3.11}$$

For a Gaussian signal their value is zero. The values of skewness and kurtosis of the speech signals especially voiced is non-zero. Kurtosis is useful in detecting voiced parts of speech [103]. The HOS of the LP residual signals are used as they are effective in low signal to noise ratio (SNR) conditions [82]. Their values for speech signal is different from that of the Gaussian noise and they are immune to Gaussian noise. Hence they can be used in low SNR for speech detection [82].

5. Pitch and Formants: Pitch is relevant in detecting voiced speech [28, 94, 118]. Pulse period of the glottal pulses also known as the pitch of the signal is an important time domain property of the vowel waveform. Pitch of the signal gives the periodicity of the signal. Formant frequencies are the resonant frequencies of the vocal tract. Pitch and formant frequencies are discussed in section 3.2. For short duration voiced signal can be considered periodic. Accurate estimation of pitch is of importance as error can occur which will lead to error in classification [104].

6. Spectral slope, spectral roll-off point and spectral centroid: Spectral slope of the vowel waveform is an important frequency domain property of the vowels (voiced speech). It has a spectral slope roll-off of 12 dB/octave in the frequency range of 0.8 -1 kHz [87, pg. 116]. A 6 dB/octave boost is introduced during the radiation of the speech from the lips. This overall results in a slope that has a 6 dB/octave roll-off [87, pg. 116]. Spectral slope is calculated by taking the difference of the logarithmic amplitudes of the first two formants. It can be used to distinguish between voiced speech and other signals [55].

   Another feature is the spectral roll-off point also known as spectral roll-off. Spectral roll-off point is the point below which $c\%$ of the magnitude distribution of the spectrum of the signal, obtained by the Fourier transform, is present. The spectral roll-off point is the bin number (P_{SR}) that satisfies the following equation [113, 381]:

$$\sum_{i=0}^{P_{SR}} |X(i)| = \frac{c}{100} \sum_{i=0}^{K-1} |X(i)| \tag{3.12}$$

where P_{SR} is the spectral roll-off point and $X(i)$ is the signal in the frequency domain and K is the number of Discrete Fourier Transform (DFT) coefficients.

The spectral roll-off point provides a measure of the concentration of the spectral energy of the signal and the skewness of the spectrum. The spectrum with brighter sounds has a right-skewed shape and hence the spectral roll-off value is high [118, 113]. It distinguishes between voiced and unvoiced signals. For voiced speech most of the energy is contained in the low frequency band (left skewed) whereas for unvoiced speech most of the energy is contained in the higher frequency band (right skewed). The value of c is taken as 95% [99].

Spectral centroid is defined as the balancing point of the power spectrum of the signal. It also helps to distinguish between voiced and unvoiced speech signals [99]. Spectral centroid of the signal frame is given by:

$$SC = \frac{\sum_{i=0}^{K-1} i|X(i)|}{\sum_{i=0}^{K-1} |X(i)|} \tag{3.13}$$

where K is the length of the signal frame and $X(i)$ denotes the frequency spectrum of the signal.

Spectral centroid gives a measure of the shape of the spectrum where high value of spectral centroid corresponds to the presence of high energy in the higher frequency band indicating the presence of unvoiced speech [113, pg. 381].

7. Normalized autocorrelation coefficient of the signal at unit sample delay: This value is obtained by normalizing the autocorrelation of the signal frame for a single sample delay. It is equal to [16]:

$$AC = \frac{\sum_{i=1}^{K} x(i)x(i-1)}{\sqrt{\sum_{i=1}^{K} x^2(i) \sum_{i=0}^{K-1} x^2(i)}} \tag{3.14}$$

Sub-band	1	2	3	4	5	6	7	8	9	10
Critical band center frequency (Hz)	100	200	300	400	500	600	700	800	900	1000

Sub-band	11	12	13	14	15	16	17	18	19	20
Critical band center frequency (Hz)	1148	1318	1514	1737	1995	2291	2630	3020	3467	4000

Table 3.1: Center frequency of 20 critical bands

where K is the signal frame length.

The normalized autocorrelation coefficients of the signal at unit sample delay measures the correlation between adjacent signal samples. Since adjacent signal samples of voiced sounds are highly correlated, for voiced sounds its value is higher compared to other sounds.

8. Cepstral Coefficients and Mel-Frequency Cepstral Coefficients (MFCC): Cepstrum features have also been utilized for speech recognition [61, pg. 398], speaker recognition [87, pg. 340], speech and audio classification [94, 35]. The cepstrum of a signal is equal to the inverse Fourier transform of the logarithm of the Fourier transform of the signal [113, pg. 375]. The cepstrum of the signal can be written as:

$$c(n) = F^{-1}[log_{10}|X(\omega)|] \tag{3.15}$$

where $F^{-1}[.]$ represents the inverse Fourier Transform operator and $c(n)$ are the cepstral coefficients for signal $x(n)$. The obtained coefficients are an approximate value of the cepstral coefficients [94, pg. 377].

The first LP predictor coefficient and the first sample of the Cepstrum of the

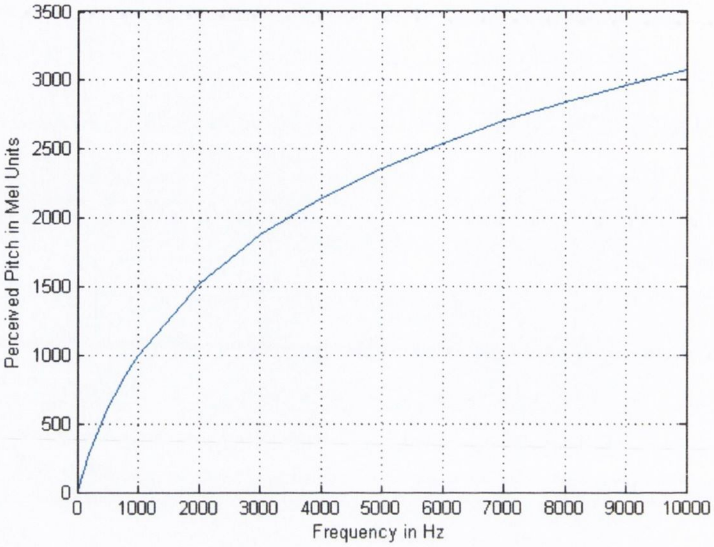


Fig. 3.4: Relation between the signal frequency in Hz and perceived signal frequency in Mel units.

signal are equal [104]. The variance of the cepstra of the noise is lower than that of the speech [49]. Other cepstral features employed for VAD include the cepstral distance [49], LPC cepstral coefficients [94].

Mel Frequency Cepstral Coefficients (MFCC) is another important signal feature [118, 94, 80]. MFCC is useful for endpoint detection [73] and word isolation [33] in speech recognition. It is a data reduction process that in essence preserves the speech signal information [117]. Perceived pitch or frequency is measured in *mel* unit. Physical frequency of the signal and *mel* are not linearly related. For frequencies below 1 kHz their relation is linear and for frequencies above 1 kHz they are logarithmically related [61, pg. 380]. Figure 3.4 shows the relationship between the perceived frequency in mel units and the real frequency of the signal in Hz. An approximate relation between the real frequency (measured in Hz) and the perceived frequency (measured in mel) scales is given by [61, pg. 380]:

$$F_{mel} = \frac{1000}{log\ 2}\left(1 + \frac{F_{Hz}}{1000}\right) \tag{3.16}$$

where F_{mel} and F_{Hz} are the perceived and actual frequencies.

Tones of different frequencies form complex sounds as perceived by the auditory system. The auditory system cannot distinguish between the tones of sound that are present in a particular band known as the critical band. The perception of a particular frequency in a critical band is influenced by the frequencies present around the frequency in the critical band [61, pg. 381]. Table 3.1 presents the center frequencies of first twenty critical bands. For complex sounds having bandwidth smaller than the critical bandwidth around the center frequency, the perceived sound is a single tone at the center frequency of the critical band. The loudness of this sound is a scaled average of the loudness of the tones present in the sound. The bandwidth of the critical band depends on the frequency [113, pg. 378]. It increases uniformly for frequencies below $1kHz$ with a bandwidth of $100Hz$ and increases logarithmically for frequencies above $1kHz$ [61, pg. 381-382].

MFCC are constructed from the frequency spectrum of the signal such that they reflect the sounds as perceived by the auditory system [113, pg. 378]. The are equal to the inverse Fourier transform of the coefficients obtained by taking the sum of the weighted log energy of the frequencies in each critical band around the mel frequencies of the band [61, pg. 382].

In order to compute the MFCC coefficients, first frequency spectrum of the signal is obtained by taking the Fourier transform of the input signal. The log magnitude of the DFT coefficients for each critical band is given by:

$$Y(k_i) = \sum_{k} log_{10}|X(k)|H_i\left(\frac{f_s k}{K'}\right) \tag{3.17}$$

where $Y(k_i)$ is the sum of weighted log magnitude of the frequency components of the signal $(X(k))$ for the i^{th} critical band, $i = 0, 1, 2, \cdots, K^{'}$. $K^{'}$ is the total number of point used in the computation of the DFT of the signal. Range of k is limited to the bandwidth of the i^{th} critical band [61, pg. 383], [113, pg. 380]. $H_i(.)$ is the triangular weighting window for the i^{th} critical band.

The relation between i^{th} band center frequency (f_i) and sampling rate (f_s) is given by:

$$f_i = k_i \frac{f_s}{K'} \tag{3.18}$$

The coefficients obtained by the weighted sum of the log magnitude of the DFT coefficients of each critical band are Inverse Discrete Fourier Transformed (IDFT) to get the mel-frequency cepstral coefficients [61, pg. 383-384], [113, pg. 380]. The equation for mel-frequency cepstral coefficients is:

$$c_{mel}(n) = \frac{1}{K'} \sum_{m=0}^{K^{'}-1} Y(m) exp\left(j\frac{2\pi mn}{K'}\right) \quad where \quad n = 0, 1, 2, \cdots, K-1. \tag{3.19}$$

and

$$Y(m) = \begin{cases} Y(k_i) & (for \quad m = k_i, where \quad i = 1, 2, \cdots, M) \\ 0 & (otherwise) \end{cases} \tag{3.20}$$

M is the number of critical bands and K is the length of the signal frame.

The mel-frequency cepstral coefficients are very significant in the area of speech as well as audio recognition and classification and are described as the most powerful features for these applications [113, pg. 380].

3.4 Feature Selection for Classification

Features are the characteristics of the input signal that are employed for distinguishing between different signal classes during the classification process. As described in section 1.2.1 feature selection/evaluation is the first step in designing a VAD system. In this section the process of feature selection/evaluation is discussed. The algorithms that can be used to obtain the best feature subset are also explained.

3.4.1 Feature Selection

Pattern classification problems generally involve solving two crucial problems. They are the selection/evaluation of features and decision rule for classification. The problem of feature selection and evaluation is more complex and important compared to the problem of finding the decision function for classification [87, pg. 171]. Feature Selection involves identifying the features that are relevant for a specific classification problem and presenting them in a compact and intelligible manner by discarding unimportant information [87, pg. 175]. Feature evaluation comprises of determining the significance of each feature towards classification. In the existing literature [48, 81, 84, 51, 119] concerned with selection of features for classification the problem of feature evaluation is discussed as feature selection. So in this thesis the term feature selection is used for feature evaluation.

While designing a classifier an early step is to identify the features that are relevant to the classification problem and hence can be used for classification. Desirable properties of features adopted for classification include the ease in measurement, statistical independence, stability for longer time periods, insensitivity to the influence of other (external) variable and possessing different values for different classes [87, pg. 175].

The supervised machine learning algorithms are trained by example data for each class known as the training data. Each data example of the training data is a vector consisting of the values of the features and the class label associated with it. All the features together form the feature set. The performance of the classifier is then assessed by another set of data (that is not a part of the training data) called as testing data in this work.

If the training data size is infinite then additional features improve the accuracy of classification since more training data for different scenarios is available for training. For finite training data size in spite of the features being statistically independent the performance of the classifier, beyond a certain stage, degrades with additional features [81]. This is known as the 'peaking phenomenon'. To avoid the peaking phenomenon, feature selection is used to improve the classifiers performance.

The objective of feature selection is to provide an optimal feature subset from the feature set such that it improves the prediction rate of a classifier [48]. Moreover feature selection improves the computational complexity by reducing the number of features to be calculated, economy by reducing the cost of measuring undesired features and also the understanding of the classification problem [81]. In cases where computation complexity and economy are not relevant, feature selection is still important as the analysis presented by Amir et al. [81] suggests that the usage of too many features gives rise to the classification error that is equal to the that obtained by chance.

The main aim of feature selection is to select the features from the feature set such that it increases the interclass distance and decreases the intra-class variance for better classification [113, pg. 214]. There are several approaches for performing feature selection which use different selection criteria to perform the task of feature selection. One of the approaches towards feature selection is to obtain the significance of each feature individually based on its discriminating ability by statistical hypothesis testing [113, pg. 216] or Receiver Operating Characteristics (ROC) curves [113, pg. 223] and form the feature subset by including the features from the feature set that are able to separate the classes best. The main drawback of this approach is that it ignores the correlation of the features that can result in poor classification [113, pg. 224].

In case of multi-class pattern classification, machine learning algorithms such as Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) can be employed for feature selection [48]. These algorithms take into consideration the correlation between the features. They select the feature subset from the feature set based on its ability to meet the class separability criteria. The class separability criterion depends on the algorithm being used for feature selection. In the case of DA the class separability criteria is to reduce the variance within the group and increase the distance between the centroids of different groups. In the case of support vector machines this criteria is to keep the ratio of the difference between the margin of hyperplane that separates the classes, with all the features and the one with selected features to the margin of hyperplane with all the original features as small as possible [15, pg. 192].

In some cases for feature selection, the classification hit rate of the classifier for which the feature selection is being done is taken as the criteria for feature selection instead of the class separability criteria. Classification results are presented using the parameters *classification Hit Rate* (HR) [54, pg 83] representing the percentage of data correctly classified into a particular class and *False Alarm Rate* (FAR) representing the percentage of data of other classes that are incorrectly classified into a particular class.

Classification hit rate is defined as:

$$\frac{No.\ of\ correctly\ classified\ data}{Total\ No.\ of\ data} \times 100 \tag{3.21}$$

Classification error rate is defined as:

$$\frac{No.\ of\ incorrectly\ classified\ data}{Total\ No.\ of\ data} \times 100 \tag{3.22}$$

Hit Rate (HR) for a class is defined as:

$$\frac{No.\ of\ correctly\ classified\ data\ of\ type\ k}{Total\ No.\ of\ data\ of\ type\ k} \times 100 \tag{3.23}$$

False Alarm Rate (FAR) for a class is defined as:

$$\frac{No.\ of\ incorrectly\ classified\ data\ into\ type\ k}{Total\ data\ of\ other\ types} \times 100 \tag{3.24}$$

In this approach the feature subset that gives the minimum error rate is utilized for classification [113]. This approach is suitable for small training data set only as for large training data set this approach is computationally intensive. Cross-validation also known as the leave one out method can be performed on the training data to obtain the hit rate [15, pg. 194]. In this method all the class data sample but one is used to form the classification rule. The excluded data is then classified using the classification rule and the group assigned by the classifier for the data is compared to its actual group. This method is applied on all the data of the training set and the classification hit rate is computed.

The resultant feature subset from feature selection process using different feature selection algorithms may not be the same. The feature subset obtained from feature selection for classification depends on the algorithm and the approach used for feature selection.

3.4.2 Backward and Forward selection

As described in the previous section feature selection involves identifying the features that are relevant for a particular classification problem. The approaches that can be taken for evaluating the relevance of a feature or feature set were discussed in section 3.4.1. In this section the algorithms that can be used to decide on the feature subsets from the feature set that is input to the feature selection algorithm are presented. The multi-class feature selection algorithms mentioned in section 3.4.1 utilize the feature subsets as the input to come up with the feature subset that is best for classification purpose. These algorithms are used in the feature selection algorithms mentioned in section 3.4.1.

This problem of deciding on feature subset to be considered in the feature selection process becomes crucial when the feature set is large. This is because, large number of features give rise to large number of feature subsets to be considered for feature selection. This problem is made simple if the size of the feature subset to be selected is known *a priori*. To select l features from the feature set of m features a total of $\frac{m!}{l!(m-l)!}$ feature subsets have to be evaluated [113, pg. 234]. Since the feature subset knowledge (the number of features that will be present in the selected feature subset) is not known *a priori*, evaluating all the feature subsets for all the possible variable combinations is computationally cumbersome [87, pg. 180-181]. To overcome this problem suboptimal searching techniques (also known as the greedy search techniques) such as the forward selection and the backward selection are employed [113, pg. 243], [48], [87, pg. 180-181].

Backward selection starts with an assumption that all the features in the feature set are significant. The class separability criteria for the feature set is computed. In the first iteration, the class separability or selection criteria for all the possible subsets with one less feature are found. The feature subset with the best value of the criteria is retained. This means that the feature that is not included in the selected subset is assumed to be insignificant and discarded. Discarded features are not considered further. This process is repeated until removal of another feature does not improve the accuracy of the classifier or when certain predetermined condition is met [87, pg. 181]. This method reduces the number of feature combination searches to $1 + \frac{1}{2(m(m+1)-l(l+1))}$ [113, pg. 234].

Forward selection starts with an empty feature set. The class separability or selection criteria for each feature are computed and the feature with the best value of the criteria is included in the feature subset. Features included in the feature subset are always included in all the further iterations. In the next iteration the class separability or selection criteria is computed for all the feature combinations of the selected features and other features. The subset with the best value of the criteria is included in the feature subset. This process is repeated until addition of another feature does not improve the accuracy of the classifier or when a predetermined condition is reached [87, 181]. This method reduces the number of feature combination searches to $\frac{lm-l(l-1)}{2}$ [113, pg.

235].

The choice of backward selection or forward selection depends on the problem. Backward selection is comparatively computationally efficient than forward selection when the value of l is closer to m than to 1 [113, pg. 235]. It is argued that forward selection comes up with weaker subsets as the significance of features is not obtained relative to other features that are yet to be included in the feature subset. On the other hand, by using backward selection features that are most important in classification may be discarded as they do not perform well with some other features in the feature set [48].

3.5 Classifier for Voice Activity Detection

This section discusses the formation of the decision function for VAD.

In video conferencing or a classroom environment, an audio signal is captured with microphones. The microphones can be the clip-on microphones or a microphone array. Apart from the speech signals other undesirable sounds such as cough, sneeze, paper shuffle and background stationary noise from the computer, projector, heating and ventilation system, air-conditioner in the room are also captured. Voice Activity Detection (VAD) is a useful to distinguish between the speech signals and other signals in order to process the signal appropriately for the application.

In the case of a clip-on microphone due to their closeness to the signal source background stationary noise level may not be very high when compared to the speech signal but sounds of breathing, paper shuffle are amplified along with the speech signal. Thus background noise in this case is not a significant problem. In the case of signals captured by a microphone array background stationary noise and other undesirable sound levels can be comparable to the captured speech signal which makes the VAD difficult.

The number of features required for VAD depends on the number of classes of the signal. The decision function for classification is based on the threshold value of one or more features. This decision function is used to classify the signal into a particular class by comparing the feature value with the threshold values of the decision function. For classifying the signal into speech and background noise (varying level) features such as signal energy levels, zero-crossing rate and periodicity may be employed.

From the VAD results by Tanyer et al. in [112], signal energy gives an overall Hit Rate (HR) of less than 80% at 10 dB SNR and approximately 90% at 50 dB at SNR. Zero-crossing rate of the signal results in a classification HR of approximately 70% at 10 dB SNR and approximately 85% at 50 dB SNR. Periodicity of the signal gives a HR of approximately 90% and 96% at 10 and 50 dB respectively. Lie et al. [75] use the logarithm of the kurtosis of the LP residual of the signal and low band to full band energy ratio to detect speech and noise frames in presence of car and street noise. The maximum HR for speech signal was 95.7% at 18 dB in the presence of street noise and 96.1% in the presence of car noise.

The main advantage of the user defined threshold approach is that the user has more control over the VAD as the threshold value for the features is user defined. The main disadvantage of this approach is the decision regarding the threshold value. For small amount of data the threshold for different features can be found by visual inspection but this data may not contain enough examples to generalize the decision function for that particular scenario. For large amount of data obtaining the threshold values may not be simple. Similarly using large number of features for classification can be very tedious. For a small class set setting a threshold value for the features may not be a difficult task as the feature set required may not be very large. But when the class set is large then a larger feature set may be required to distinguish between the signal classes. Getting threshold values for a larger class set with larger feature set can be complicated.

Another approach for VAD is to employ the machine learning pattern classification algorithms. Based on the approach adopted by the algorithms for pattern classification the machine learning algorithms can be broadly divided into two groups namely the unsupervised machine learning algorithms and the supervised machine learning algorithms. Unsupervised machine learning algorithms take unlabeled data and arrive at the possible number of class clusters contained in the data based on pre-defined pattern set or cost function provided by the user [38, 17]. Clustering is an example of this algorithm.

The supervised machine learning algorithms require training data to construct a decision function that distinguishes between signals of different classes predefined by the user. They train on the labeled data known as the training data. Based on the feature values for different classes decision function is formulated to classify the signal in the pre-defined classes. The user controls the classification process by providing the training data that is crucial in creating the decision function.

The advantage of using the supervised machine learning algorithms is that they can handle large feature sets. Another advantage is that a larger feature set can be considered initially and the feature subset consisting of the most significant features for classification can be finally utilized. The main disadvantage of these algorithms is that to get a decision function for acceptable classification rate the training data provided should be sufficiently large to train for each class and include as many scenarios as possible from the environment for which the classifier is being trained. Another disadvantage is that some of these algorithms may be parametric. Parametric classifiers assume the distribution of the data. Based on the feature values for different classes decision function is formulated to classify the signal in the predefined classes. Discriminant Analysis (DA), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and logistic regression are the well-known algorithms of the supervised machine learning algorithm category.

3.6 Classification Using Supervised Machine Learning Algorithms

This section involves the discussion regarding the decision rule for classification which constitutes the second stage of the pattern classification using the supervised machine learning algorithms. It gives a brief introduction to the concepts of the Gaussian distribution and the Bayes decision theory that are a part of some of the supervised machine learning algorithms used for pattern classification. This is followed by a description of the supervised machine learning algorithms. The three supervised machine learning algorithms described in this work are Discriminant Analysis (DA), the Artificial Neural Network (ANN) and the Support Vector Machine (SVM). The performance of these three supervised machine learning algorithms is evaluated for Voice Activity Detection (VAD) in this thesis.

3.6.1 Multivariate Gaussian Distribution

Pattern Classifiers are either Parametric or Non-parametric. Parametric classifiers assume knowledge of the distribution of the data and exploit this information to form the decision rule for classification. The multivariate normal or Gaussian density is the commonly used data distribution. Figure 3.5 shows a two dimensional Gaussian distribution. It's significance is due to analytical simplicity and also since it is appropriate in situations where the feature vectors of a particular class are continuous-valued and randomly corrupted forms of a standard vector [38, pg. 31].

The univariate normal or Gaussian density function of a input variable x is given by:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{3.25}$$

where μ and σ^2 are the mean (expected value) and variance of the input x respectively.

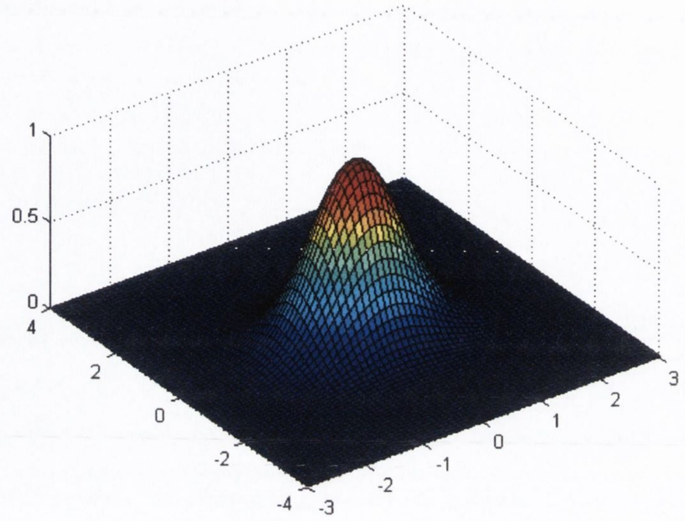


Fig. 3.5: Plot of 2-dimensional Gaussian Distribution.

The Gaussian distribution is described by its mean (μ) and variance (σ^2). It is denoted by $N(\mu, \sigma^2)$ [38, pg. 32].

μ, the mean of the input x is:

$$\mu = \int_{-\infty}^{\infty} xp(x)dx \tag{3.26}$$

and σ, the variance of the input x is:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \tag{3.27}$$

The multivariate Gaussian density function is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{3.28}$$

where $\mathbf{x}$ is l element column vector, $\boldsymbol{\mu}$ is the l element mean vector and $\mathbf{\Sigma}$ is a $l \times l$ dimensional covariance matrix. $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix and $\mathbf{\Sigma}^{-1}$ is its inverse.

The covariance matrix is given by:

$$\mathbf{\Sigma} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \tag{3.29}$$

where $E[.]$ is the expected value of signal.

Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) assume that the class has a multivariate Gaussian distribution.

3.6.2 Bayes Decision Theory

Existing pattern classification algorithms such as the DA use Bayes Decision Theory to find the probability of the input data belonging to a particular class. This probability is used to classify a signal into one of the pre-specified classes. Bayes Decision Theory takes into consideration the *a priori* or *prior* probability $P(c_i)$ of an event in order to obtain the *posterior* probability of an event. The *a priori* probability is the information that the user is aware of regarding a particular event taking place. For classification purpose Bayes formula is given by:

$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{p(\mathbf{x})} \tag{3.30}$$

In equation 3.30, $\mathbf{x}$ is a l dimensional data vector. The term $P(c_i|\mathbf{x})$ is the *posterior* probability that indicates the likelihood of occurrence of an event c_i for a given input vector $\mathbf{x}$. $p(\mathbf{x}|c_i)$ is the likelihood of the event conditional probability density function for $\mathbf{x}$ conditioned on true event being c_i and i is the total number of events. The term $p(\mathbf{x})$ in the denominator is a scaling factor that ensures that the sum of the *posterior* probabilities for all the possible events is unity [38, pg. 24]. The numerator in equation 3.30 is critical in the determination of the *posterior* probability of an event. In case of the $P(c_i)$ being equal for all the events the decision is based on the term $p(\mathbf{x}|c_i)$ and vice versa. Bayes Decision Theory combines both the terms in order to obtain minimum probability of error [38, pg. 23].

The data of unlabeled data is assigned to the class i by the following rule:

$$P(c_i|\mathbf{x}) > P(c_j|\mathbf{x}) \quad for\ all\ i \neq j \tag{3.31}$$

The value of $P(c_i|\mathbf{x})$ is found by equation 3.30. The data is assigned to the class that has the maximum *posterior* probability [90, pg. 71].

3.6.3 Supervised Machine Learning Algorithms

This section presents the theory of the Discriminant Analysis which are parametric classifiers and the non-parametric classifiers like the cascade neural networks and the Support Vector Machine (SVM).

Parametric Classifiers

In this section, a brief description of the parametric classifiers considered for classification in this work is presented.

Discriminant Analysis: Discriminant Analysis classifiers are parametric classifiers and assume that the signal class has a multivariate Gaussian distribution. Discriminant analysis when used to predict the class of the given data is known as Predictive Discriminant Analysis (PDA) and when used to find the class difference is known as the Descriptive Discriminant Analysis (DDA) [54, pg. 28]. PDA algorithms are used here for Voice Activity Detection (VAD). DA involves training the classifier by finding a classification function. This classification function is found from an example data set whose group membership is already known also called the training data. The classification function in turn is used to classify data whose class is unknown based on the classification rule. Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA) and Mahalanobis Distance (MD) are the important algorithms belonging to the PDA category.

1. **Quadratic Discriminant Analysis (QDA)**: QDA is quadratic in the unlabeled data. Assuming that the classes have a multivariate normal distribution with respect to the input data the likelihood of conditional probability density function for $\mathbf{x}$ conditioned on true class being c_i is given by:

$$p(\mathbf{x}|c_i) = \frac{1}{(2\pi)^{\frac{l}{2}}|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp\left[\frac{-1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right] \tag{3.32}$$

where $i = 1, 2, \cdots, k$ is the total number of pre-defined classes, $\boldsymbol{\mu}_i$ is the expected value of data vector for class c_i, $\mathbf{\Sigma}_i$ is a $l \times l$ covariance matrix, $|\mathbf{\Sigma}_i|$ is the determinant of the covariance matrix and $\mathbf{\Sigma}_i^{-1}$ is the inverse of the covariance matrix.

The covariance matrix for i^{th} class is given by:

$$\mathbf{\Sigma}_i = E\left[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^T\right] \tag{3.33}$$

From the Bayes formula for posterior probability in equation 3.30, the likelihood of occurrence of class c_i for a given input $\mathbf{x}$ is given by:

$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{p(\mathbf{x})} \tag{3.34}$$

Substituting equation 3.33 in equation 3.34 and taking the logarithm, $ln(.)$ of the resultant $P(c_i|\mathbf{x})$ the quadratic discriminant function is given by:

$$QF_i(\mathbf{x}) = \ln\ P(c_i) - \frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] - \frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i| \tag{3.35}$$

The term $\frac{-l}{2}\ln 2\pi$ is a constant. When the *a priori* probability of all the classes is equal, $\ln\ P(c_i)$ is equal for all the classes. Thus neglecting both the terms equation 3.35 can be rewritten as:

$$QF_i(\mathbf{x}) = -\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] - \frac{1}{2}\ln|\mathbf{\Sigma}_i| \tag{3.36}$$

The unlabeled data is assigned to the group c_i such that:

$$QF_i(\mathbf{x}) > QF_j(\mathbf{x});\ for\ all\ i \neq j \tag{3.37}$$

where $i, j = 1, 2, \cdots, k$

2. **Linear Discriminant Analysis (LDA)**: LDA is a special case of the QDA where the covariance matrix for all the classes is equal and a common pooled covariance matrix is taken [81], [113, pg. 21-22]. LDA is a named so as it is linear in the unlabeled data.

   Consider the equation 3.35. When expanded the equation is written as:

$$\begin{aligned} QF_i(\mathbf{x}) = & -\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_i^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}_i^{-1}\mathbf{x} \\ & + \ln P(c_i) - \frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i| \end{aligned} \tag{3.38}$$

   When the covariance matrix of the classes are equal then in equation 3.38 the first term which is the quadratic in the input data can be neglected as it is same for all the classes. The constant terms $-\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}_i|$ can be neglected as they are equal for the classes. Thus equation 3.38 can be rewritten to give the linear discriminant analysis function:

$$LF_i(\mathbf{x}) = \boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(c_i) \tag{3.39}$$

   The unlabeled data is assigned to the group c_i such that:

$$LF_i(\mathbf{x}) > LF_j(\mathbf{x});\ for\ all\ i \neq j \tag{3.40}$$

   where $i, j = 1, 2, \cdots, k$

3. **Mahalanobis Distance (MD)**: The MD algorithm is the simplest of the DA algorithms. It measures the distance of the input data from the centroid of each

class. It is a special case of the function provided in equation 3.35. When all the classes have equal probability of occurrence ($P(c_i)$) and all the classes have same covariance matrix then equation 3.35 can be rewritten as:

$$QF_i(\mathbf{x}) = \frac{-1}{2}[(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)] \tag{3.41}$$

For non-diagonal covariance matrix maximizing $QF_i(\mathbf{x})$ of equation 3.41 is equivalent to minimizing the norm of the covariance matrix [113, pg. 25]. This distance measure is known as the Mahalanobis distance given by:

$$D_i(\mathbf{x}) = [(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)]^{\frac{1}{2}} \tag{3.42}$$

For diagonal covariance matrix the equation 3.42 is reduced to the Euclidean distance given by:

$$D_i(\mathbf{x}) = ||\mathbf{x} - \boldsymbol{\mu}_i|| \tag{3.43}$$

The unlabeled data is assigned to the class for which the $D_i(\mathbf{x})$ value is minimum indicating that the data is closest to the centroid of that class [54, pg. 55-56].

LDA and QDA are among the most popular classifiers as they are suitable for different applications. LDA and QDA are computationally demanding since for a large number of features large numbers of unknown parameters have to be computed in high dimensional feature space. Moreover, they need a large number of training data to provide acceptable classification results [113, pg. 28].

Non-Parametric Classifiers

In this work a number of non-parametric classifiers are also considered as decision engines. A brief outline of each of the schemes follows.

1. **Cascade Correlation Multilayer Artificial Neural Networks:** Artificial Neural Networks (ANN) attempt to act as simplified models of the human brain. They are usually nonlinear and non-parametric machine learning classifiers. ANN consists of an input layer, an output layer and layers between the input and output layer known as the hidden layers. All the layers consist of processing elements. A processing element maybe connected to other processing element and to itself. The topology of the neural networks is defined by the interconnection between the processing elements [90, pg. 101].

   The output of each processing element is a weighted sum of the inputs into the element and is dependent on the threshold function known as the activation function being used by the neural network. The value of the weighted sum of inputs is compared with the threshold function value and output is generated accordingly [90, pg. 102]. The threshold function is known as the activation function. Most commonly used activation function is the sigmoid function. The output range of the sigmoid activation function is between 0 and 1. 0 represents low input values to the processing element and 1 represents high value input to the processing element [83]. Using the processing elements, ANN constructs the discriminant functions for classification. The number of discriminant functions and their shape is dependent on the topology of the ANN [90, pg. 101].

   ANN learns from the training data that consists of the examples related to the classification problem. During training, the training data is input and output is computed. The output is compared to the desired response and the error is obtained. The error value is then used to adjust the weights of the processing elements using the training algorithms. For changing the weights, all the training data is used and the process is repeated till the convergence criteria is satisfied. Determining the shape/number of discriminant functions and the location of the discriminant function in the pattern space are the two prominent issues encountered in designing the ANN for classification with minimum error [90, pg. 101].

The cascade correlation also known as the cascade architecture is an iterative construction algorithm that is employed to build an artificial neural network [96, pg. 172]. Initially the input and output processing elements are connected directly without any hidden layers. The processing elements of the hidden layers are later included in the network one at a time [96, pg. 172], [38, pg. 329].

The processing elements in the hidden layer have two type of weights associated with them. The first type of weights connects the new processing element to the input layer processing elements and also to the output of the previously added processing elements in the hidden layer. The weights associated with the connection between the previously added processing elements and the newly added processing element is -1. This is done in order to stop the new hidden processing element from learning function that have already been taken care of by the previously added hidden processing elements [38, pg. 330].

The weights connecting the input of newly added processing element to the output of the input layer processing elements are trained such that the correlation between the newly added processing elements output and the residual error signal at the networks output prior to the inclusion of the new processing element is maximum. The value of these weights once computed is not changed [113, pg. 152].

The second type of weights connects the new processing elements to the output layer processing elements. These weights undergo adaptive training and hence are updated such that the sum of squares error cost function is minimized. The training process is complete when a pre-specified network performance value is obtained [113, pg. 152].

Cascade correlation architecture has a layered structure and does not perform global optimization i.e. it does not update all the weights periodically. Updating weights periodically results in smaller networks that have better generalization ability compared to the cascade architecture but take more time to train [96, pg. 172-173]. Thus training cascade correlation neural network is faster since all the weights are not updated at a given time [38, pg. 330].

The Next section provides a brief description of the third type of machine learning algorithm used in this thesis for performance comparison for Voice Activity Detection (VAD).

2. **Support Vector Machine:**

   Support Vector Machine (SVM) is a non-parametric supervised machine learning algorithm that is used for pattern classification. It is a supervised machine learning algorithm since it trains on data with labeled class and classifies the unlabeled data based on the training. SVM uses training data to find an optimal hyperplane that separates different classes of the training data. This is done by transforming the data into higher dimension space. The dimension of the transformed data space is greater than that of the original data. A non-linear mapping function ($\phi(.)$) is used to transform the training data to the higher dimension space such that the data of two classes can always be separated by an optimal separating hyperplane [38, pg. 258]. The performance of SVM is not affected by the dimension of the training data.

   Consider the input training data represented by ($\mathbf{x}_i$) where $i = 1, 2, 3, \cdots, m$; and m is the number of input training data examples. Using the nonlinear mapping function $\phi(.)$, the input data is transformed to a higher dimension space. The transformed data is represented by $\mathbf{y}_i$. Thus the relation between the input data ($\mathbf{x}_i$) and transformed data ($\mathbf{y}_i$) is given by:

$$\mathbf{y}_i = \phi(\mathbf{x}_i); \qquad i = 1, 2, \cdots, m. \tag{3.44}$$

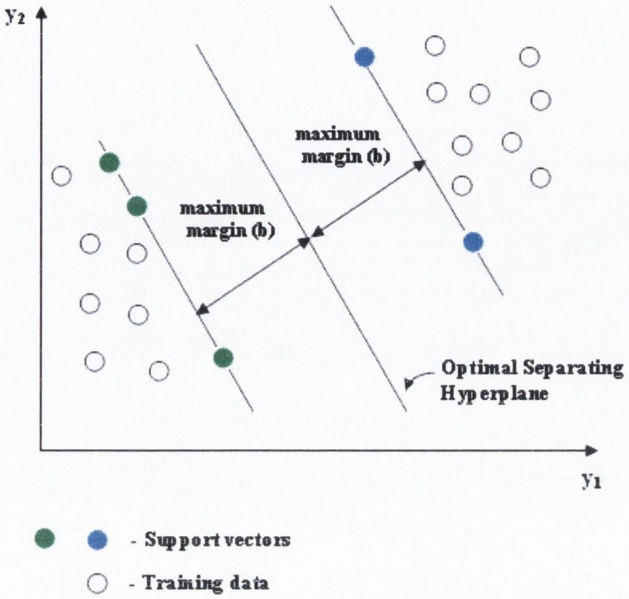


Fig. 3.6: Example of separable problem of SVM in a 2 dimensional space

The discriminating function obtained that separates the data classes into one of the two classes is given by:

$$f(\mathbf{y}) = \mathbf{w}^T\mathbf{y} \tag{3.45}$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_m]$ is the weight associated with each element in the transformed vector $\mathbf{y}_i$. The class is also given a numerical value. In the two class case let the values of the classes (z_i) be ±1.

The association of the transformed data and the class is given by:

$$For\ \ z_i = 1, \quad f(\mathbf{y}_i)\ >\ 1\ and\ For\ \ z_i = -1, \quad f(\mathbf{y}_i)\ <\ 1 \tag{3.46}$$

The separating hyperplane separates the two classes based on the following rule:

$$z_i\ f(\mathbf{y}_i) \geq 1 \qquad i = 1, 2, \cdots, m. \tag{3.47}$$

The distance between the separating hyperplane and the class boundary is called the margin (b). The aim of the training algorithm is to find a separating hyperplane that maximizes the value of the margin (b). The equation to be satisfied is:

$$\frac{z_i\, f(\mathbf{y}_i)}{||\mathbf{w}||} \geq b \tag{3.48}$$

where the aim is to find an optimal hyperplane by selecting the values of weight vector ($\mathbf{w}$) such that the value of the margin b is maximized.

The optimal separating hyperplane maximizes the distance between itself and the class boundary of the training data of the classes that it separates (see figure 3.6). The generalization ability of the SVM depends on the location of the separating hyperplane. Thus if the optimal hyperplane is chosen as the separating hyperplane then the generalization ability of the classifier is maximized provided that the unlabeled data follows the same probability rule as the training function and also if the training data does not contain any outliers [15, pg. 16].

The solution obtained for equation 3.48 has several hyperplanes. A constraint is applied on the solution of equation 3.48 where $b\ ||\mathbf{w}|| \ = 1$. This constraint is applied so that the value of margin b is maximized while the value of weight vector $\mathbf{w}$ is minimized. Thus finding the optimal hyperplane is a constrained optimization problem solved by Lagrangian multiplier method [38, pg. 262]. The constrained problem ($L(a, \boldsymbol{\alpha})$) is represented as:

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m} \alpha_i[z_i \mathbf{w}^T \mathbf{y}_i - 1] \tag{3.49}$$

where α_i is a multiplier bounded by $\alpha_i \geq 0$. Equation 3.49 is solved to obtain the weight vector and the value of multiplier α_i. The aim is to obtain weight vector that minimizes the equation 3.49 and the value of multipliers that maximize the equation.

The problem when reformulated gives rise to the optimization problem that maximizes [38, pg. 264], [78]:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j z_i z_j \mathbf{y}_i^T \mathbf{y}_j \tag{3.50}$$

where the constraint on the equation 3.50 is:

$$\sum_{i=1}^{m} z_i \alpha_i = 0 \qquad and \qquad \alpha_i \geq 0, \qquad i = 1, \cdots, m \tag{3.51}$$

All the training data is not used to obtain the position of the optimal hyperplane. The data that is used to train the SVM in order to maximize the distance between the class boundary and the hyperplane are called support vectors. Support vectors provide most of the information for classification purpose. The support vectors determine the location of the hyperplane and their inclusion or exclusion during the training influences the location of the hyperplane and the width of the margin [34].

The decision function is given by:

$$f(\mathbf{x}) = sign\left(\sum_{i} z_i \alpha_i \mathbf{y}_i^T \mathbf{y} + b\right) \quad where \quad i \in S \tag{3.52}$$

where S is the indices of the set of support vectors [15, 24].

If the optimal separating hyperplane is able to separate the training data without any errors then the expected value of the probability of error while classifying unlabeled data is bounded by the value that depends on support vectors only. The bound on the probability of error ($Pr(error)$) [38, pg. 263] is given by:

$$E[Pr(error)] \leq \frac{E[D_{sv}]}{m} \tag{3.53}$$

where $E[.]$ is the expectation operator, D_{sv} is the number of support vectors and m is the number of training data.

The generalization ability of the SVM will be better if the optimal hyperplane is constructed with smaller number of support vectors compared to the size of the training data [34]. Higher generalization ability leads to lower classification error. Thus for the right choice of non-linear transformation function $\phi(.)$, in the presence of smaller number of support vectors the probability of error will be smaller [38, pg. 263].

Kernel functions are applied on the data to transform them so that they are linearly separable [52]. They employ scalar products of the mappings, $\phi(.)$ in feature spaces [78]. These functions are given by the relation:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T.\phi(\mathbf{x}_j) \tag{3.54}$$

The decision function is given by:

$$f(\mathbf{x}) = sign\left(\sum_i z_i\alpha_i K(\mathbf{x}, \mathbf{x}_j) + b\right) \quad where \quad i \in S \tag{3.55}$$

where S is the indices of set of support vectors [15, 26].

Four basic kernels used in SVM are given below:

(a) Linear kernel: This kernel is used if the data can be classified linearly in the input space. The data need not be mapped into higher dimensional space. The linear kernel is given by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j \tag{3.56}$$

(b) Radial Basis Function (RBF): The Radial Basis Function kernel is given by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2),\ \gamma > 0 \tag{3.57}$$

where γ is a parameter that controls the radius [15, 27].

(c) Polynomial kernel: The polynomial kernel is given by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d, \ \gamma > 0 \tag{3.58}$$

where d is the degree of the polynomial kernel.

(d) Sigmoid kernel: The sigmoid kernel is given by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + r) \tag{3.59}$$

The two class classification problem can be extended to multi-class classification problem. SVM can be trained as "one against all", "one against one" or "directed acyclic graph" to obtain the decision functions. The decision functions provide the probability of the unlabeled data belonging to a particular class.

In the case of "one against all", the number of SVM models constructed is equal to the number of classes in the classification problem. The training data is divided into two groups for each SVM model. The class for which the decision function is being obtained is treated as one of the classes whereas the rest of the classes in the classification problem are grouped together to form the second class. The decision function for each class provides the probability of the unlabeled data belonging to the respective class. The unlabeled data is assigned to the class for which the probability of the data belonging to that class is maximum [53].

For the "one against one" case the number of SVM's constructed is $\frac{k(k-1)}{2}$ where k is the number of classes in the classification problem. The SVM's are trained by taking a pair of classes and ignoring the rest of the classes. The SVM's are trained for all the combinations of the classes taking two different classes at a time. In this case a voting scheme is used to assign the unlabeled data to a particular class. Every constructed SVM model indicates the class that the data might belong to compared to the other class. The vote of the class to which the SVM model indicates the unlabeled data belongs to is incremented. The accumulated votes for all the classes is compared and the unlabeled data is assigned to the class with maximum votes [53].

In the Directed Acyclic Graphs (DAG) approach the SVM's are constructed similar to the "one against one" method. They differ in the class assignment procedure for the unlabeled data. The DAG method uses a directed acyclic graph for class assignment. The probability of the signal belonging to one of the first two classes is compared and the class that has smaller probability is eliminated as a prospective class of the unlabeled data. The process of eliminating the classes continues till only two classes are remaining. The probability of the unlabeled data belonging to one of the two classes is compared. The unlabeled data is assigned the class with higher probability of the data belonging to that class [53].

3.7 Summary

This chapter discussed different aspects of designing a Voice Activity Detection (VAD) system using a machine learning algorithm. The speech signal characteristics were discussed. The characteristics of speech are useful in identifying the features that can be employed in the design of the voice activity detection (VAD) system. The features used in different VAD algorithms were discussed next. Feature selection and the choice of the classifier are the two problems encountered while designing a VAD system. Feature selection deals with the selection of a feature subset from the feature set in a convenient but reasonable manner in order to improve the classifiers performance. Identifying the classification algorithm involves deciding on the classification algorithm that performs the best for the VAD with the selected features. Finally a description of the supervised machine learning algorithms such as the Discriminant Analysis (DA), the cascade correlation Artificial Neural Network (ANN) and the Support Vector Machine (SVM) was provided.

Chapter 4

DOA Studies

The primary contribution of this thesis is the modified YIN based TDE and the Weighted YIN based TDE algorithms, a YIN [36] based approach, for TDE. Studies comparing the performance of the modified YIN based TDE algorithm and the unbiased cross-correlation algorithm are presented in [105]. This chapter presents the simulation results to evaluate the performance of the modified YIN based TDE and the Weighted YIN based Time Delay Estimation (TDE) algorithms relative to the WCC, MAMDF and GCC-PHAT TDE algorithms. The chapter starts with a description of the modified YIN based TDE and the Weighted YIN based TDE algorithms. This is followed by a section on the metrics used to evaluate the performance of the TDE algorithms through simulations and the simulation setup section. The simulation results are presented in the next section. The chapter concludes with an observation and conclusion section.

4.1 YIN for TDE

For the TDE using the YIN algorithm [36] (discussed in section 2.12) only the difference function, cumulative mean normalized difference function and the parabolic interpolation steps are employed. The signal $x(n)$ and its shifted version $x(n+\tau)$ in the equation 2.59 are replaced by the signals received by a pair of microphones $x_1(n)$ and $x_2(n)$ for TDE. The difference between the signals is squared and summed to give the difference function for different time delays. The time delay (τ) value associated with the lowest difference function value is taken as the estimated time delay.

The difference function of the YIN algorithm is written as:

$$d(\tau) = \sum_{n=0}^{N-1} (x_1(n) - x_2(n+\tau))^2 \tag{4.1}$$

The equation when expanded gives:

$$\sum_{n=0}^{N-1} x_1^2(n) + \sum_{n=0}^{N-1} x_2^2(n+\tau) - 2\sum_{n=0}^{N-1} x_1(n)x_2(n+\tau) \tag{4.2}$$

The first two terms in the equation represent the energy of the two signals and the third term is equal to the cross-correlation of the two signals. This algorithm does not give the minima at the same τ value at which the cross-correlation function gives the maxima. This is due to different values of the energy of the second signal for different time shifts. This algorithm may however suffer from the same problem that is encountered in the cross-correlation i.e. it may be biased to certain τ values due to the presence of the sum of the squares of more sample difference terms compared to other τ values. To overcome this problem the function is normalized such that the normalized difference function term associated with the actual delay has the least value [36].

In this step, for a particular τ value, the difference function at each time delay is divided by the normalization function. Normalization function is equal to the mean of the difference function associated with delay values starting from zero delay to the current delay. Since the YIN algorithm is originally meant to find the fundamental frequency, the zero delay τ term has been avoided by starting the normalization at first delay. To make the algorithm applicable to the TDE, the normalization starts with the zero delay instead of first delay value. For the YIN algorithm, the function obtained by dividing the difference function values for different time delays by the corresponding normalization function is called the Cumulative Mean Normalized Difference Function (CMNDF).

The Cumulative Mean Normalized Difference Function (CMNDF) for the modified YIN based TDE algorithm represented by R_{YDF12} is given by:

$$R_{YDF12}(\tau) = \begin{cases} 1 & (\tau = 0) \\ \frac{d(\tau)}{\frac{1}{\tau+1}\sum_{i=0}^{\tau} d(i)} & (for\ \tau > 0) \\ \frac{d(\tau)}{\frac{1}{|\tau|+1}\sum_{i=\tau}^{0} d(i)} & (for\ \tau < 0) \end{cases} \tag{4.3}$$

The delay τ value associated with the lowest R_{YDF12} function value is taken as the estimated delay. The main difference between the cross-correlation algorithm and the modified YIN based TDE algorithm is that the former is highly sensitive to amplitude changes whereas the latter is not as it takes into account the amplitude changes in terms of the energy changes associated with different time shifts [36].

Absolute threshold is used in the YIN algorithm for period calculation to prevent getting higher period values instead of the actual period as they may have the minimum value of the CMNDF. This stage is useful in the period calculation as theoretically no upper limit has been set on the search range of the period. In practice the period range is limited by the frame size used during implementation. In the case of TDE, the upper limit of the time delay is dependent on the formula used to find the direction of arrival using the time delay value as discussed in section 2.2. Since the time delay limit is very small compared to the signal frame size used this step is not included in the modified YIN based TDE algorithm.

Finally second-order Lagrange interpolation also known as the parabolic interpolation can be included to improve the resolution of the estimated time delay. The parabolic interpolation for modified YIN based TDE algorithm is similar to the one used in [32] and is presented in section 2.5. For the purpose of fractional TDE, the values of x_1, x_2, x_3 in equation 2.21 of section 2.5 are replaced by the values of $\tau - 1$,

τ and $\tau + 1$ respectively where τ is the estimated time delay. Similarly $f(x_1)$, $f(x_2)$, $f(x_3)$ are replaced by the values of $R_{YDF12}(\tau - 1)$, $R_{YDF12}(\tau)$ and $R_{YDF12}(\tau + 1)$ respectively. Between $\tau - 1$, τ four points are interpolated and between τ, $\tau + 1$ four points are interpolated. Finally the time delay corresponding to the minimum value of R_{YDF12} for all the delays between $\tau - 1$ and $\tau + 1$ is taken as the estimated time delay.

4.2 Weighted YIN

Another algorithm proposed is the Weighted YIN based TDE algorithm. It is inspired by the Weighted Cross Correlation (WCC) algorithm proposed by [32] given by the equation 2.56 of section 2.10.3. For TDE the Weighted YIN based TDE algorithm combines the GCC-PHAT algorithm and the modified YIN based TDE algorithms. The Weighted YIN based TDE algorithm is given by:

$$R_{WYDF}(\tau) = \frac{R_{GCC}(\tau)}{R_{YDF12}(\tau) + \epsilon} \tag{4.4}$$

where $\tau = 0, \pm1, \pm2, \cdots$; $R_{GCC}(\tau)$ is given by equations 2.41 and 2.44. $R_{YDF12}(\tau)$ is given by the equation 4.3 respectively. A small positive value ϵ is added to prevent division overflow. The τ value that corresponds to the maximum value of the R_{WYDF} function is taken as the estimated time delay. Similar to the modified YIN based TDE algorithm the resolution of the DOA estimate can be improved by applying parabolic interpolation on the estimated time delay to obtain the fractional sample delays.

4.3 Performance Evaluators for Time Delay Estimator

The accuracy of the algorithm is used to compare the performance of the algorithms for TDE [101]. The total number of delay estimates at each time delay by each algorithm is presented in the form of histograms or bar plot for different signal to noise ratio (SNR). A good TDE is expected to have higher estimated delays at the expected time delays.

The expected value of the estimated time delay for an unbiased estimator is equal to the true value. In practice estimators may not be unbiased. Even though an unbiased estimator does not imply that the estimator is good still a biased estimator is not preferable [91]. The performance of the estimators is commonly compared by the *Mean Squared Error (MSE)* [91, 32].

The bias of the estimator is given by the equation:

$$Bias = E[\hat{\tau}] \ - \ \tau \tag{4.5}$$

$E[.]$ is the expectation operator, $\hat{\tau}$ is the estimated delay and (τ) is the actual or expected delay value.

Mean Squared Error (MSE) is equal to the square of the difference between the estimated value $(\hat{\tau})$ and the expected value τ of the parameter. It is given by:

$$MSE = E[(\hat{\tau} \ - \ \tau)^2] \tag{4.6}$$

MSE provides an estimate of the average variance of the error around the expected delay value. A good TDE algorithm has a smaller MSE value. The comparison of the performance of the TDE algorithms is presented as bar plots providing the accuracy of the algorithms and the MSE plots.

4.4 Simulation Setup

The impulse response of the room for the simulations to compare the performance of the Generalized Cross-Correlation Phase Transform, modified YIN based TDE algorithm, Weighted YIN based TDE algorithm, Modified Average Magnitude Difference Function and Weighted Cross-Correlator time delay estimator was generated by the

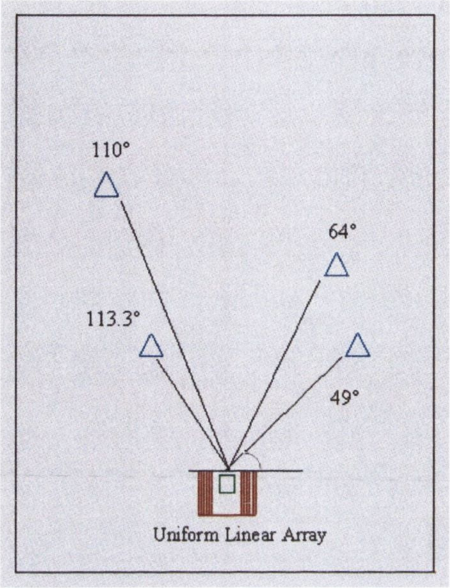


Fig. 4.1: Source locations for simulations.

EASE software [5]. The dimensions of the room are $6m \times 3.5m \times 2.5m$ and the reverberation time ($RT60$ introduced in section 2.6) is 0.5 s. The dimensions and the reverberation time of the room are approximately equal to that of the classroom for which the experimental results are presented in chapter 6.

A Uniform Linear Array (ULA) consisting of 4 microphones with the adjacent microphones spaced 3.4 cm apart was placed (considering the plan view of figure 4.1) in the center of the room horizontally and a meter vertically from the wall. The source locations are shown in figure 4.1. The signal source locations were: $49°, 64°, 110°, 113.3°$. The signal source and the microphone array were on the same plane. The impulse response generated by the EASE software [5] was convolved with the clean speech signal from the TIMIT database [6] to obtain the source signal. The four source locations consisted of two female source locations corresponding to $49°$ and $110°$ and two male source locations corresponding to $64°$ and $113.3°$.

The white Gaussian noise generated in the Cool Edit Pro 2.0 package [7] was used

d (cm)	Delays in terms of signal samples and corresponding angles								
	-4	-3	-2	-1	0	1	2	3	4
3.4	-	-	-	129°	90°	51°	-	-	-
6.8	-	161°	129°	108.4°	90°	71.6°	51°	19°	-
10.2	147°	129°	115°	102°	90°	78°	65°	51°	33°

Table 4.1: Delay values in signal sample number and their corresponding source angles for different separation between microphone pair (d) at sampling frequency = 16kHz

as the background noise. The sampling rate of the source and the noise signal was 16 kHz. The velocity of sound was taken as 342 m/s. The Signal to Noise Ratio (SNR) is introduced in section 2.6. The signal power in equation 2.22 is the power of the source signal and the noise power corresponds to the power of the noise signal generated in Cool Edit Pro 2.0. The SNR values used in the simulations are 10, 15 and 20 dB. These values were chosen taking into consideration the minimum required SNR value in a classroom as mentioned in section 1.1. Signal frames of 20 ms were taken for the simulations with no overlap between the adjacent frames.

The performance of the Phase Transform Generalized Cross-Correlator (GCC-PHAT), modified YIN based TDE, Weighted YIN based TDE, Modified Average Magnitude Difference Function (MAMDF) and Weighted Cross-Correlator (WCC) was compared using the accuracy of the time delay estimates and the mean squared error (MSE) plot. The accuracy of the algorithms is displayed in the bar plots for each TDE algorithm for different SNR values. The MSE plot for the algorithms is also provided. The x-axis of the MSE plot corresponds to the SNR value and the y-axis corresponds to the MSE.

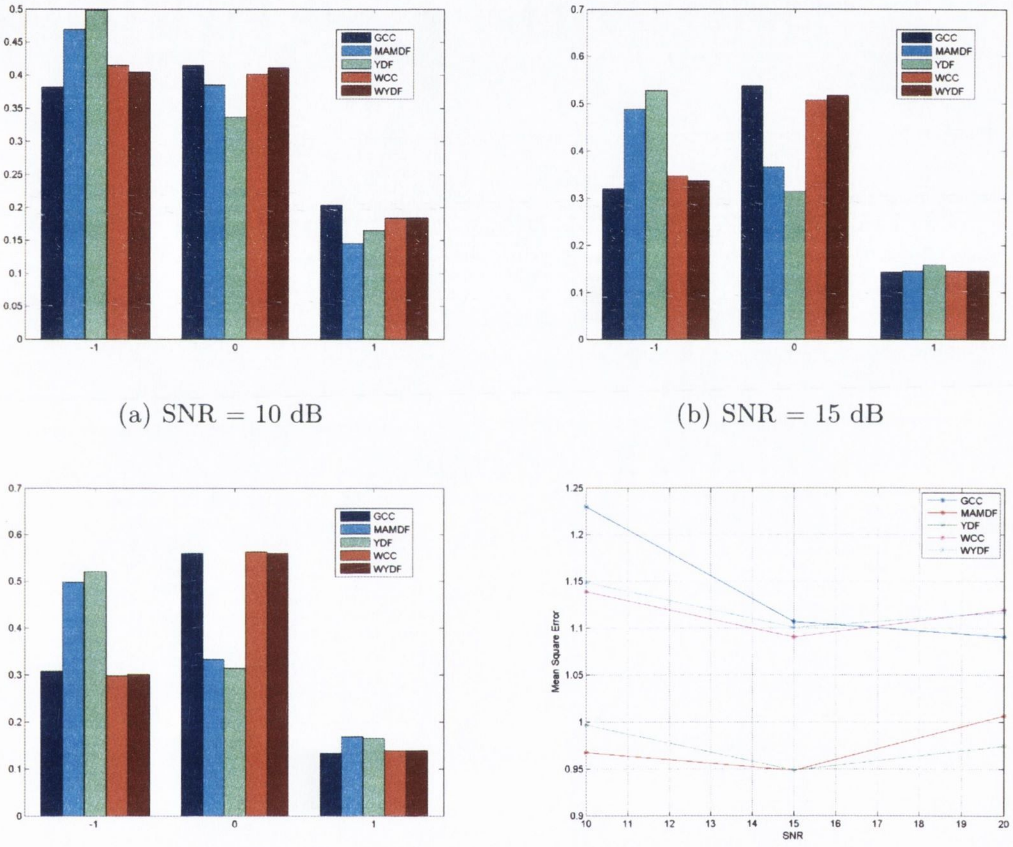
4.5 Simulation Results

This section presents the simulation results comparing the performance of the Generalized Cross-Correlation Phase Transform (GCC-PHAT), modified YIN based TDE (YDF), Weighted YIN based TDE (WYDF), Modified Average Magnitude Difference Function (MAMDF) and Weighted Cross-Correlator (WCC). The GCC-PHAT algorithm is labeled as the GCC in the plots.

Using equation 2.14 the possible angles for microphone separation of 3.4 cm are 129°, 90° and 51° corresponding to the delays -1, 0 and 1. A negative delay indicates that the second microphone in the microphone pair receives the signal before the reference microphone. A positive delay indicates that the reference microphone receives the signal before the second microphone in the microphone pair. Zero delay indicates that both the microphones in the microphone pair receive the signal simultaneously.

Figure 4.2(a), 4.2(b), 4.2(c) presents the bar plot for time delay estimates at 10, 15, 20 dB respectively for female speech source at an angle 110° when the separation between the microphones (d) in the ULA is 3.4 cm. The expected delay is between -1 and 0. The delay is taken as -1. The accuracy of the modified YIN based TDE algorithm (YDF) is the best for all the SNR values followed by the MAMDF algorithm. The accuracy of both the algorithms improves with the increase in SNR. The GCC-PHAT, WCC and WYDF algorithms performance degrades with the increase in SNR. The MSE plot in figure 4.2(d) shows that the MSE of the modified YIN based TDE algorithm and the MAMDF algorithm is the lower than other algorithms considered in this work.

When the separation between the microphones (d) is increased to 6.8 cm, the valid angles from the table 4.1 are 161°, 129°, 108.4°, 90°, 71.6°, 51° and 19° corresponding to the delays -3, -2, -1, 0, 1, 2 and 3 respectively. The expected delay for the female source is between -1 and 0. Since 110° is closer to 108.4°, the expected delay value is taken as -1. From the bar plots of figure 4.3(a), 4.3(b), 4.3(c) peak of all the algorithms is at -1



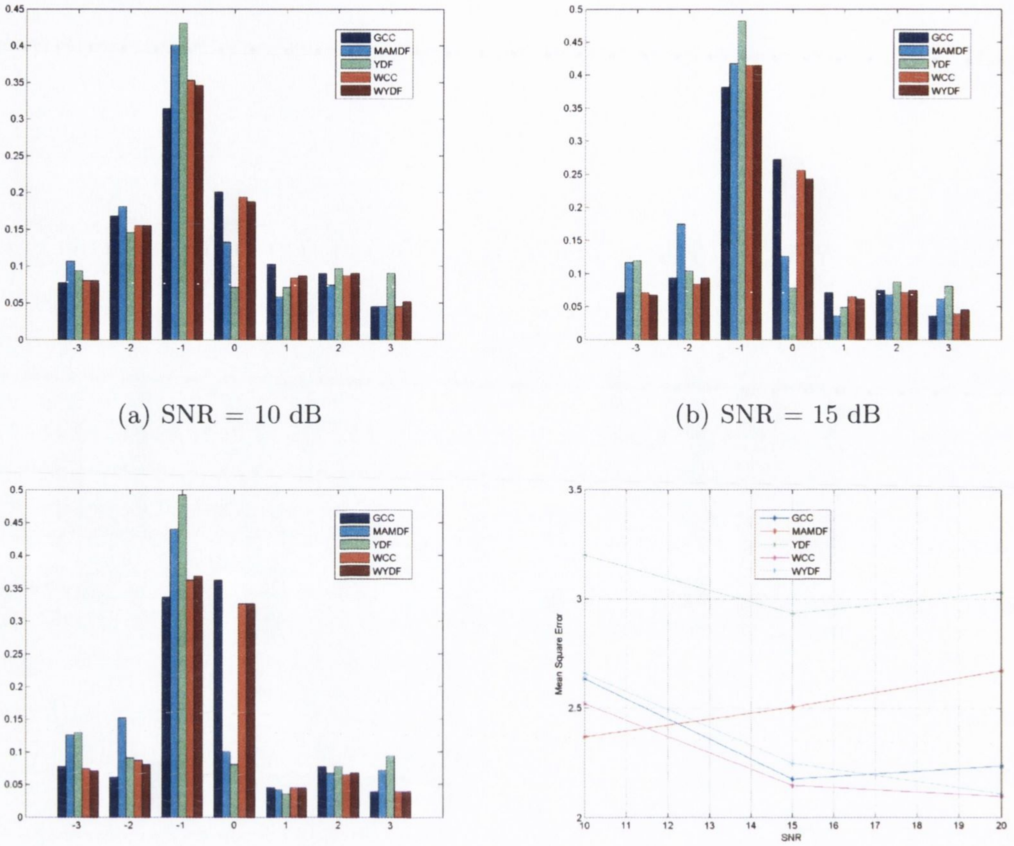
(a) SNR = 10 dB

(b) SNR = 15 dB

(c) SNR = 20 dB

(d) MSE plot

Fig. 4.2: Plot for female speaker with microphone separation (d) = 3.4 cm



(a) SNR = 10 dB

(b) SNR = 15 dB

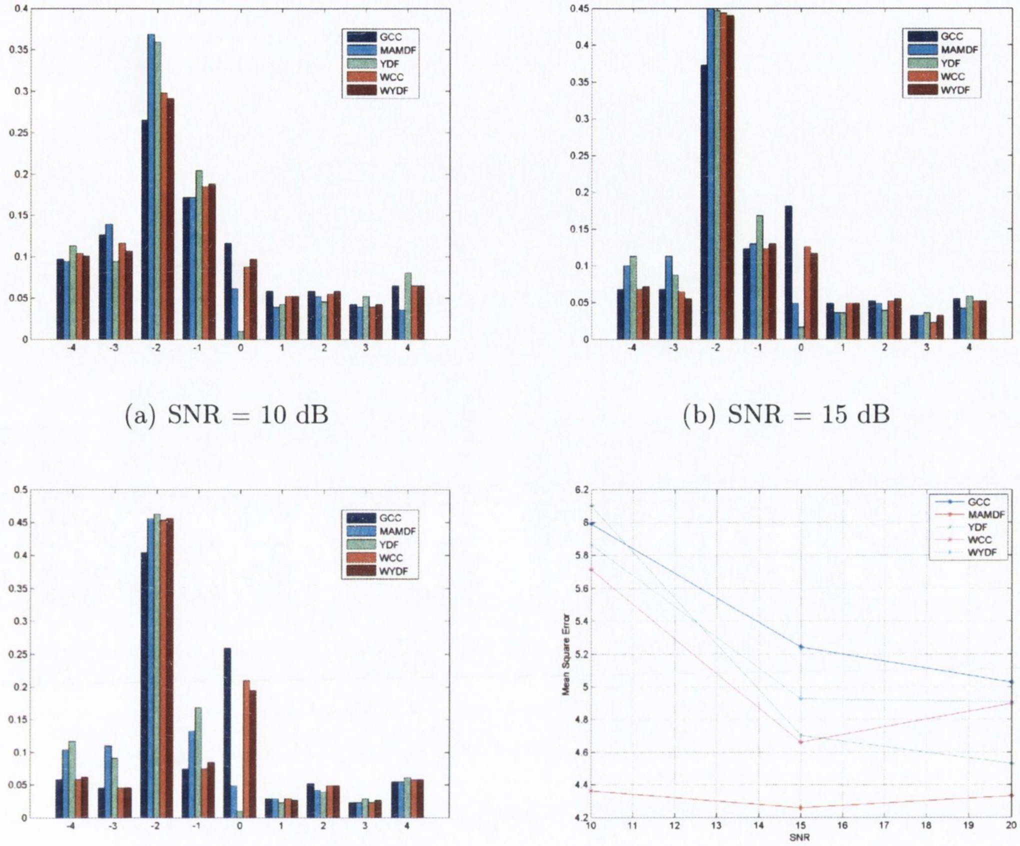
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.3: Plot for female speaker with microphone separation (d) = 6.8 cm

delay. At the -1 delay, the modified YIN based TDE (YDF) algorithm has the highest accuracy followed by the MAMDF. WCC and Weighted YIN based TDE (WYDF) algorithm have similar performance and GCC-PHAT has the lowest accuracy. From the MSE plot in figure 4.3(d), the WCC, Weighted YIN based TDE algorithm and GCC-PHAT have the lowest MSE and the modified YIN based TDE algorithm has the maximum MSE.

From the table 4.1, at d = 10.2 cm the expected time delay for the female source is between -1 and -2. The delay is closer to -2 hence the expected delay is taken as -2. Figures 4.4(a), 4.4(b), 4.4(c) present the accuracy of the algorithms at 10, 15 and 20 dB



(a) SNR = 10 dB

(b) SNR = 15 dB

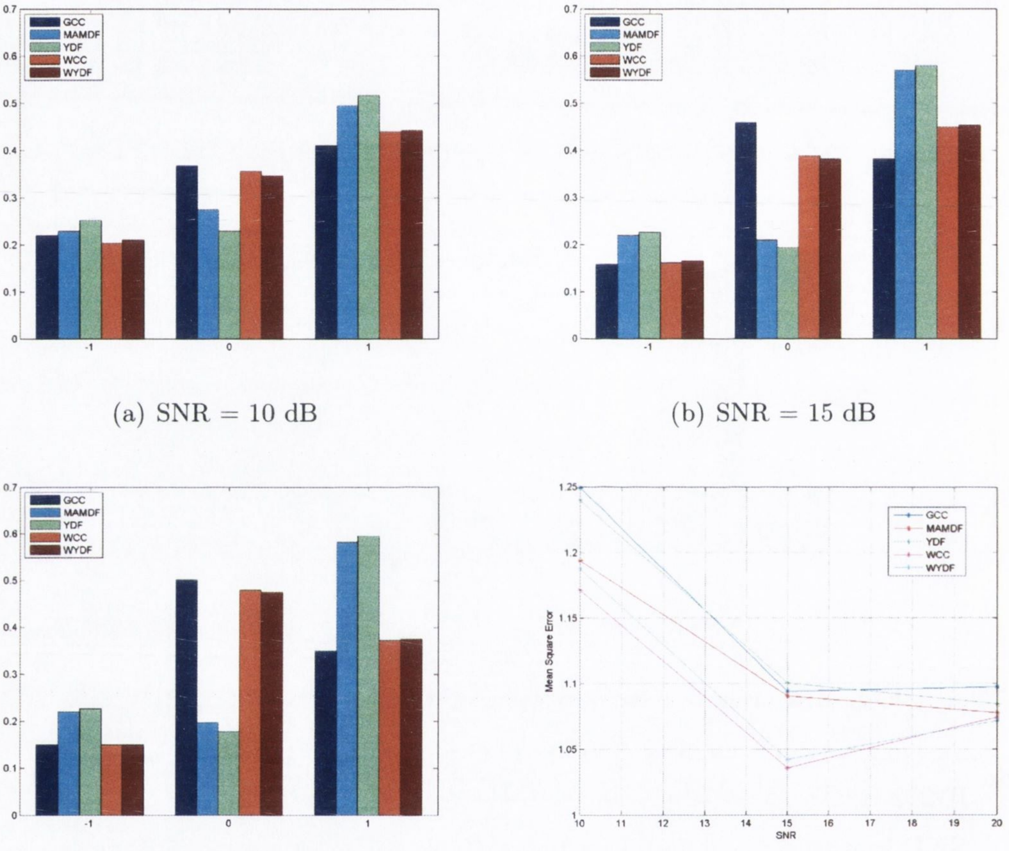
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.4: Plot for female speaker with microphone separation (d) = 10.2 cm

respectively. At l0 dB SNR, MAMDF has the highest accurate delay estimates. As the SNR increases modified YIN based TDE (YDF) algorithm, Weighted YIN based TDE (WYDF) algorithm and WCC perform equally well to MAMDF. The GCC-PHAT algorithm has the lowest accuracy. The MSE plot for different SNR is presented in figure 4.4(d). The GCC-PHAT algorithm has the highest MSE whereas the MAMDF algorithm has the lowest MSE. The MSE of the modified YIN based TDE algorithm decreases with the increase in SNR and has the lowest MSE after MAMDF at 20 dB.

Figure 4.5 shows the accuracy and MSE plots for second female source position at 49°. Figures 4.5(a), 4.5(b), 4.5(c) are the bar plots at 10, 15 and 20 dB respectively



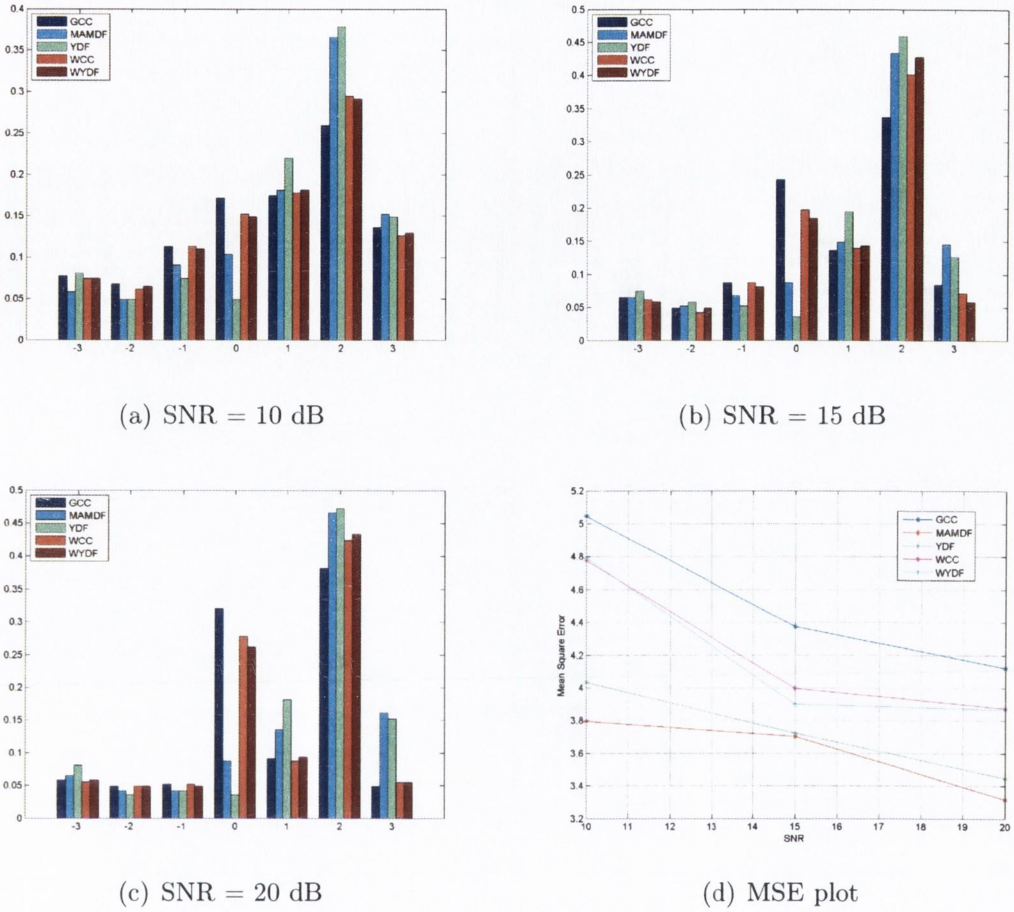
(a) SNR = 10 dB

(b) SNR = 15 dB

(c) SNR = 20 dB

(d) MSE plot

Fig. 4.5: Plot for female speaker with microphone separation (d) = 3.4 cm



(a) SNR = 10 dB

(b) SNR = 15 dB

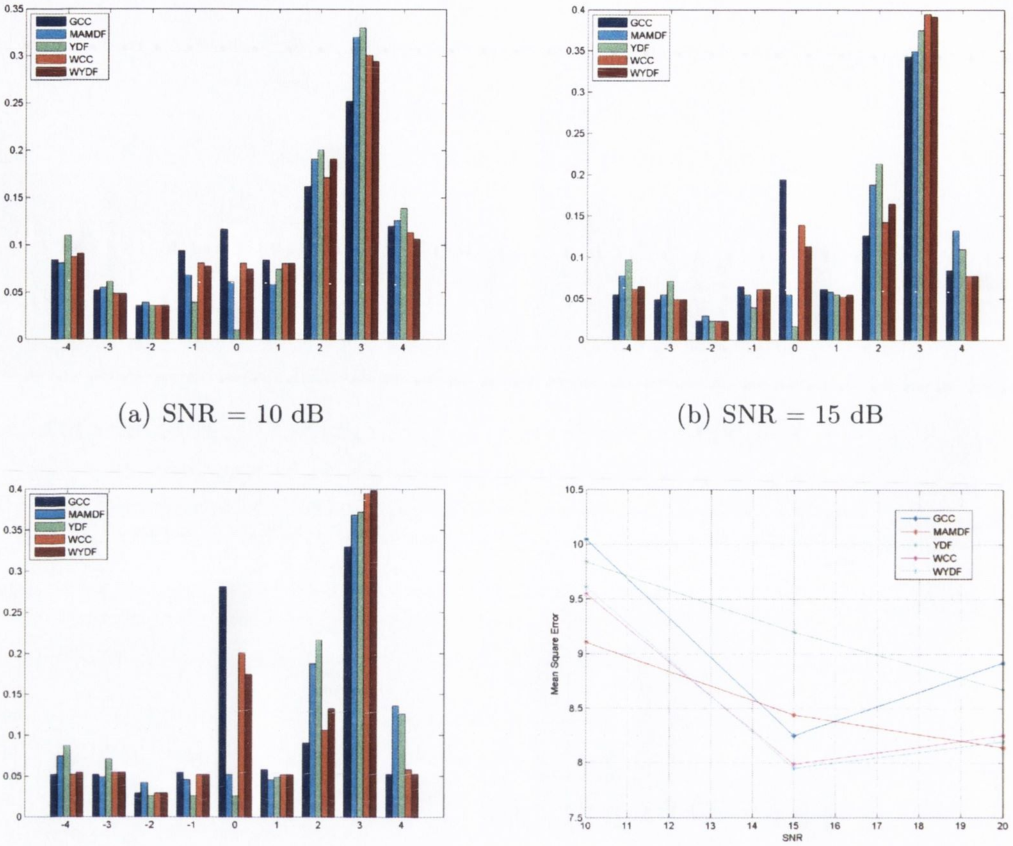
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.6: Plot for female speaker with microphone separation (d) = 6.8 cm

for d = 3.4 cm. From table 4.1, the expected delay is 1. Modified YIN based TDE algorithm (YDF) has the highest accuracy at all the SNR with MAMDF having the second highest accuracy. GCC-PHAT algorithm has the lowest accuracy. From the MSE plot of figure 4.5(d), Weighted YIN based TDE (WYDF) algorithm and WCC have the lowest MSE. All the other algorithms have similar MSE values. All the algorithms except GCC-PHAT have similar MSE value at 20 dB.

For d = 6.8 cm, the accuracy and MSE plots are presented in figure 4.6. From table 4.1 the expected delay at this position is 2. The bar plots in figure 4.6(a), 4.6(b), 4.6(c) for 10, 15 and 20 dB respectively show that modified YIN based TDE (YDF) algorithm



(a) SNR = 10 dB

(b) SNR = 15 dB

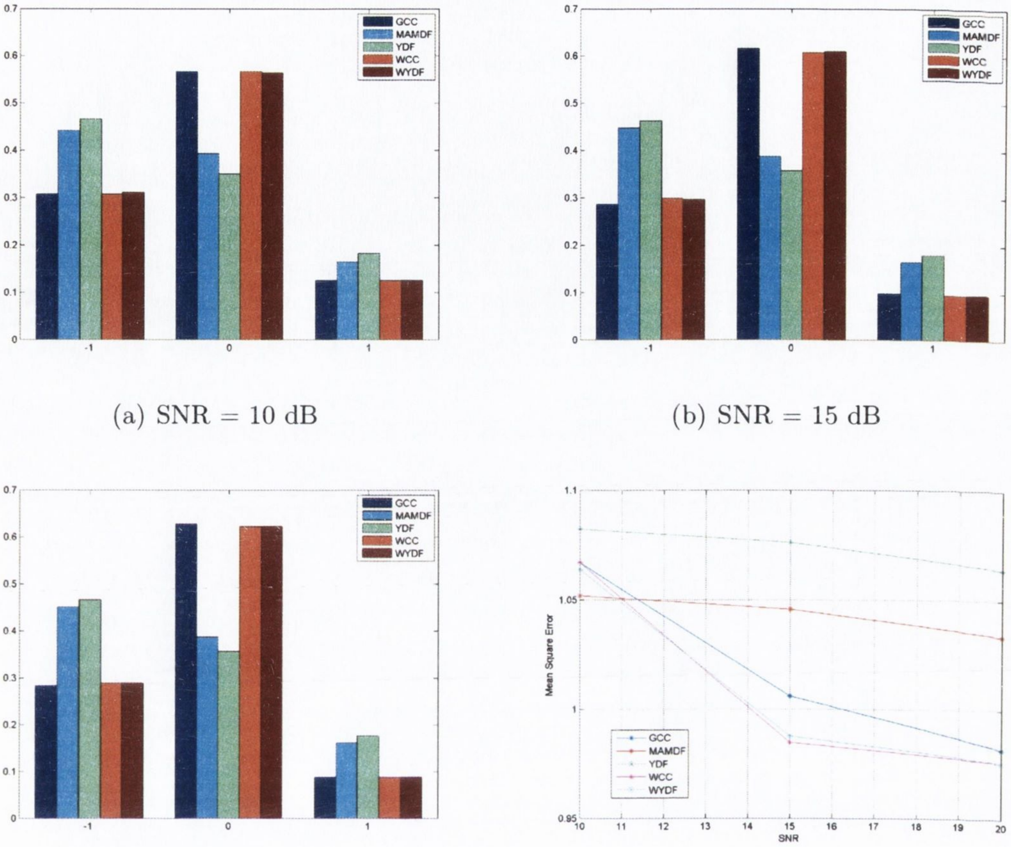
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.7: Plot for female speaker with microphone separation (d) = 10.2 cm

has the highest accuracy followed by the MAMDF algorithm. At high SNR, Weighted YIN based TDE (WYDF) algorithm performs better than WCC. GCC-PHAT has the lowest accuracy. The MSE is lowest for MAMDF followed by the modified YIN based TDE. At 15 dB SNR both the algorithms have approximately equal MSE value. GCC-PHAT algorithm has the highest MSE at all SNR. The WCC and Weighted YIN based TDE algorithm have similar MSE values.

The accuracy and MSE plots for $d = 10.2$ cm are presented in figure 4.7. From table 4.1 the expected delay at this position is 3. The bar plots in figure 4.7(a), 4.7(b), 4.7(c) for 10, 15 and 20 dB respectively show that the modified YIN based TDE (YDF) al-



(a) SNR = 10 dB

(b) SNR = 15 dB

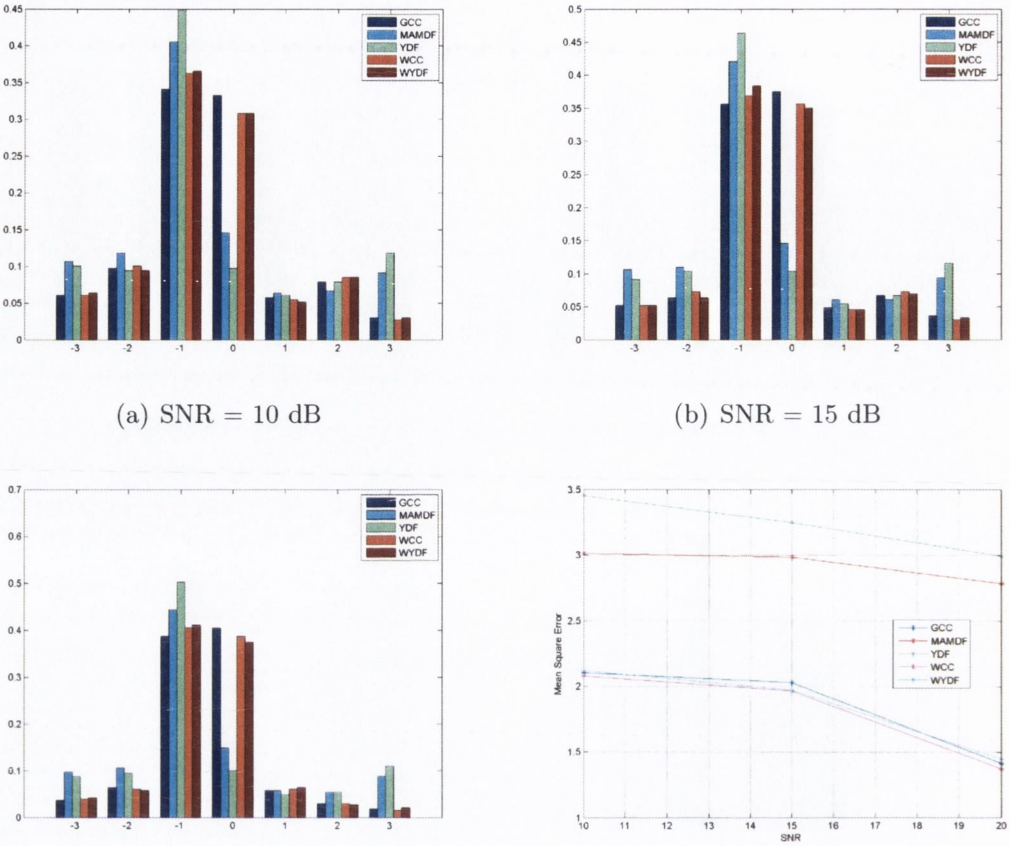
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.8: Plot for male speaker with microphone separation (d) = 3.4 cm

gorithm has the highest accuracy at 10 dB but at 15 and 20 dB the WCC and the Weighted YIN based TDE (WYDF) algorithm respectively have highest accuracy compared to the other algorithms. GCC-PHAT has the lowest accuracy. The MSE is lowest for the Weighted YIN based TDE algorithm and WCC. Modified YIN based TDE algorithm has the highest MSE that decreases with the increase in SNR.

For $d = 3.4$ cm, the accuracy and MSE plots for male source at 113.3° are presented in figure 4.8. The expected delay at this position is between -1 and 0. Since the angle is closer to 129° compared to 90° in value, the expected delay is taken as -1. The bar plots in figure 4.8(a), 4.8(b), 4.8(c) for 10, 15 and 20 dB respectively show that



(a) SNR = 10 dB

(b) SNR = 15 dB

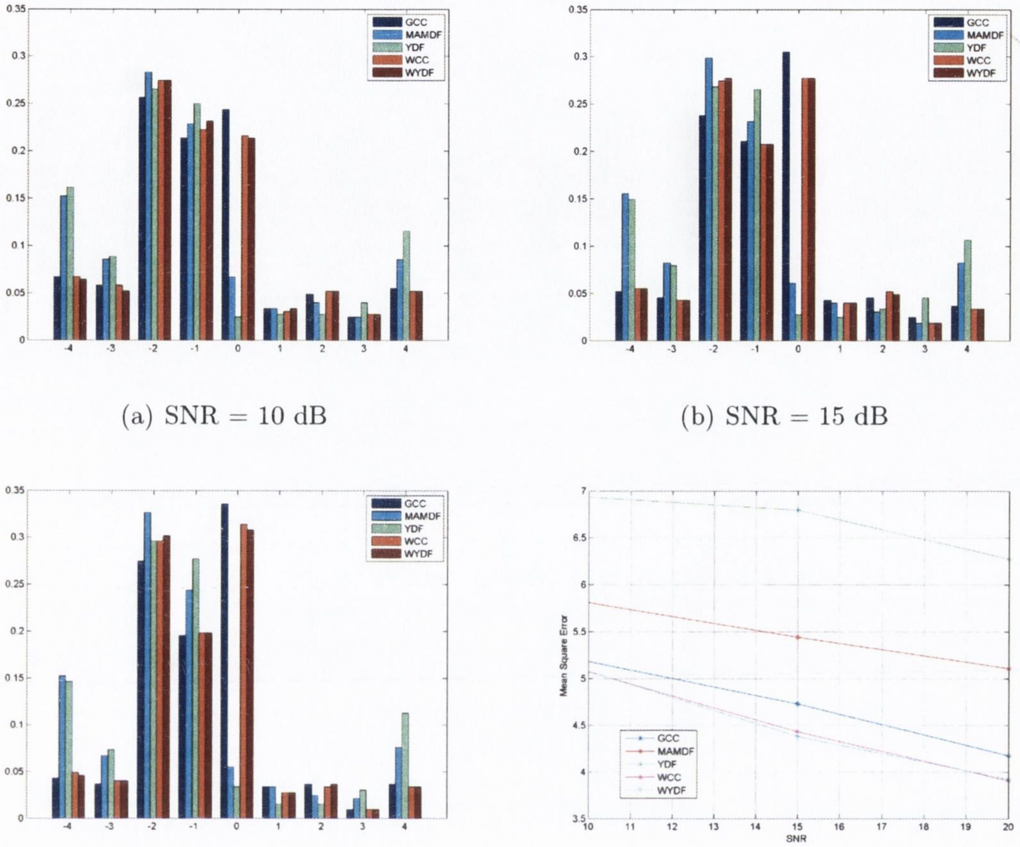
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.9: Plot for male speaker with microphone separation (d) = 6.8 cm

modified YIN based TDE (YDF) algorithm has the highest accuracy closely followed by the MAMDF algorithm. At high SNR, Weighted YIN based TDE (WYDF) algorithm performs slightly better than the WCC. GCC-PHAT has the lowest accuracy. From the MSE plot in figure 4.8(d) Weighted YIN based TDE algorithm and the WCC have the lowest MSE whereas the modified YIN based TDE algorithm has the highest MSE. The MSE of all the algorithms decreases with the increase in SNR.

For d = 6.8 cm, the accuracy and MSE plots are presented in figure 4.9. The expected delay at this position is between -2 and -1. Since the difference between the angle and 108.4° is smaller than the difference between the angle and 129° the expected



(a) SNR = 10 dB

(b) SNR = 15 dB

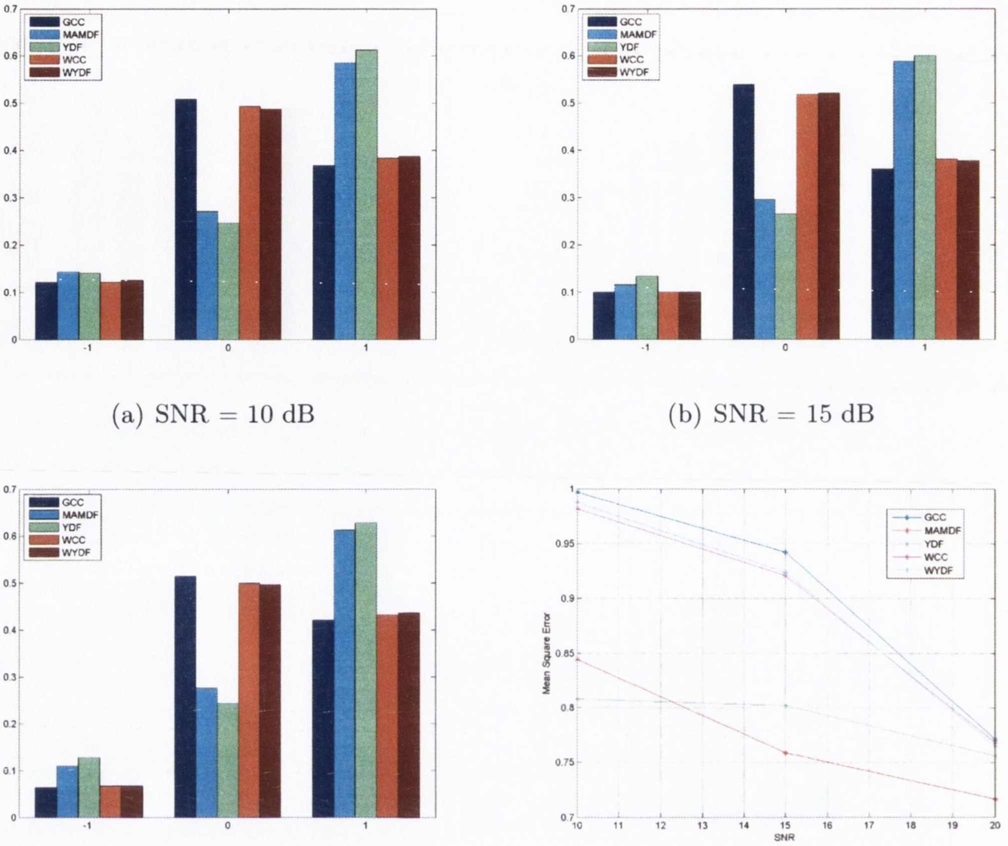
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.10: Plot for male speaker with microphone separation (d) = 10.2 cm

delay is taken as -1. The bar plots in figure 4.9(a), 4.9(b), 4.9(c) for 10, 15 and 20 dB respectively show that the modified YIN based TDE (YDF) algorithm has the highest accuracy followed by the MAMDF algorithm. At all the SNR values, Weighted YIN based TDE (WYDF) algorithm performs better than WCC. GCC-PHAT has the lowest accuracy. From figure 4.9(d), the MSE is lowest for the GCC-PHAT, the WCC and the Weighted YIN based TDE algorithm. It is highest for the modified YIN based TDE algorithm. The MSE for all the algorithms decreases with an increase in the SNR value.

The accuracy and MSE plots are for d = 10.2 cm are presented in figure 4.10.



(a) SNR = 10 dB

(b) SNR = 15 dB

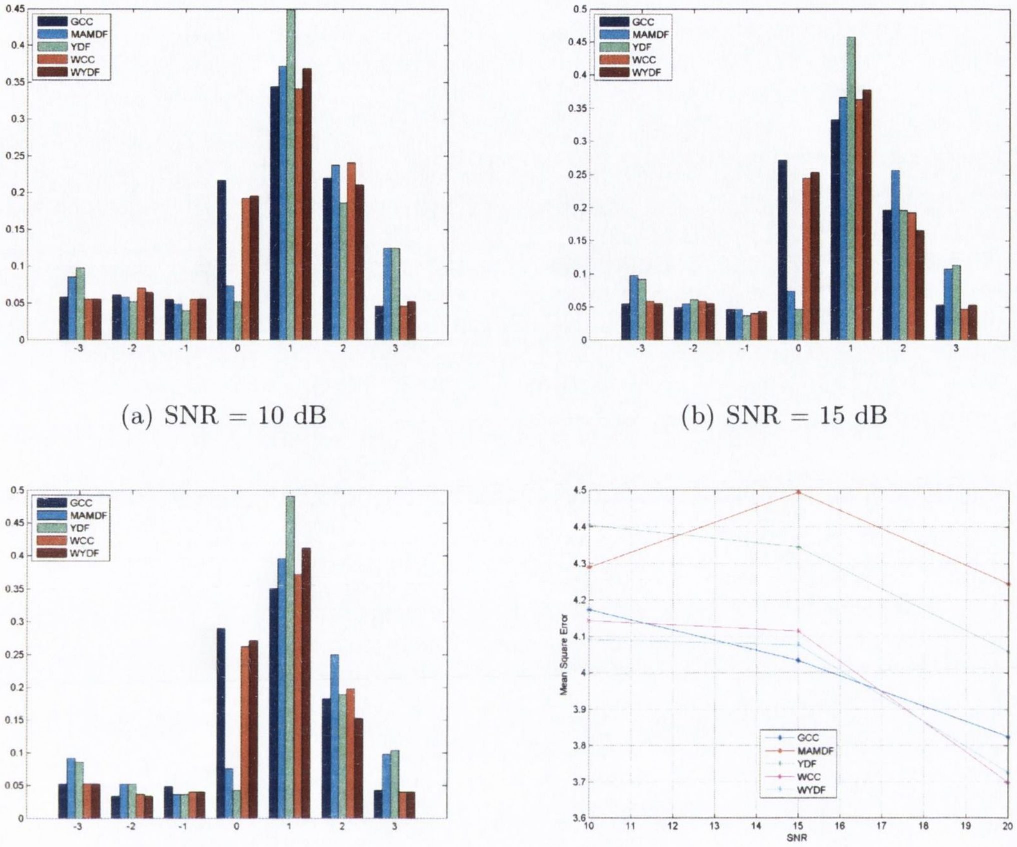
(c) SNR = 20 dB

(d) MSE plot

Fig. 4.11: Plot for male speaker with microphone separation (d) = 3.4 cm

The expected delay at this position is taken -2 since 113.3° is closer in value to 115° compared to 102°. The bar plots in figure 4.10(a), 4.10(b), 4.10(c) for 10, 15 and 20 dB respectively show that MAMDF algorithm has the highest accuracy at 10, 15 and 20 dB. The performance of the Weighted YIN based TDE (WYDF) algorithm and the WCC is similar and next to the MAMDF algorithm. GCC-PHAT has the lowest accuracy. The MSE is lowest for the Weighted YIN based TDE algorithm and WCC. The modified YIN based TDE (YDF) algorithm has the highest MSE. The MSE for all the algorithms tends to decreases with an increase in the SNR.

For d = 3.4 cm, the accuracy and MSE plots for male source at 64° are presented



(a) SNR = 10 dB

(b) SNR = 15 dB

(c) SNR = 20 dB

(d) MSE plot

Fig. 4.12: Plot for male speaker with microphone separation (d) = 6.8 cm

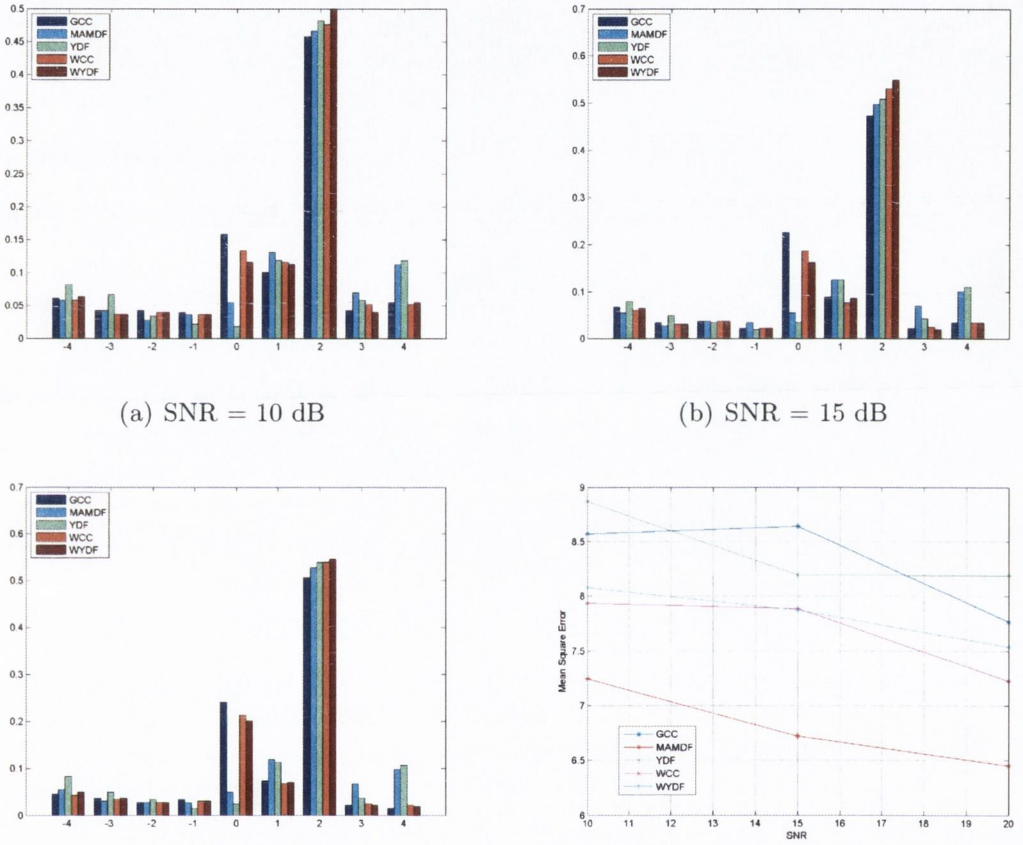
in figure 4.11. The expected delay at this position is taken as 1. The bar plots in figure 4.11(a), 4.11(b), 4.11(c) for 10, 15 and 20 dB respectively show that the modified YIN based TDE (YDF) algorithm has the highest accuracy followed by the MAMDF algorithm. The Weighted YIN based TDE (WYDF) algorithm and the WCC algorithm have similar performance. GCC-PHAT has the lowest accuracy. From the MSE plot in figure 4.11(d) MAMDF has the lowest MSE followed by the modified YIN based TDE algorithm. The Weighted YIN based TDE algorithm, the WCC and the GCC-PHAT have the highest MSE. The MSE of all the algorithms decreases when the SNR value increases.

For d = 6.8 cm, the accuracy and MSE plots are presented in figure 4.12. The expected delay at this position is between 1 and 2. The angle being closest to the angle for delay = 1, the expected delay value is taken as 1. The bar plots in figure 4.12(a), 4.12(b), 4.12(c) show the accuracy of the algorithms for SNR of 10, 15 and 20 dB respectively. The modified YIN based TDE (YDF) algorithm has the highest accuracy at all the SNR values. At 10 dB, the Weighted YIN based TDE (WYDF) algorithm and the MAMDF algorithms have the second highest accuracy whereas the GCC-PHAT and the WCC have the lowest accuracy. At 15 and 20 dB SNR, the Weighted YIN based TDE algorithm has the second highest accuracy and the GCC-PHAT has the lowest accuracy at this SNR. From figure 4.12(d) the MSE is highest for MAMDF. The GCC-PHAT, the WCC, the Weighted YIN based TDE algorithm have low MSE values. The MSE of all the algorithms decreases as the SNR value increases.

From the table 4.1, at d = 10.2 cm the expected time delay for the male source is 2. Figures 4.13(a), 4.13(b), 4.13(c) present the accuracy of the algorithms at 10, 15 and 20 dB respectively. The Weighted YIN based TDE (WYDF) algorithm has the highest accuracy at all the SNR values. At 10 dB SNR, the modified YIN based TDE (YDF) algorithm has the second highest accuracy. At 15 dB SNR the WCC has the second highest accuracy and at 20 dB SNR both the WCC and the modified YIN based TDE algorithms have the second highest accuracy. GCC-PHAT has the lowest accuracy at all the SNR values. The MSE plot for different SNR is presented in figure 4.13(d). The MAMDF algorithm has the lowest MSE. The MSE of all the algorithms decreases with an increase in the SNR value.

4.6 Observations and Conclusion

The number of valid delays is dependent on the separation between the microphones (d) (see equation 2.14 of section 2.2 and table 4.1). The number of valid delays increases



(a) SNR = 10 dB

(b) SNR = 15 dB

(c) SNR = 20 dB

(d) MSE plot

Fig. 4.13: Plot for male speaker with microphone separation (d) = 10.2 cm

with the increase in the separation between the microphones in the microphone pair. More number of delays improves the resolution of the angle estimates as the difference between the possible angle estimates is reduced. As mentioned in section 2.10.2, adjacent samples of speech signal are highly correlated. This strong correlation leads to incorrect delay estimation which will results in the decrease in the accuracy of the estimate and increase the MSE.

From the results presented in section 4.5 it is observed that an increase in the distance between the microphone pair (d) results in a decrease in the accuracy of the algorithms. The MSE of the algorithms for TDE increases with the decrease in the SNR. This is due to the presence of more noise in the signal which results in incorrect time delay estimates. Similarly increase in the SNR decreases the MSE value of the algorithms.

Among the three values of the separation between the microphones (d) (3.4 cm, 6.8 cm and 10.2 cm) considered in this work, 6.8 cm is a good compromise in terms of accuracy and angle resolution. This is based on the consideration that the TDE accuracy of the algorithms at this value is greater than that at 10.2 cm but lower than that at 3.4 cm. It also provides four more angles to improve the resolution in angles compared to the distance value of 3.4 cm and two angles less than the distance value of 10.2 cm.

From the simulation results when the inclination of the algorithms in the incorrect time delay estimates is compared the GCC-PHAT, the WCC and the Weighted YIN based TDE (WYDF) algorithm are biased towards the zero delay corresponding to an angle of 90°. These algorithms tend to have a large number of delay estimates at this value. The modified YIN based TDE (YDF) algorithm and the MAMDF algorithms tend to have large number of incorrect time delay estimates at the extreme delay values.

From the simulation results in section 4.5 the GCC-PHAT, the WCC and the Weighted YIN based TDE algorithms have small MSE. This is due to their bias towards the zero delay. The WCC and the Weighted YIN based TDE algorithms are better than the GCC-PHAT algorithm but not as good as the modified YIN based TDE algorithm and the MAMDF in terms of the accuracy. The modified YIN based TDE algorithm and the MAMDF emerge as strong contenders for the DOA estimation using TDE. In terms of the MSE, the MAMDF algorithm is slightly better than the modified YIN based TDE algorithm as it has lower MSE but in terms of the accuracy the modified YIN based algorithm is better than the MAMDF algorithm.

Chapter 5

VAD Feature Selection Studies

The secondary contribution of this thesis is a Voice Activity Detection (VAD) system that classifies the signal into four classes for source localization and post-processing e-learning. As described in section 3.4, identifying the features for classification and applying feature selection on these features to obtain the best feature subset is first of the two steps in designing a voice activity detection (VAD) system. The first section of this chapter introduces a Teager Energy Operator (TEO) based feature called the Mel-Spectrum Teager Energy (MSTE) coefficients for classification. The next section describes the simulation setup that was used to generate the data for feature selection using features described in chapter 3. This is followed by a section presenting the results for supervised machine learning algorithm selection comparing the performance of Discriminant Analysis (DA), Cascade Artificial Neural Network (ANN) and Support Vector Machine (SVM) classifiers for VAD. Feature selection results for the feature subset consisting of the proposed feature and other features from chapter 3 that enhance the classification performance is presented next. This is followed by a section presenting the simulation results for classification comparing the performance of the selected classification algorithms using the selected feature set. In the end observations and conclusion of the chapter is presented.

5.1 Mel-Spectrum Teager Energy Coefficients

Teager Energy based features are used in signal processing applications such as speech recognition [123, 56, 124] and noise suppression [122]. Teager energy gives a measure of the energy required to produce a signal [63]. Teager's Energy Operator (TEO) for a signal is non-linear and was introduced by Kaiser [63]. It is based on the fundamentals of simple harmonic motion where the energy required to produce sinusoidal oscillations is proportional to the square of the product of the amplitude and frequency of the signal. TEO of the signal is given by:

$$\Psi[x(t)] \triangleq [\dot{x}(t)]^2 - x(t)\ \ddot{x}(t) \tag{5.1}$$

where $x(t)$ is the input signal, $\dot{x}(t)$ is $\frac{dx(t)}{dt}$ and $\ddot{x}(t)$ is $\frac{d^2x(t)}{dt^2}$.

For the discrete signal, $x(n)$, TEO ($\Psi_d[x(n)]$)is given by:

$$\Psi_d[x(n)] \triangleq x(n)^2 - x(n-1)\ x(n+1) \tag{5.2}$$

where $x(n)$ is the input signal. TEO represents the local property of the signal as for each point the computed TEO involves only three signal samples which are the signal sample and its adjacent samples.

For oscillating input signal $x(n) = Acos(\Omega n + \phi))$, the TEO is given by:

$$\Psi_d[x(n)] \triangleq A^2 sin^2(\Omega) \tag{5.3}$$

where A is the amplitude of the signal, ϕ is the initial phase in radians and Ω is the digital frequency of the signal in *radians/sample* [63]. Ω is related to the sampling frequency f_s and the signal frequency f by $\Omega = \frac{2\pi f}{f_s}$. The relation $\Psi[x(n)] = A^2 sin^2(\Omega)$ holds only if $\Omega < \frac{\pi}{2}$. $\frac{\pi}{2}$ is equivalent to $\frac{f_s}{4}$.

The Mel-Spectrum Teager Energy (MSTE) coefficients used in this work involves finding the average TEO of the signal spectral sub-bands. The signal spectrum sub-bands are the ones used in computing the MFCC introduced in section 3.3. Coefficients similar to MSTE have been introduced by Firas et al. [56, 57] as an intermediate stage for computing the TEOCEP coefficients and have been named as sub-signal average Teager energy. The difference between the sub-signal average Teager energy and MSTE coefficients is that the sub-bands for both the coefficients and the approach adopted to compute the coefficients is different. The sub-bands used to compute the TEOCEP coefficients are the ones proposed by Erzin et al. [41]. Bandpass filters are used to obtain the sub-band and compute the sub-signal average Teager energy in the time domain. All the sub-signal average Teager energy is then log compressed and inverse Discrete Cosine Transformed (DCT) to obtain the TEOCEP coefficients.

The Mel-Spectrum Teager Energy coefficients for VAD are computed by a simpler approach as introduced in the paper by Kavanagh et al. [65]. Firstly, the signal spectrum ($X(k)$) is obtained by the Short Term Fourier Transform(STFT) of the signal. The TEO of the spectrum is computed on the magnitude (amplitude) spectrum of the signal using the relation between the TEO, amplitude and signal frequency in equation 5.3 resulting in the non-linear energy spectrum of the signal. The non-linear energy spectrum of the signal is divided into sub-bands based on the mel-scale. The center frequencies of the critical bands for MFCC presented in table 3.1 form the limiting frequencies of each sub-band. The average energy of each sub-band is the computed to obtain the MSTE coefficient for each sub-band. The MSTE coefficient for each sub-band is given by:

$$\Psi_{mel}(\alpha) = \frac{1}{\beta} \sum_{k=l_{(\alpha)}}^{p_{(\alpha)}} \Psi_d(k) \tag{5.4}$$

where α are the number of mel sub-bands and $\alpha = 1, 2, \cdots, 20$. $l(_{\alpha})$ and $p(_{\alpha})$ are the lower and upper limit of α mel sub-band respectively. β is the difference between the upper and lower limit of α mel sub-band. For a signal frame, twenty MSTE coefficients are computed.

Another variation in using the MSTE coefficients for VAD is that the coefficients are computed on the LP residual of the signal instead of computing on the actual signal. LP residual is discussed in section 3.3.

5.2 Setup

The VAD system aims to distinguish between four type of signals. They are the voiced speech, unvoiced speech, non-stationary noise and the stationary noise. The voiced speech and the unvoiced speech sample signals were obtained from the TIMIT [6] database. Speech signals were labeled as voiced speech and unvoiced speech from the table available in [64]. The stationary (background) noise was the white Gaussian noise generated from the Cool Edit Pro 2.0 software [7]. The stationary noise is included to represent signals such as computer fan noise and air conditioner noise. The non-stationary noise signals were recorded in a classroom. These signals included human produced sounds like coughs, sneezes, clearing of throat and non-human produced sounds generated by coins, keys, shutting of doors, plastic bags, books falling and paper shuffle.

Two data sets consisting of the features for all the four classes were generated for simulations. The data set used to train the classification algorithms is called the training data and the data set used to evaluate the performance of the classification algorithms is called the testing data. The data included in the training data and the testing data were different. The signals used in this work were sampled at 16 kHz.

Features were calculated on 20ms long signal frames with no overlap. Feature selection was performed by Linear Discriminant Analysis (LDA) in the SPSS software [8]. Both forward and backward selection methods were employed to find the best feature subset. The features of different signal classes were generated in MATLAB [9]. For DA classifier, the DA implementation in the statistical toolbox [10] of MATLAB was used. The libsvm [11] was used for the SVM classifier. The FANN [12] library was used for the neural network classifier. The ANN was created using the cascade-classifier supervised training algorithm [43] which determines the size and the topology of the ANN itself.

The parameters used to evaluate the performance of the classifiers for VAD are the Hit rate (HR) and False Alarm Rate (FAR) for each class and the overall Hit Rate for VAD. The HR and the FAR are introduced in section 3.4.1.

5.3 Supervised Machine Learning Algorithm Selection for VAD

This section describes the process that was employed to select the supervised machine learning algorithms for VAD in this work. Initially VAD simulations study was performed on the supervised machine learning algorithms discussed in section 3.6.3 with a few signal features to compare their performance for classifying the signal into four classes. The study provides an indication of the performance of the algorithms for VAD in this work. Different feature set (consisting of additional features and some of the features from the studies in this section) was employed to compare the performance of the selected supervised machine algorithms for VAD in section 5.4. Discriminant Analysis (Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), Mahalanobis Distance (MD)), cascade Artificial Neural Networks (ANN), radial Support Vector Machine (SVM), linear SVM, polynomial SVM and sigmoid SVM supervised machine learning algorithms were considered for VAD.

The signal features considered were the short term signal energy, zero-crossing rate, second and third LP predictor coefficients, absolute value of the difference between the first two PARCOR coefficients, low to high frequency band energy ratio, LP residual energy, skewness, kurtosis, pitch, first three formants and spectral slope. The pitch and the first three formants were computed using the MATLAB code available in the speech processing toolbox COLEA [13].

	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise	
	HR	FAR	HR	FAR	HR	FAR	HR	FAR
QDA	89.4	12.9	82.8	8.1	72.1	13.9	42.8	3.7
LDA	89.4	7.4	81.9	8.1	62.6	4.8	94.5	6
MD	88.6	12.4	90	13.7	69.1	15.7	0.1	0.4
ANN	81.1	16.7	89.7	26.5	0	0	98.7	8.3
Radial SVM	89	5.2	83.4	7.2	77	10.8	70.6	2.4
Linear SVM	89	4.8	82.5	7	77.3	19.6	35.8	2.1
Polynomial SVM	88.1	5.7	83	7.9	74.8	14	56.1	2.6
Sigmoid SVM	34.9	3.5	57.3	7.3	16.8	5	99.9	52.2

Table 5.1: The Hit Rate (HR) and False Alarm Rate (FAR) for classifying the test data into four classes.

VAD experiments were performed on audio signal consisting of four types of signals namely voiced speech, unvoiced speech, non-stationary noise and stationary noise. The training data was approximately 10.5 minutes long. The number of cases for the noise and speech class in the training data was approximately equal. Testing data was 10 minutes long. The frames consisting of two or more signal classes were not included in both the training and the testing data.

From the forward and backward feature selection using LDA in SPSS, all the features were found to be significant for VAD. These features were used as input to the machine learning algorithm for training and testing. The results for each machine learning algorithm presenting the Hit Rates (HR) and False Alarm Rates (FAR) for all the classes and the overall HR are presented in table 5.1. HR and FAR are described in section 3.4.1.

A good classifier generally has a high HR for every class, a low FAR for every class and a high overall HR. A high HR for each class indicates that the signals are being correctly classified by the classifier into their respective classes with high accuracy. A low FAR for a class indicates that signals of other class are not being incorrectly classified into the class. High overall HR indicates that the classifier has high classification accuracy.

From the simulation results in table 5.1 QDA had a low HR of 42.8% for the stationary noise. MD had a 0.1% HR for stationary noise. The neural network was not able to classify any of the non-stationary noise frames correctly as it had a HR of 0%. Linear SVM had a low HR of 35.8% for the stationary noise and a high FAR of 19.6% for non-stationary noise. Polynomial SVM had a low HR of 56.1% for stationary noise and a high FAR of 14% for non-stationary noise. Sigmoid SVM had a high HR of 99.9% for only stationary noise. But it also had a high FAR of 52.2% for the stationary noise signal. Based on the studies to classify the signal into four classes QDA, MD, neural network, polynomial, linear and Sigmoid SVM were not considered further for the VAD system in this work.

5.4 Features and Feature Selection Results

The features considered in the feature set after the simulation on the supervised machine learning algorithm selection for VAD are:

1. The MSTE coefficients.
2. MFCC.
3. Log energy.
4. Zero-crossing rate.
5. Second and third LP predictor coefficients.
6. Log of LP residual energy.
7. Skewness and kurtosis of the LP residual signal.

8. Absolute value of the difference of the first two PARCOR coefficients.

9. Skewness and kurtosis of the signal.

10. Log of the low to high band energy.

11. Log of the low to high band energy product.

12. Log of low band to full band energy ratio.

13. Log of the low band to full band energy product.

14. Pitch.

15. The first three formants.

16. Spectral centroid.

17. Spectral roll-off point.

18. Normalized autocorrelation coefficient at unit sample delay.

19. Spectral slope.

20. First ten Cepstrum coefficients.

The details of these features were presented in section 3.3.

Initial classification results comparing the performance of the LDA and radial SVM indicated that the MSTE coefficients and MFCC on their own are able to classify the signal into four classes with a high HR for each class and overall classification. Thus features that improved the classification rate along with MSTE coefficients or MFCC were added to the feature subset using forward selection in LDA. MFCC coefficients were computed using the MATLAB module from [14].

The features that were found to be significant in the feature selection process and were included in the feature subset are:

1. Log energy.

2. Zero-crossing rate.

3. Second and third LP filter coefficients.

4. Skewness of the LP residual.

5. Spectral centroid.

6. The absolute difference between the first two PARCOR coefficients.

5.5 VAD Four Class Classification Results

The simulation results comparing the performance of the LDA and the radial SVM with 3 feature sets have been presented at 10, 15 and 20 dB SNR where the classification algorithms were trained at the same SNR values. The training and the testing data set were a subset of the training and testing data used in section 5.3. This was done in order to have approximately equal number of data for all the classes in both the data sets. The training data was 31.4 s long whereas the testing data was 31.7 s long. The Signal to Noise Ratio (SNR) is introduced in section 2.6. The signal power in equation 2.22 is the power of the source signal and the noise power corresponds to the power of the noise signal generated in Cool Edit Pro 2.0. The SNR values used in the simulations are 10, 15 and 20 dB similar to the values used in TDE simulations of chapter 4.

Classification results at each SNR with six different feature set are presented for Linear Discriminant Analysis (LDA) and Radial SVM (RSVM). The first feature set consisted of only MFCC, the second feature set had only MSTE coefficients and the third feature set called the Other Feature Set (OFS) consisted of log energy, zero-crossing rate, skewness of the LP residual, spectral centroid, absolute difference between

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	89.98	4.97	82.13	6.26	75.41	5.66	100	0.25	87.13
MSTE	92.36	5.15	86.6	6.68	74.86	2.46	100	0.75	88.77
OFS	78.28	4.63	84.37	3.04	81.42	8.03	100	2.86	85.99
MFCC + OFS	92.36	3.17	88.59	3.55	92.35	2.13	100	0.08	93.31
MSTE + OFS	92.84	3.26	87.59	3.98	92.1	1.89	100	0.08	93.12
MSTE + MFCC + OFS	93.08	2.74	87.84	3.64	93.44	2.13	100	0.08	93.56

Table 5.2: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 10 dB.

the first two PARCOR coefficients and the second and third LP coefficients. The fourth feature set consisted of the MFCC and the features of the OFS. The fifth feature set had the MSTE coefficients with the features of the OFS. The final feature set consisted of all the MFCC, MSTE coefficients and features of the OFS.

From table 5.2, both MFCC and MSTE coefficients on their own to be able to classify signals of all the classes better than the non-stationary noise class signals at 10 dB SNR when LDA is employed for classification. In the case of speech signals MSTE coefficients classify both voiced and unvoiced signals better than the MFCC. Other Feature Set (OFS) classifies unvoiced signals better than the MFCC and non-stationary noise signals better than both MFCC and MSTE coefficients. The performance of the MFCC is improved when OFS is included in the feature set. Similarly the performance of MSTE coefficients improves when OFS are included in the feature set. In both the cases the False Alarm Rate (FAR) also is reduced for all the classes. The improvement in classification of the non-stationary noise signal is noticeable. The classification performance with all the features together (MFCC, MSTE coefficients and OFS) is

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	89.49	4.89	82.88	6.43	75.68	5.66	100	0	87.26
MSTE	91.41	5.06	87.59	7.02	75.14	2.54	100	0.3	88.83
OFS	75.89	4.2	83.13	3.05	82.79	8.2	100	3.96	85.36
MFCC + OFS	92.36	3	89.08	3.55	93.17	1.97	100	0	93.62
MSTE + OFS	92.36	2.49	89.09	4.57	93.72	2.46	100	0	92.81
MSTE + MFCC + OFS	91.65	2.32	88.59	3.55	95.08	2.21	100	0.25	93.75

Table 5.3: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 15 dB.

marginally improved compared to the case where OFS is used with either MFCC or MSTE coefficients.

Table 5.3 presents the classification results using LDA for all the six feature sets at 15 dB SNR. The performance of MFCC and MSTE coefficients is similar to their performance at 10 dB SNR. Except for the non-stationary noise signal the features on their own are able to classify signals of other classes with HR greater than 82%. The MSTE coefficients classify both voiced and unvoiced signals better than the MFCC. When OFS is included in the feature set the HR increases for all the classes and the FAR decreases for all the classes except the stationary noise where the FAR is already 0% compared to when only MFCC features are used on their own. Similarly the performance of MSTE coefficients improves when OFS are included in the feature set. Here too the improvement in classification of the non-stationary noise signal is more than in the case of other classes. The classification performance with all the features together (MFCC, MSTE coefficients and OFS) is marginally improved compared to the case where OFS is used with either MFCC or MSTE coefficients.

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	89.26	4.63	82.88	6.77	75.41	5.74	100	0	87.19
MSTE	91.65	4.72	87.34	7.45	74.59	2.71	100	0.25	88.7
OFS	70.88	2.74	83.87	3.21	78.96	8.86	100	7.24	83.34
MFCC + OFS	91.41	2.32	90.32	3.64	95.36	1.8	100	0	94.19
MSTE + OFS	91.65	2.4	90.32	3.98	94.81	1.39	100	0.08	94.13
MSTE + MFCC + OFS	91.65	1.89	89.08	3.64	96.45	2.13	100	0.08	94.2

Table 5.4: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for LDA at SNR of 20 dB.

Table 5.4 presents the classification results using LDA for all the six feature sets at 20 dB SNR. The performance of MFCC and MSTE coefficients is similar to their performance at 10 and 15 dB SNR. In this case too the MSTE coefficients classify both voiced and unvoiced speech signals better than the MFCC. When OFS is included with MFCC in the feature set the HR increases for all the classes and the FAR decreases for all the classes except stationary noise where the FAR is already 0% compared to when only MFCC features are used on their own. Including the OFS feature in the feature set along with MSTE coefficients increases the FAR for stationary noise slightly. The improvement in HR of the non-stationary noise is higher compared to that of other classes. The classification performance with all the features together (MFCC, MSTE coefficients and OFS) is marginally improved compared to the case where OFS is used with either MFCC or MSTE coefficients.

From table 5.5, both MFCC and MSTE coefficients are able to classify signals of all the classes better than the non-stationary noise signal at 10 dB SNR when radial SVM is employed for classification. The classification rates are lower than that of the LDA at

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	90.93	4.63	81.64	6.85	76.5	5.17	100	0	87.51
MSTE	92.12	4.89	85.61	6.43	78.42	3.04	100	0	89.27
OFS	79.71	3.17	84.62	3.05	90.98	8.37	100	0.42	88.64
MFCC + OFS	92.36	3.6	90.07	4.91	87.71	1.4	100	0	92.62
MSTE + OFS	92.84	3.69	88.59	3.64	90.71	1.97	100	0	93.06
MSTE + MFCC + OFS	92.36	3.26	91.81	4.06	90.16	1.15	100	0.08	93.63

Table 5.5: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 10 dB.

the same SNR. In case of the speech signals MSTE coefficients classify the voiced speech better than the MFCC. OFS classifies non-stationary signals better than the MFCC and the MSTE coefficients. The performance of the MFCC is improved when OFS is included in the feature set with a decrease in the FAR for unvoiced speech. The HR for non-stationary noise increases. Similarly the performance of the MSTE coefficients improves when OFS is included in the feature set. In both the cases the False Alarm Rate (FAR) is also reduced for all the classes except for the stationary noise where the FAR remains unchanged. The improvement in classification HR of the non-stationary noise signal is noticeable. The classification performance with all the features together (MFCC, MSTE coefficients and OFS) is marginally improved compared to the case where OFS is used with either the MFCC or the MSTE coefficients.

Table 5.6 presents the classification results using radial SVM for all the six feature sets at 15 dB SNR. The performance of MFCC and MSTE coefficients is similar to their performance at 10 dB SNR. In this case too the MSTE coefficients classify both voiced and unvoiced signals better than the MFCC. OFS are unable to classify voiced speech

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	90.69	4.55	81.39	6.77	76.78	5.41	100	0	87.45
MSTE	92.12	4.89	87.1	6.68	77.87	2.46	100	0	89.53
OFS	76.37	2.49	83.87	2.96	93.17	9.84	100	0.42	88.08
MFCC + OFS	92.12	3.09	90.32	4.48	89.62	1.72	100	0	93.06
MSTE + OFS	92.6	3	89.08	3.64	92.62	1.97	100	0	93.57
MSTE + MFCC + OFS	92.84	3.09	91.32	3.72	92.35	1.07	100	0	94.13

Table 5.6: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 15 dB.

as well as the signals of other classes and has a higher FAR for non-stationary noise compared to other signal classes. When OFS and the MFCC are included in the feature set the HR for voiced speech, unvoiced speech and non-stationary noise increases and the FAR for these classes decreases compared to using only MFCC. In the case of MSTE coefficients with OFS in the feature set, the FAR for voiced speech, unvoiced speech and non-stationary noise decreases and the HR for these classes increases compared to the performance of only MSTE coefficients. The classification performance with all the features together (MFCC, MSTE coefficients and OFS) is marginally improved compared to the case where OFS is used with either MFCC or MSTE coefficients.

Table 5.7 presents the classification results using radial SVM for all the six feature sets at 20 dB SNR. Among the MFCC, MSTE coefficients and OFS feature sets, MSTE coefficients have the best overall HR. It also has the best HR for voiced speech and unvoiced speech class. OFS has the highest HR for the non-stationary noise class. Both MFCC and MSTE coefficients have lower HR for non-stationary noise compared to the HR for other classes. In the case of OFS with MFCC or MSTE coefficients, the HR

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
MFCC	90.69	4.46	80.89	6.68	77.05	5.66	100	0	87.38
MSTE	92.12	4.72	87.1	6.6	78.7	2.46	100	0	89.72
OFS	75.66	2.4	84.37	3.05	93.44	9.76	100	0.51	88.08
MFCC + OFS	92.36	2.92	90.32	4.48	89.62	1.81	100	0	93.12
MSTE + OFS	92.6	2.83	89.33	3.72	92.9	1.89	100	0	93.69
MSTE + MFCC + OFS	92.6	2.74	91.56	3.81	93.44	0.98	100	0	94.39

Table 5.7: The hit rate (HR), false alarm rate (FAR) for each class and overall hit rate for Radial SVM at SNR of 20 dB.

increases and FAR decreases for all the classes except the stationary noise where the HR and FAR remains unchanged. There is a marginal improvement in the classification with all the features together (MFCC, MSTE coefficients and OFS) compared to the case where OFS is used with either MFCC or MSTE coefficients.

5.6 Observations and Conclusion

From the classification tables in section 5.4 it can be seen that for all SNR's the corresponding overall classification HR for all MFCC, MSTE coefficients and OFS feature sets in LDA and radial SVM are below 90%. When OFS is included in the feature set with MFCC or MSTE coefficients the overall HR is above 90%. When all the three features are included in the feature set, the overall HR for both the classifiers improves marginally compared to the classifiers performance when OFS with MFCC or MSTE coefficients feature sets are employed. All the feature sets are able to classify the stationary noise signal with a high HR and very low FAR in case of both the classification algorithms.

For the other three classes (voiced speech, unvoiced speech and non-stationary noise), both MFCC and MSTE coefficients have a higher HR for voiced speech compared to the unvoiced speech and have the lowest HR for the non-stationary noise for both the classification algorithms. MSTE coefficients have a higher HR for both the classes (voiced and unvoiced speech) compared to the MFCC. OFS has a high HR and a high FAR for the non-stationary signal class for both the classification algorithms. It's HR for non-stationary noise is higher for radial SVM compared to that of the LDA. It has the lowest HR for the voiced speech for both the classification algorithms.

MFCC with OFS improves the HR for the voiced, unvoiced and non-stationary noise classes compared to the HR of that of only MFCC. The improvement is more noticeable in the case of non-stationary noise and unvoiced speech signal classes. MSTE coefficients with OFS also have improved HR for non-stationary noise and unvoiced speech classes compared to the HR with only MSTE coefficients. The performance of MFCC with OFS, MSTE coefficients with OFS is similar to each other. The overall classification HR is marginally improved when MFCC, MSTE coefficients and OFS are included in the feature set.

For MFCC with OFS, LDA and radial SVM have similar HR for all the classes. In the case of MSTE coefficients, both linear and radial SVM have similar overall HR. LDA has a better HR than radial SVM for non-stationary noise for feature sets of MFCC with OFS, MSTE coefficients with OFS and OFS with MFCC and MSTE coefficients. The overall HR for both the classifiers is almost the same with the HR of LDA marginally better than that of the radial SVM. Thus both the classifiers are suitable for VAD. Since MSTE coefficients computation is simpler compared to the MFCC and all the features (MFCC, MSTE coefficients and OFS) together perform marginally better than the performance of MSTE coefficients with OFS thus MSTE coefficients with OFS feature set is suitable for VAD.

Chapter 6

DOA System

This chapter presents the experimental results for TDE when applied to the voiced speech segments of the audio signal recorded in a classroom. The DOA system consists of the Time Difference Of Arrival (TDOA) based DOA estimator and a supervised machine learning algorithm based VAD system. The time delay estimator is the modified YIN based TDE from chapter 4. The voiced speech is identified by the Linear Discriminant Analysis (LDA) supervised machine learning algorithm and selected signal features of chapter 5. The DOA system is called MINDER (Modified yIN based Doa Estimator for e-leaRning). Experimental setup is described in the first section of this chapter. This is followed by sections on the experimental results for the MINDER and the conclusion of the chapter.

6.1 Experimental Setup

An audio signal consisting of speech by both male and female talkers, silence and non-stationary noise was played through a loudspeaker and recorded by a ULA in a classroom situated next to a busy street. The audio signal was of 4.5 s duration of which the voiced speech was approximately 1.1 s. The windows of the room were closed

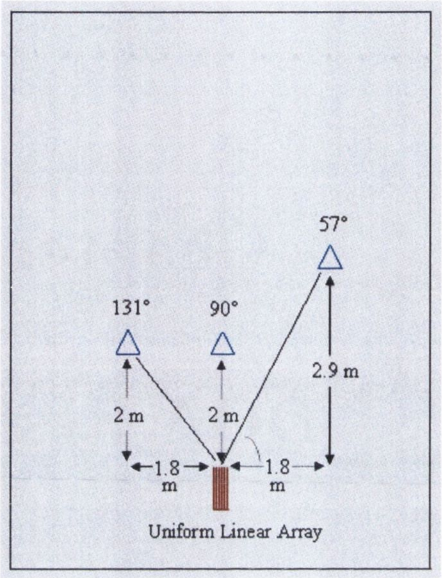


Fig. 6.1: Source locations for DOA system experiment.

due to the presence of Heating, Ventilating, Air Conditioning (HVAC) systems. The closed windows reduced the noise from the street. The sampling rate of the signals was 16 kHz. The silence in the audio signal was introduced to record the background noise signal present in the room. The background noise in the room consisted of the noise from HVAC and a low level traffic noise. The non-stationary noise consisted of the paper shuffle, male and female cough signals. The dimensions of the room were $6m \times 3.5m \times 2.5m$ and the $RT60$ of the room was 0.5 s. The simulation results for TDE for the same room are presented in section 4.5 of chapter 4.

A Uniform Linear Array (ULA) consisting of two microphones with the microphones spaced 6.8 cm apart was placed in the room. The source locations are shown in figure 6.1. The source locations were chosen such that they covered three areas (left, center and right) of the room with respect to the ULA. The signal source locations were: position 1: 57°, position 2: 90° and position 3: 131° with respect to the ULA. The signal source and the microphone array were on the same plane. The source at

Feature Set	Voiced Speech		Unvoiced Speech		Non-Stationary Noise		Stationary Noise		Overall
	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR
LDA with Mel-TEC + OFS feature set	100	1.75	71.88	3.13	68.5	3.97	96.97	12	84.82

Table 6.1: Hit Rate (HR), False Alarm Rate (FAR) for each class and overall HR for the experimental signal using LDA.

position 1 was beside the window and hence closest to the traffic noise.

Signal frames of length 20 ms were taken for DOA calculation with no overlap between the adjacent frames. The TDE was performed on the signal frames identified as voiced speech by the LDA based VAD system. The features included in the feature set were the MSTE coefficients and the Other Feature Set (OFS) of chapter 5. The features in the OFS included the log energy, zero-crossing rate, skewness of the LP residual signal, spectral centroid, absolute difference between the first two PARCOR coefficients and the second, third LP coefficients. The separation distance between the microphones in the microphone pair was set at 6.8 cm based on the simulation results of chapter 4. It is a compromise between the 3.4 cm separation where the number of delays is less to provide higher angle resolution and 10.2 cm separation where the number of delays is more, leading to higher angle resolution, but also has a lower accuracy in TDE. Parabolic interpolation (discussed in section 2.5) in steps of 0.2 sample value was used to improve the angular resolution.

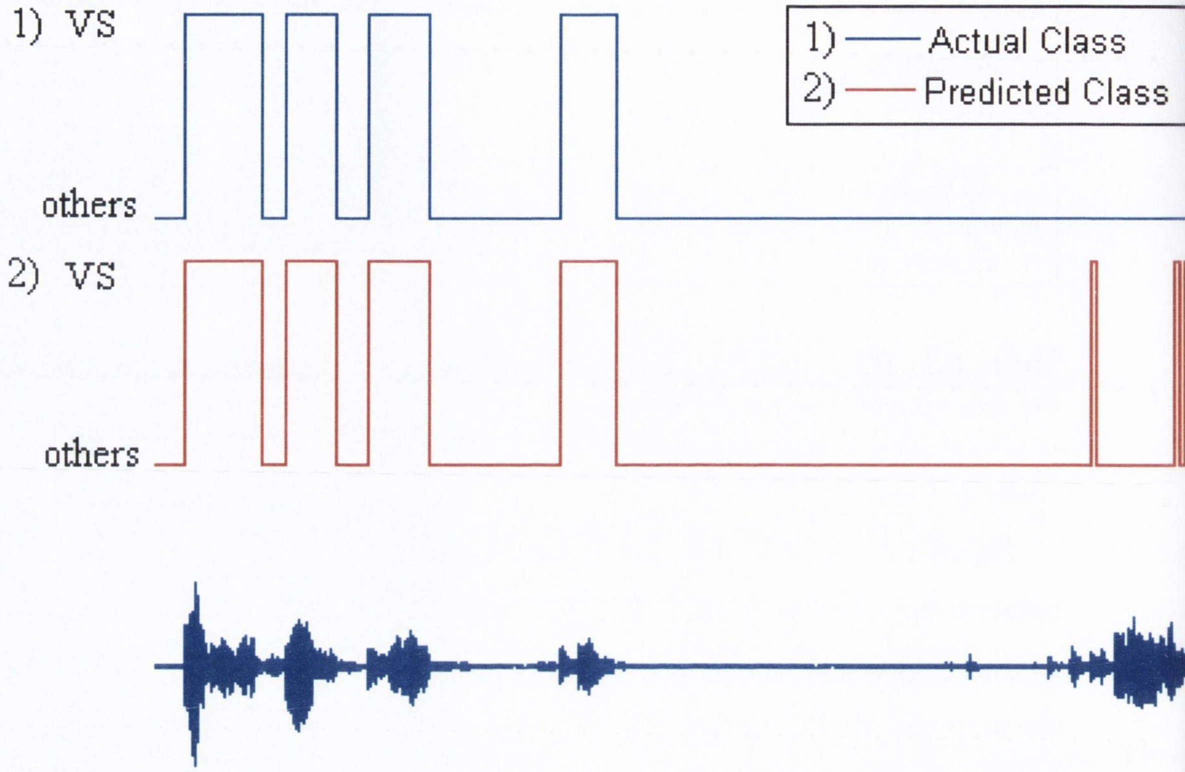


Fig. 6.2: VAD Experimental results classifying the audio signal into Voiced Speech (VS) and others (consisting of the unvoiced speech, stationary noise and the non-stationary noise signals).

6.2 Direction of Arrival System Experimental Results

Table 6.1 presents results for classifying the audio signal into four classes. Figure 6.2 shows the classification of the signal into voiced speech and others (unvoiced speech, stationary noise and non-stationary noise together) classes. All the voiced speech signals were identified correctly. Some of the non-stationary noise signal frames (cough) were classified as the voiced speech resulting in the FAR of 1.75%.

The TDE was performed only on the voiced speech signal. The experimental results for TDE are presented as histograms of the delay estimates in terms of signal samples, indicating the accuracy of the estimate. For the source at position 1 (57°), the signal reaches the reference microphone before the other microphone in the microphone pair. The expected delay for the microphones separated by 6.8 cm is 1.71 samples, which when rounded to the nearest integer value gives 2 samples without interpolation and is 1.6 or 1.8 with interpolation in steps of 0.2 samples.

Figure 6.3 presents the results of TDE on the source signal at position 1. The peak of the delay estimates without VAD is at the expected sample delay of 2. The maximum number of the delay estimates with VAD is also at the expected sample delay of 2. With VAD, incorrect delay estimates at sample delay -3 are eliminated. The sample delay estimates with VAD and parabolic interpolation are equal in number at the fractional sample delays of 1.6 and 2.2. Sample delays of 1.6 and 2.2 correspond to 59.8° and 46° respectively. Note that rouding the interpolated values yields the integer accurate estimate of 2.0 samples.

The experimental results of TDE for the source at 90° are presented in figures 6.4. The expected delay is 0 with and without parabolic interpolation. The peak of the histogram is at the expected delay without VAD and parabolic interpolation. The histogram of delay estimates with VAD on voiced speech results in a peak at the expected delay. With VAD the incorrect delays due to other signals at delay values of 3 and -3 are eliminated. Applying parabolic interpolation on the delay estimates results in a maximum number of delay estimates at the expected delay. But here the number of delay estimates at the expected delay is less than the number of delay estimates without parabolic interpolation. Large numbers of delay estimates have values of 0.2 and -0.2. 0.2 and -0.2 correspond to 86.4° and 93.6° respectively.

Figure 6.5 presents the TDE results for source at an angle of 131° (position 3). The expected delay is -2.1. Without parabolic interpolation the expected result is -2 and with parabolic interpolation the expected delay is -2 or -2.2. This position being closest

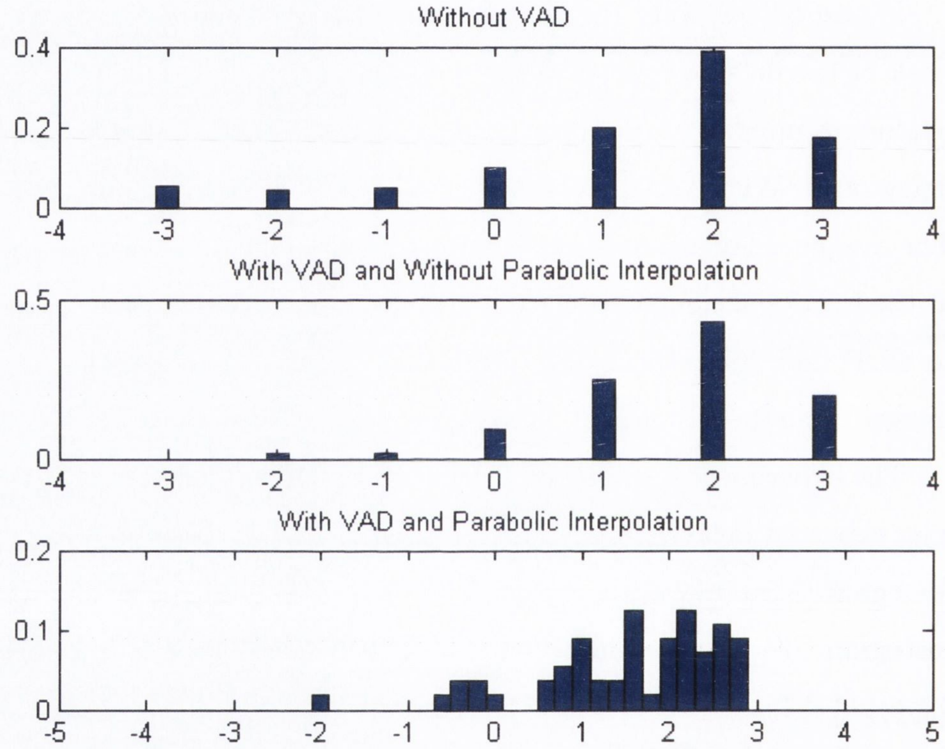


Fig. 6.3: Position 1: TDE results for source at 57°.

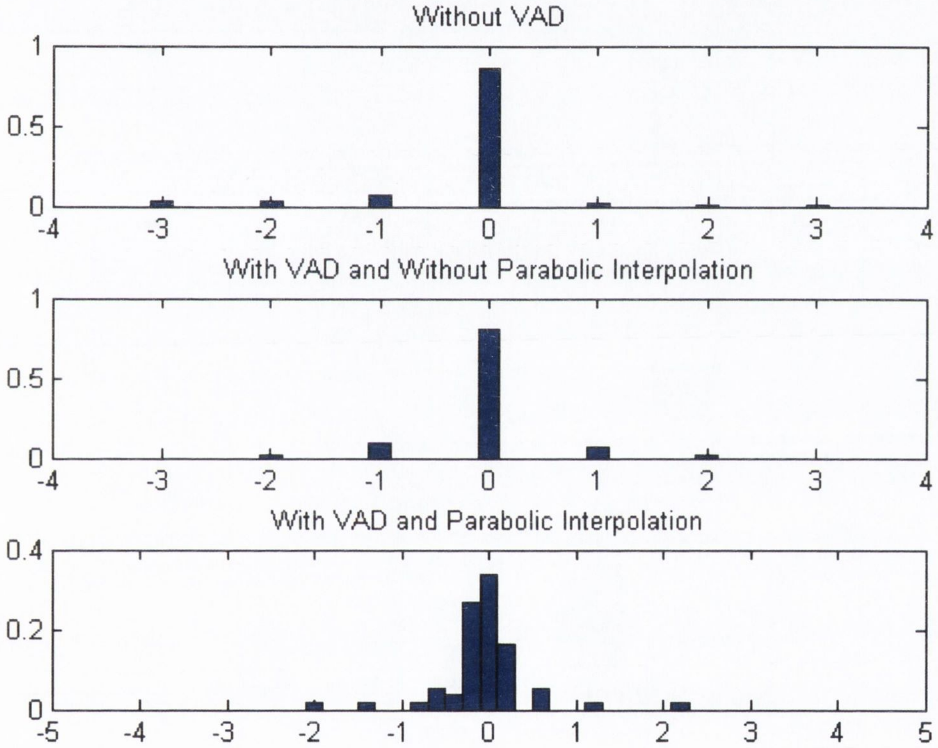


Fig. 6.4: Position 2: TDE results for source at 90°.

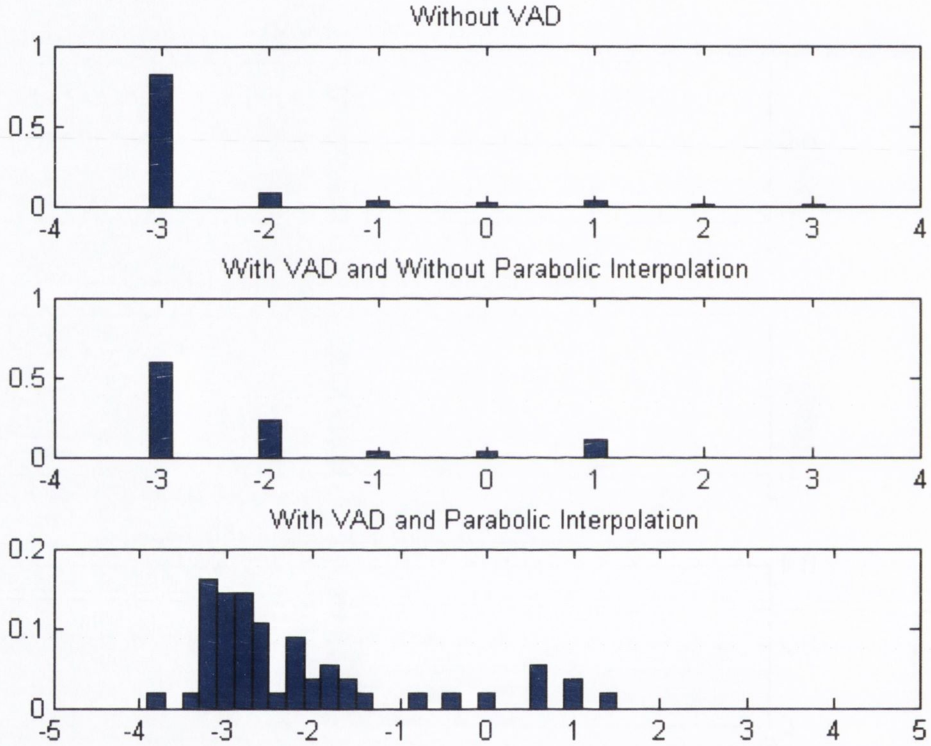


Fig. 6.5: Position 3: TDE results for source at 131°.

to the window, the delay estimates for it are more affected by the traffic noise than the delay estimates for the other two source positions. Without VAD, the peak of the delay estimates is at the sample delay of -3 (which is towards the traffic noise). With VAD, the peak of the delay estimates is still at -3 but the difference between the number of delay estimates at -3 and -2 is reduced. When parabolic interpolation is applied on the delay estimates, the number of delay estimates at delay value of -3 decreases. The delay estimates around the value of 2.2 increases. For this source position the benefits of applying VAD before TDE is noticeable.

6.3 Conclusion

This chapter demonstrated the performance of the MINDER combining the modified YIN based TDE algorithm of chapter 4 and the VAD system of chapter 5 for DOA in a classroom environment. The experimental results of VAD for classifying the signal into all the four classes was presented in section 6.1. From the results it is observed that in the presence of low background noise the proposed VAD system was able to classify voiced speech with a high HR and low False Alarm Rate (FAR). The FAR was due to the presence of human produced non-stationary noise signal. The HR for the unvoiced speech and non-stationary noise signals were lower than the other two class signals. Stationary noise had a high HR of approximately 97% but it also had a very high FAR of 12%.

Since the TDE for DOA estimation was performed on the voiced speech frames, the VAD using LDA and MSTE with OFS features was found to be suitable for this application as it was able to identify the voiced speech signal with high accuracy and low FAR. It was also shown that by applying VAD to identify voiced speech signal frames in order to perform time delay estimation only on voiced speech reduced the number of incorrect TDE computations due to presence of other type of signals.

The time delay estimates using the modified YIN based TDE on voiced speech provided a greater number of correct delay values at the expected sample delay without parabolic interpolation. When parabolic interpolation was employed to improve the angular resolution, its effect on the results was not alike in all the cases. Employing parabolic interpolation improved the resolution in the obtained angles as the fractional sample delay estimates moved towards the expected fractional delay values.

This chapter successfully combined the proposed VAD system and the proposed TDE algorithm for DOA estimation in a classroom.

Chapter 7

Conclusion

This chapter first presents the summary of the thesis followed by the future work.

7.1 Thesis Summary

Time Difference Of Arrival (TDOA) approach is suitable for Direction Of Arrival (DOA) estimation for a single source. As discussed in section 2.10, the DOA estimation using the TDOA approach is suitable for e-learning environment. In the TDOA approach first the relative time difference between the signals to reach a microphone pair in the microphone array is computed. This information is used to compute the DOA of the source signal. Cross-correlation on the microphone pair signals can be used to compute the time delay value. However, presence of noise and reverberation in a classroom leads to incorrect DOA estimation of the signal source location using Cross-Correlation (CC). Furthermore the performance of Time Difference Of Arrival (TDOA) algorithms such as Generalized Cross-Correlation Phase Transform (GCC-PHAT), Average Magnitude Difference Function (AMDF) while giving some improvement over simple CC also deteriorate in the presence of both high noise and reverberation.

This thesis presented a novel approach for DOA using the modified YIN based TDE algorithm. The original YIN algorithm was proposed for estimating the fundamental frequency of a signal. As explained in section 2.12 the YIN algorithm provides the fundamental frequency estimate by finding the similarity between the signal and its delayed versions. The delay value at which the signals are most similar is taken as the fundamental frequency of the signal. The TDE is performed by on the signals received by a pair of microphones by delaying one of the signals with respect to the other and finding the similarity between the two signals at each delay value. The delay value at which the signals are most similar is taken as the time delay. Due to the similarity between the way YIN estimates the fundamental frequency of the signal and the TDE computation the thesis shows that a modified YIN algorithm is suitable for TDE.

Simulation results comparing the performance of GCC-PHAT, contemporary algorithms such as Modified Average Magnitude Difference Function (MAMDF) and Weighted cross-correlation (WCC), proposed modified YIN based TDE algorithm and the Weighted YIN based TDE algorithm were presented in section 4.1 and section 4.2 respectively. The performance of the algorithms was compared in terms of their accuracy and the Mean Square Error (MSE) for different values of SNR. The SNR values considered for simulations were 10, 15 and 20 dB because as mentioned in section 2.6 the acceptable SNR in a classroom is greater than 10 dB.

The simulation results showed that the modified YIN based TDE algorithm had the best accuracy as defined by the maximum mode of the distribution of the delay estimates but it had a higher MSE compared to the other algorithms. The contemporary MAMDF algorithm was best next to the modified YIN based TDE algorithm but had lower MSE compared to the modified YIN based algorithm. The WCC algorithm and the Weighted YIN based TDE algorithm had similar performance in both accuracy and MSE. The GCC-PHAT had the worst performance in terms of accuracy. The GCC-PHAT, WCC and the Weighted YIN based TDE algorithm were found to be biased towards the zero delay value which resulted in lower MSE whereas the other two algorithms were biased towards the extreme delay values in the valid delay value set leading to higher MSE. Since for DOA estimation, accuracy is important hence for Time Delay Estimation (TDE) in the proposed DOA system for e-learning modified YIN based TDE algorithm was used.

As mentioned in section 1.2, a pre-processing Voice Activity Detection (VAD) system that identifies part of speech that contain speech is another part of the proposed DOA system. VAD is useful in avoiding DOA computations on parts of audio signal that do not contain speech. The VAD pre-processing stage not only reduces the number of DOA computations but also prevents steering the video camera in direction of undesired signal sources.

The VAD process classified the signal into four categories namely the voiced speech, unvoiced speech, stationary noise and the non-stationary noise. The signal was classified into four classes for TDE on the voiced speech and for post processing e-learning. Supervised machine learning algorithms such as the Discriminant Analysis (DA), Artificial Neural Network (ANN) and Support Vector Machine (SVM) were considered for classification.

Initially for VAD, supervised machine learning algorithms such as the Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA) and the Mahalanobis Distance (MD) belonging to the DA category; cascaded ANN and radial, linear, polynomial and sigmoid kernel SVM were considered for VAD. The features used in the initial classification to decide on the supervised machine learning algorithm suitable for classifying the signal into four classes are mentioned in section 5.3 and discussed in section 3.3.

Features were computed on 20ms long signal frames. Forward and backward feature selection using DA was performed on the feature set to assess the significance of the features in classification and remove the features that do not contribute towards classification. All the features were found to be significant. The classification algorithms were trained and then tested for classification. The training data and the testing data were different. For training and testing purpose all the signal frames consisting of two or more signal classes were excluded. Based on the classification Hit Rate and False Alarm Rate for each class in section 5.3 only LDA and radial kernel SVM were further considered for inclusion in the VAD stage of the DOA system.

This thesis also proposes a novel mel-spectrum and Teager Energy Operator based feature called the Mel-Spectrum Teager Energy (MSTE) coefficients. These features are presented in section 5.1. The performance of these features was compared with the frequently used Mel Frequency Cepstrum Coefficients (MFCC) in speech signal processing applications. As shown is section 5.5 the MSTE coefficients were found to perform as well as or marginally better than the MFCC for classifying the signal into four classes at the considered SNR of 10, 15 and 20 dB for both LDA and radial SVM. The advantage of these coefficients is that they are simpler to compute than the MFCC.

The MSTE coefficients on their own were not able to effectively classify all the classes especially the non-stationary noise class signal. The MFCC coefficients also suffered from this problem. Forward feature selection in DA was performed on the feature set mentioned in section 5.4. Forward feature selection was employed because the aim of performing feature selection was to find the features that perform well with the MSTE coefficients in the feature set. The features that were found to be significant for classification along with the MSTE coefficients are presented in section 5.4. All the selected features except the MSTE were grouped into Other Feature Set (OFS).

The results of the performance of the MSTE coefficients with the OFS and MFCC with OFS using LDA and radial SVM are presented in section 5.5 for 10, 15 and 20 dB SNR. Both the feature sets performed equally well for classification into four classes with a high HR and low FAR for all the classes. The classification results for classification with only OFS was also presented to assess their classification without

MSTE coefficients. OFS was able to classify non-stationary noise signal well but had a low HR for the voiced speech signal. The performance of all the features together was also presented in section 5.5. All the features together performed marginally better than only MSTE with OFS or MFCC with OFS. Based on the simulation results MSTE with OFS was included in the final feature set for VAD.

Comparing the classification results of LDA with radial SVM in section 5.5 for the above mentioned feature sets (only MFCC, only MTSE, only OFS, MFCC with OFS, MTSE with OFS and all the three together), the performance of LDA was found to be marginally better than that of the radial SVM hence LDA was chosen for VAD. Thus the final VAD stage in the proposed DOA system consisted of the LDA machine learning algorithm with MTSE and OFS features.

Finally in chapter 6 the LDA classification algorithm with MTSE and OFS features for VAD was combined with the TDE using the modified YIN based TDE algorithm to form the DOA system. The results of the experimental results of the DOA system on the data recorded in a classroom are presented in section 6.2. In the experimental results the frames consisting of two or more signal classes were not included. The LDA was trained with the training data and the signal recorded in the classroom was classified into four classes.

TDE was performed on the signal frames classified as voiced speech. The VAD system was able to identify all the voiced speech frames and had a low False Alarm Rate (FAR). The FAR was due to the classification of some of the non-stationary noise signals into voiced class. Parabolic interpolation (described in section 2.5) was used on the TDE to improve the resolution of the angle estimates.

7.2 Future Work

From the experimental results for TDE in section 4.5 it was concluded that modified YIN based TDE has higher accuracy than the other algorithms but there is still a scope for improving the accuracy of the TDE. In section 4.6 it was also mentioned that modified YIN based TDE is biased towards the extreme delay value of the TDE when the delay estimate is incorrect. In this context, further investigation is required to add post TDE logic correction in order to improve the accuracy of the TDE. The DOA information from different microphone pairs can be used to find the source location.

Another aspect of the TDE was using the parabolic interpolation to improve the resolution in the angle. The experimental results with parabolic interpolation were presented in section 6.2. The parabolic interpolation was not as helpful in all the cases. Further research into the suitable interpolation method can be done.

Since the work concentrated on finding the DOA based on TDE on voiced speech, hence even though the signal was classified into four classes the primary concern was to identify voiced speech signals with high HR and low FAR. This was achieved in the work. Further research into identifying undesirable noises such as cough, sneeze, paper shuffle with high accuracy so that they can be effectively removed automatically, without much human intervention, for post-processing e-learning so that the students are able to access the lectures online or on a Compact Disc (CD) is required.

A limitation of the VAD work presented in this thesis is that the frames consisting of two or more signal classes (non-overlapping) have not been included for training or testing. In real time application these frames will be classified into one of the four classes depending on the training algorithm. This will result in incorrect VAD decision. Further research on identifying the frames consisting of two or more classes so that they can be classified into the other class (not voiced speech) to reduce incorrect VAD decisions is required.

Moreover for post-processing e-learning the word boundaries have to be exactly identified which is not required in the case of TDE. This is because a few ms of missing signal can change the meaning of the words that are being spoken whereas a camera can be steered a few ms after the words have been spoken. The frames with mixed signals can be further worked on to identify the word boundaries to improve the post-processing e-learning.

Another aspect that was not considered in this work was the overlap of signals of different classes (different class signals are present simultaneously) in a frame. All the frames that were considered in the work consisted of a single class signal. For VAD simulations in different SNR, white Gaussian noise was taken as the background noise. Scenarios where the voiced speech is present with other signal class in the background were not included. Further investigations on classifying the signals consisting of speech with non-stationary noise in the background will be desirable to make the VAD system more robust to identify speech especially voiced for TDE. Since the MSTE coefficients perform on par with the MFCC, their application in speech recognition and speaker recognition applications using the Hidden Markov Models should be explored.

Bibliography

[1] http://www.blackboard.com/us/index.Bb, last accessed: June 2007.

[2] http://www.echalk.de/, last accessed: June 2007.

[3] http://iddl.vt.edu/instructors/ivc/intheroom.php, last accessed: Dec 2007.

[4] http://www.asha.org/default.htm, last accessed: June 2007.

[5] http://www.renkus-heinz.com/ease/index.html, last accessed: June 2007.

[6] http://www.mpi.nl/world/tg/corpora/timit/timit.html, last accessed: June 2007.

[7] http://www.adobe.com/special/products/audition/syntrillium.html, last accessed: June 2007.

[8] http://www.spss.com/, last accessed: June 2007.

[9] http://www.mathworks.com/, last accessed: June 2007.

[10] http://www.mathworks.com/products/statistics/, last accessed: June 2007.

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/, last accessed: June 2007.

[12] http://leenissen.dk/fann/, last accessed: June 2007.

[13] `http://www.utdallas.edu/~loizou/speech/colea.htm`, last accessed: June 2007.

[14] `http://labrosa.ee.columbia.edu/matlab/rastamat/`, last accessed: July 2007.

[15] Shigeo Abe, *Support vector machines for pattern classification (advances in pattern recognition)*, Springer-Verlag London Limited, 2005.

[16] Bishnu S. Atal and Lawrence R. Rabiner, *A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing **24, Issue 3.** (1976), 201–212.

[17] A Bendiksen and K Steiglitz, *Neural networks for voiced/unvoiced speech classification*, International Conference on Acoustics, Speech and Signal Processing, vol. 1, Apr 1990, pp. 521–524.

[18] Adil Benyassine, Eyal Shlomot, Huan-Yu Su, Dominique Massaloux, Claude Lamblin, and Jean-Pierre Petit, *Itu-t recommendation g.729 annex b: A silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications*, IEEE Communications Magazine (1997, Issue 9.), 64–73.

[19] F Beritelli, S Casale, G Ruggeri, and S Serrano, *Performance evaluation and comparison of g.729/amr/fuzzy voice activity detectors*, IEEE Signal Processing Letters **9, Issue 3.** (2002), 85–88.

[20] Francesco Beritelli, Salvator Casale, and Alfredo Cavallaero, *A robust voice activity detector for wireless communications using soft computing*, IEEE Journal on Selected Areas in Communications **16, Issue 9.** (1998), 1818–1829.

[21] Stanley T. Birchfield, *A unifying framework for acoustic localization*, European Signal Processing Conference, Sep 2004, pp. 1127–1130.

[22] Stanley T. Birchfield and Daniel Kahn Gillmor, *Acoustic source direction by hemisphere sampling*, IEEE conference on Acoustics, Speech and Signal Processing, vol. 5, 2001, pp. 3053–3056.

[23] Sylvio R. Bistafa and John S. Bradley, *Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics*, The Journal of Acoustical Society of America **107, Issue 2.** (2000), 861–875.

[24] Michael Brandstein and Darren Ward (eds.), *Microphone arrays: Signal processing techniques and applications (digital signal processing)*, Springer, 2001.

[25] Michael S. Brandstein, *A pitch-based approach to time-delay estimation of reverberant speech*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct 1997.

[26] Michael S Brandstein, *Time-delay estimation of reverberated speech exploiting harmonic structure*, Acoustical Society of America **105, Issue 5.** (1999), 2914–2919.

[27] Michael S. Brandstein, John E. Adcock, and Harvey F. Silverman, *A practical time-delay estimator for localizing speech sources with a microphone array*, Computer, Speech and Language **9** (1995), 153–159.

[28] M. J. Carey, E. S. Parris, and H Lloyd-Thomas, *A comparison of features for speech, music discrimination*, IEEE conference on Acoustics, Speech and Signal Processing, vol. 1, Mar 1999, pp. 149–152.

[29] G. Clifford Carter, *Coherence and time delay estimation*, Proceedings of the IEEE **75, Issue 2.** (1987), 236–255.

[30] G. Clifford Carter, Albert H. Nuttall, and Peter G. Cable, *The smoothed coherence transform*, Proceedings of the IEEE **61, Issue 10.** (1973), 1497–1498.

[31] J. C. Chen, Kung Yao, and R. E. Hudson, *Source localization and beamforming*, IEEE Signal Processing Magazine **19, Issue 2.** (2002), 30–39.

[32] Jingdong Chen, Jacob Benesty, and Yiteng (Arden) Huang, *Performance of gcc- and amdf-based time-delay estimation in practical reverberant environments*, EURASIP Journal on Applied Signal Processing (2005), 25–36.

[33] Etienne Cornu, Nicolas Destrez, Alain Dufaux, Hamid Sheikhzadeh, and Robert Brennan, *An ultra low power, ultra miniature voice command system based on hidden markov models*, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, 2002, pp. IV–3800– IV–3803.

[34] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine Learning **20** (1995), no. 3, 273–297.

[35] V. Davidek, J. Sika, and J. Stusak, *Implementing a noise cancellation system with the tms320c31*, First European DSP Education and Research Conference (ESIEE), Sept. 1996.

[36] Alain de Cheveigne and Hideki Kawahara, *Yin, a fundamental frequency estimator for speech and music*, Acoustical Society of America **111, Issue 4.** (2002), 1917–1930.

[37] Paulo S. R. Diniz, Eduardo A. B. da Silva, and Sergio L. Netto, *Digital signal processing system analysis and design*, Cambridge University Press, 2002.

[38] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern classification*, second ed., John Wiley and Sons, Inc., 2001.

[39] Khaled El-Maheh and Peter Kabal, *Comparison of voice activity detection algorithms for wireless personal communications systems*, IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, May 1997, pp. 470–473.

[40] Dong Enqing, Liu Guizhong, Zhou Yatong, and Zhang Xiaodi, *Applying support vector machines to voice activity detection*, International Conference on Signal Processing, vol. 2, Aug 2002, pp. 1124–1127.

[41] Engin Erzin, A. Enis Cetin, and Yasemin Yardimci, *Subband analysis for robust speech recognition in the presence of car noise*, International Conference on Acoustics, Speech and Signal Processing, vol. 1, May 1995, pp. 417–420.

[42] P. A. Estevez, N. Becerra-Yoma, N. Boric, and J. A. Remirez, *Genetic programming-based voice activity detection*, Electronics Letters **41, Issue 20.** (2005), 1141–1143.

[43] Scott E Fahlman and Christian Lebiere, *The cascade-correlation learning architecture*, Advances in Neural Information Processing Systems, vol. 2, 1990, pp. 524–532.

[44] Stefaan Van Gerven and Fei Xie, *A comparative study of speech detection methods*, EUROSPEECH, 1997, pp. 1095–1098.

[45] Lal C. Godara, *Application of antenna arrays to mobile communications, part ii: Beam-forming and direction-of-arrival considerations*, Proceedings of the IEEE, vol. 85, 1997, pp. 1195–1245.

[46] Lal Chand Godara, *Smart antennas*, CRC Press, 2004.

[47] Scott Matthew Griebel, *A microphone array system for speech source localization, denoising and dereverberation*, Ph.D. thesis, Harvard University, 2002.

[48] Isabelle Guyon and Andre Elisseeff, *An introduction to variable and feature selection*, The Journal of Machine Learning Research **3** (2003), 1157–1182.

[49] J. A Haigh and J. S Mason, *Robust voice activity detection using cepstral features*, IEEE Region 10 Conference on Computer, Communication, Control and Power Engineering, vol. 3, Oct 1993, pp. 321–324.

[50] B.V. Harsha, *A noise robust speech activity detection algorithm*, Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing., 2004.

[51] Lothar Hermes and Joachim M. Buhmann, *Feature selection for support vector machines*, Proceedings of International Conference on Pattern Recognition, vol. 2, Sep 2000, pp. 712–715.

[52] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, *A practical guide to support vector classification*, `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`, last accessed: June 2007.

[53] Chih-Wei Hsu and Chih-Jen Lin, *A comparison of methods for multiclass support vector machines*, IEEE Transactions on Neural Networks **13, Issue 2.** (2002), 415–425.

[54] Carl J Huberty, *Applied discriminant analysis*, John Wiley and Sons Inc., 1994.

[55] K. Itoh and M. Mizushima, *Environmental noise reduction based on speech/non-speech identification for hearing aids.*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Apr 1997, pp. 419–422.

[56] Firas Jabloun and A. Enis Cetin, *The teager energy based feature parameters for robust speech recognition in car noise*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Mar 1999, pp. 273–276.

[57] Firas Jabloun, A. Enis Cetin, and Engin Erzin, *Teager energy based feature parameters for speech recognition in car noise*, IEEE Speech Processing Letters **6, Issue 10.** (1999), 259–261.

[58] Giovanni Jacovitti and Gaetano Scarano, *Discrete time techniques for time delay estimation*, IEEE Transactions on Signal Processing **41, Issue 2.** (1993), 525–533.

[59] Sabina Jeschke, Lars Knipping, Raul Rojas, and Ruedi Seiler, *Intelligent chalk-system for modern teaching: in math, science and engineering areas*, `http://kazan.inf.fu-berlin.de/echalk/docs/ChalkSystems_ASEE2006_Preprint.pdf`, last accessed: Sept 2007.

[60] Don H. Johnson and Dan E. Dudgeon, *Array signal processing: Concepts and techniques*, Simon and Schuster, 1992.

[61] John R. Deller Jr, John G Proakis, and John H L Hansen, *Discrete-time processing of speech signals*, Macmillan Publishing Company, 1993.

[62] Jean-Claude Junqua, Ben Reaves, and Brian Mak, *A study of endpoint detection algorithms in adverse conditions: Incidence on a dtw and hmm recognizer*, Eurospeech Second European Conference on Speech Communication and Technology, Sep 1991, pp. 1371–1374.

[63] James F. Kaiser, *On a simple algorithm to calculate the 'energy' of a signal*, International Conference on Acoustics, Speech and Signal Processing, vol. 1, Apr 1990, pp. 381–384.

[64] Supphanat Kanokphara and Julie Carson-Berndsen, *Articulatory-acoustic-feature-based automatic language identification*, `http://muster.ucd.ie/pubs/kanokphara_multiling06.pdf`, last accessed: June 2007.

[65] Darren F. Kavanagh and Frank Boland, *A non-linear operator based method for harmonic feature extraction from speech signals*, IEEE International Conference on Signal Processing and Communications, Nov 2007.

[66] M. Kashif Saeed Khan, Wasfi G. Al-Khatib, and Muhammad Moinuddin, *Automatic classification of speech and music using neural networks*, MMDB '04: Proceedings of the 2nd ACM international workshop on Multimedia databases, ACM Press, 2004, last accessed: June 2007, pp. 94–99.

[67] Yusuke Kida and Tatsuya Kawahara, *Voice activity detection based on optimally weighted combination of multiple features*, Interspeech, Sep 2005, pp. 2621–2624.

[68] H.-I. Kim and S.-K. Park, *Voice activity detection algorithm using radial basis function network*, IEEE Electronics Letters **40, Issue 22.** (2004), 1454–1455.

[69] Hyun-Don Kim, Jong-Suk Choi, Munsang Kim, and Chang-Hoon Lee, *Reliable detection of sound's direction for human robot interaction*, International Conference on Intelligent Robots and Systems, vol. 3, Oct 2004, pp. 2411–2416.

[70] C.H. Knapp and G.C.Carter, *The generalized correlation method for estimation of time delay.*, IEEE Transactions on Acoustics, Speech and Signal Processing **24, Issue 4.** (1976), 320–327.

[71] Trausti Kristjansson, Sabine Deligne, and Peder Olsen, *Voicing features for robust speech detection*, Interspeech 9th European Conference on Speech Communication and Technology, Sep 2005, pp. 369–372.

[72] J. P. Kuhn, *Detection performance of the smooth coherence transform (scot)*, The IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 3, Apr 1978, pp. 678–683.

[73] Oh-Wook Kwon and Te-Won Lee, *Optimizing speech/non-speech classifier design using adaboost*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, Apr 2003, pp. I–436–I–439.

[74] I.D Lee, H.P Stern, and S.A. Mahmoud, *A voice activity detection algorithm for communication systems with dynamically varying background acoustic noise*, 48th IEEE Vehicular Technology Conference, vol. 2, May 1998, pp. 1214–1218.

[75] Ke Li, M.N.S Swamy, and M.O. Ahmad, *An improved voice activity detection using higher order statistics.*, IEEE Transactions on Speech and Audio Processing **13, Issue 5.** (2005), 965–974.

[76] Lie Lu, Hao Jiang, and HongJiang Zhang, *A robust audio classification and segmentation method*, Ninth ACM International Conference on Multimedia, Sep 2001, pp. 203–211.

[77] John Makhoul, *Linear prediction: A tutorial review*, Proceedings of the IEEE **63, Issue 4.** (1975), 561–580.

[78] Klaus-Robert Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf, *An introduction to kernel-based learning algorithms*, IEEE Transactions on Neural Networks **12, Issue 2.** (2001), 181–201.

[79] Enzo Mumolo, Massimiliano, and Gianni Vercelli, *Algorithms for acoustic localization based on microphone array in service robots*, IEEE International Conference on Robotics and Automation, vol. 3, Apr 2000, pp. 2966–2971.

[80] Wrigley S. N., Brown G. J., Wan V., and Renals S., *Speech and crosstalk detection in multichannel audio*, IEEE Transactions on Speech and Audio Processing **13, Issue 1.** (2005), 84 – 91.

[81] Amir Navot, Ran Gilad-Bachrach, Yiftah Navot, and Naftali Tishby, *Is feature selection still necessary?*, Subspace, Latent Structure and Feature Selection, Statistical and Optimization, Perspectives Workshop. Lecture Notes in Computer Science **3940** (2006), 127–138.

[82] E. Nemer, R. Goubran, and S. Mahmoud, *Robust voice activity detection using higher-order statistics in the lpc residual domain.*, IEEE Transactions on Speech and Audio Processing **9, Issue 3.** (2001), 217–231.

[83] Steffen Nissen, *Neural networks made simple*, `http://leenissen.dk/fann/`, last accessed: June 2007.

[84] Jan Nouza, *Feature selection methods for hidden markov model-based speech recognition*, International Conference on Pattern Recognition, vol. 2, Aug 1996, pp. 186–190.

[85] P.D. Olivier, *Approximating irrational transfer functions using lagrange interpolation formula*, IEE Proceedings **139, Issue 1.** (1992), 9–12.

[86] M. Omologo and P. Svaizer, *Use of the cross-power spectrum phase in acoustic event localization*, `citeseer.ist.psu.edu/omologo93use.html`, 1993, last accessed: June 2007.

[87] Thomas W. Parsons, *Voice and speech processing*, McGraw-Hill, Inc., 1987.

[88] R. Venkatesh Prasad, Abhijeet Sangwan, H.S. Jamadagni, M. C. Chiranth, Rahul Sah, and Vishal Gaurav, *Comparison of voice activity detection algorithms for*

voip, Seventh International Symposium on Computers and Communications, 2002, pp. 530–535.

[89] William H. Press, *Numerical recipes in fortran: The art of scientific computing*, Cambridge University Press, 1992.

[90] Jose C Principe, Neil R Euliano, and W Curt Lefebvre, *Neural and adaptive systems: Fundamentals through simulations*, John Wiley and Sons, Inc., 2000.

[91] Angela Elizabeth Quinlan, *Characterisitcs of direction of arrival (doa) estimators in the acoustic context*, Ph.D. thesis, University of Dublin, Trinity College, 2006.

[92] L Rabiner and M. Sambur, *Voiced-unvoiced-silence detection using the itakura lpc distance measure.*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, May 1977, pp. 323–326.

[93] Lawrence R. Rabiner and Marvin R. Sambur, *Application of an lpc distance measure to the voiced-unvoiced-silence detection problem*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 25, Issue 4., Aug 1977, pp. 338–343.

[94] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*, Englewood Cliffs ; London [etc.] : Prentice-Hall, 1978.

[95] Dmitry Zotkin Ramani Duraiswami and Larry S. Davis, *Active speech source localization by a dual coarse-to-fine search*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, 2001, pp. 3309–3312.

[96] Brian D Ripley, *Pattern recognition and neural networks*, Cambridge University Press, 1996.

[97] Abhijeet Sangwan, M.C Chiranth, H.S Jamadagni, Rahul Sah, R. Venkatesha Prasad, and Vishal Gaurav, *Vad techniques for real-time speech transmission*

on the internet, IEEE International Conference on High Speed Networks and Multimedia Communications, 2002, pp. 46–50.

[98] Kent Scarbrough, Nasir ahmed, and G. Clifford carter, *An experimental comparison of the cross correlation and scot techniques for time delay estimation*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, Apr 1980, pp. 807–810.

[99] E Scheirer and M Slaney, *Construction and evaluation of a robust multifeature speech/music discriminator*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, Apr 1997, pp. 1331–1334.

[100] Benjamin Seep, Robin Glosemeyer, Emily Hulce, Matt Linn, Pamela Aytar, and Bob Coffeen, *Classroom acoustics. a resource for creating learning environments with desirable listening conditions.*, `http://asa.aip.org/classroom/booklet.html`, last accessed: June 2007.

[101] Samir Shaltaf, *Neural network based time delay estimation*, EURASIP Journal on Applied Signal Processing **3** (2004), 378–385.

[102] Ravi R. Shenoy, *Design of e-textiles for acoustic applications*, Master's thesis, Virginia Polytechnic Institute and State University, 2003.

[103] Won-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee, and Jong-Seok Lee, *Speech/non-speech classification using multiple features for robust endpoint detection*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, 2000, pp. 1399–1402.

[104] Leah J. Siegel, *A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier*, IEEE Transactions on Acoustics, Speech and Signal Processing **27, Issue 1.** (1979), 83–89.

[105] Deepti Singh and Frank Boland, *Time delay estimation using difference function*, ISSC 2005 - IEE Irish Signals and Systems Conference, Sep 2005, pp. 174–179.

[106] ______, *Voice activity detection*, Crossroads **13, Issue 4.** (2007), 7–7.

[107] K Srinivasan and A Gersho, *Voice activity detection for cellular networks*, IEEE Workshop on Speech Coding for Telecommunications, 1993, pp. 85–86.

[108] Alex Stephenne and Benoit Champagne, *Cepstral prefiltering for time delay estimates in reverberant environments*, International Conference on Acoustics, Speech and Signal Processing, vol. 5, May 1995, pp. 3055–3058.

[109] N. Strobel and R. Rabenstein, *Classification of time delay estimates for robust speaker localization*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, Mar 1999, pp. 3081–3084.

[110] Fotios Talantzis, Anthony G. Constantinides, and Lazaros C. Polymenakos, *Estimation of direction of arrival using information theory*, IEEE Signal Processing Letters **12, Issue 8.** (2005), 561–564.

[111] S.G Tanyer and H. Ozer, *Voice activity detection in nonstationary gaussian noise*, Fourth International Conference on Signal Processing, vol. 2, 1998, pp. 1620–1623.

[112] S.G. Tanyer and H. Ozer, *Voice activity detection in nonstationary noise*, IEEE Transactions on Speech and Audio Processing **8, Issue 4.** (2000), 478–482.

[113] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern recognition*, Academic Press, 2006.

[114] Jean-Marc Valin, Francois Michaud, Jean Rouat, and Dominic Letourneau, *Robust sound source localization using a microphone array on a mobile robot*,

IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 2, Oct 2003, pp. 1228–1233.

[115] Krishnaraj Varma, *Time-delay-estimation based direction-of-arrival estimation for speech in reverberant environments.*, Master's thesis, Virginia Polytechnic Institute and State University., 2002.

[116] Barry D. Van Veen and Kevin M. Buckley, *Beamforming: A versatile approach to spatial filtering*, IEEE ASSP Magazine **5, Issue 2.** (1988), 4–24.

[117] Rivarol Vergin, Douglas O' Shaughnessy, and Azarshid Farhat, *Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition*, IEEE Transactions on Speech and Audio Processing **7, Issue 5.** (1999), 525–532.

[118] Yao Wang, Zhu Liu, and Jin-Cheng Huang, *Multimedia content analysis*, IEEE Signal Processing Magazine **17, Issue 6.** (2000), 12–36.

[119] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, *Feature selection for SVMs*, Advances in Neural Information Processing Systems, 2000, pp. 668–674.

[120] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park, and Chungyong Lee, *Robust voice activity detection algorithm for estimating noise spectrum*, IEEE Transactions on Acoustics, Speech and Signal Processing **36, Issue 2.** (2000), 180–181.

[121] Tong Zhang and C-C Jay Kuo, *Hierarchial classification of audio data for archiving and retrieving*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, 1999, pp. 3001–3004.

[122] J. Zhao, J. Kuang, X. Xie, and H. Shilei, *Noise suppression based on teager energy*

operator for improving the robustness of ASR front-end, International Workshop on Acoustic Echo and Noise Control (kyoto), Sep 2003, pp. 135–138.

[123] Junhui Zhao, Jingming Kuang, and Qionghai Dai, *Robust speech recognition in noisy backgrounds based on teager energy operator and auditory process*, Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 1, Nov 2003, pp. 550–554.

[124] Guojun Zhou, John H. L. Hansen, and James F. Kaiser, *Methods for stress classification: Nonlinear TEO and linear speech based features*, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, Mar 1999, pp. 2087–2090.